AUGMENTING LINGUISTIC SEMI -STRUCTURED DATA FOR MACHINE LEARNING - A CASE STUDY USING FRAMENET

Breno W. S. R. Carvalho[1], Aline Paes[2] and Bernardo Gonçalves[3]

[1]IBM Research, Brazil. Institute of Computing, Universidade Federal Fluminense (UFF), Niterói, RJ, Brazil.
[2]Institute of Computing, Universidade Federal Fluminense (UFF), Niterói, RJ, Brazil.
[3]IBM Research, Brazil

ABSTRACT

Semantic Role Labelling (SRL) is the process of automatically finding the semantic roles of terms in a sentence. It is an essential task towards creating a machine-meaningful representation of textual information. One public linguistic resource commonly used for this task is the FrameNet Project. FrameNet is a human and machine-readable lexical database containing a considerable number of annotated sentences, those annotations link sentence fragments to semantic frames. However, while the annotations across all the documents covered in the dataset link to most of the frames, a large group of frames lack annotations in the documents pointing to them. In this paper, we present a data augmentation method for FrameNet documents that increases by over 13% the total number of annotations. Our approach relies on lexical, syntactic, and semantic aspects of the sentences to provide additional annotations. We evaluate the proposed augmentation method by comparing the performance of a state-of-the-art semantic-role-labelling system, trained using a dataset with and without augmentation.

KEYWORDS

FrameNet, Frame Semantic Parsing, Semantic Role Labelling, Data Augmentation.

1. INTRODUCTION

A large proportion of humankind's knowledge is stored in textual form. Nevertheless, such unstructured information is hard to search, catalog, and query. To circumvent this difficulty, one needs to automate the extraction of information from texts, making it amenable for querying. It relates to the emerging area of Machine Reading [1], a task within the broader area of Natural Language Processing, NLP. Machine Reading is concerned explicitly with creating machine-friendly, yet nuanced, representations of text. A crucial task in Machine Reading is the Semantic Role Labeling task, SRL [2]. SRL consists of mapping elements of a given sentence to predefined sets of semantic roles. There are two main kinds of labeling: (i) deep labeling, i.e., the mapping of tokens of the sentence to somewhat complex semantic structures by building a composable representation of the utterance meaning; and (ii) shallow labeling, that consists of mapping the tokens to an abstract semantic role. For instance, figure 1 shows two shallow roles, namely, Content and Paradigm, which provide meaning to two subsets of tokens in the sentence. The present work is concerned with shallow labeling, which is itself far from a trivial computational

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 01-13, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101201

task and is hardly feasible without a good set of labeled sentences, whereby "good" we mean a set of sentences whose tokens are annotated with their expected deep roles in relatively good coverage.

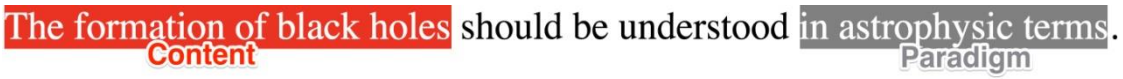


Figure 1: An example of shallow semantic roles assigned to tokens in a sentence.

One popular source of annotated sentences to support Machine Reading is FrameNet, a publicly available electronic language resource [3]. It consists of a network of concepts (called frames) such as Run, Motive and Location. Each frame is composed of frame elements, which define semantic roles in the (thereby semi-structured) domains. A key technical challenge, however, is that FrameNet's set distribution of examples forms a long tail — a few frame elements have several examples over their related frames. In contrast, most of them have only one or none example at all — making it difficult to tackle less popular frame elements. This need gets even more pressing when we target specific domains within FrameNet.

In this paper, we propose a data augmentation method to enlarge the set of annotations and its distribution in FrameNet. The technique leverages on partial structure present in the annotation of frame elements in the sentences. That is, we carry out matching of frame elements over different frames — relying on notions of lexical, syntactic, or semantic equivalence — so that sentences receive new (inferred) annotations. We take advantage of the inter-frame connections to enrich the information available in the resource.

In the next section, we describe the analyzers that enable us to process natural language sentences, the SRL method that supports our evaluation, and we provide a more detailed view on FrameNet. Then we also introduce background aspects, preparing towards our research problem. In section 3 we present the augmentation method we propose in this paper. In section 4 we report its evaluation, based on comparing the performance of a state-of-the-art semantic-role-labeling method, with and without augmentation. In section 5, we situate this work within the literature through a discussion of related work. In section 6, we conclude the paper and point challenges and future work.

2. BACKGROUND

There are three core materials used in our work: the sentence analyzers, the semantic-role-labeling method, and FrameNet itself. Boxer and spaCy are, respectively, the semantic and syntactic analyzers. Open-Sesame is the semantic-role-labeling method that supports the evaluation of our proposed method. FrameNet provides us with the annotated sentences that can support machine-reading and that we want to augment.

2.1. Boxer and Spacy: Semantic and Syntactic Analyzers

Boxer is an open-domain semantic analyzer [5] based on Combinatorial Categorical Grammars and Discourse Representation Theory. It generates a neo-Davidsonian representation of sentences. We also use it as a syntactic analyzer, the dependence tree parser, and the part-of-speech tagging system provided by the spaCy NLP library (version 2.0.11).

To process the different representations that we generate, we convert them all to a standard logical form. The Boxer analysis result is a bit tricky to normalize. Although it is already provided in first-order logic, we still need to do variable grounding, followed by Skolemization.

We also remove any negated terms and unbound variables left in order to have a simple graph structure. Figures 2 and 3 show examples of those analyzers in action.

Alice and **Bob** were **arrested**
yesterday in the beach.

pernambob(**b**),
pernamalice(**a**),
r1Time(**v**,y),
n1yesterday(y),
r1Theme(**v**,**s**),
v1arrest(**v**),
r1subset_of(**b**,**s**),
r1subset_of(**a**,**s**),
n1beach(p),
r1in(**v**,p).

Figure 2: Semantic Analysis by Boxer. Predicates (e.g., 'v1arrest') define the so-called thematic roles such as agent, theme, action etc., other semantic roles such as person name (pernam) and even nouns like beach. Every predicate (except for the person name one) is prefixed by its syntactic role as well.

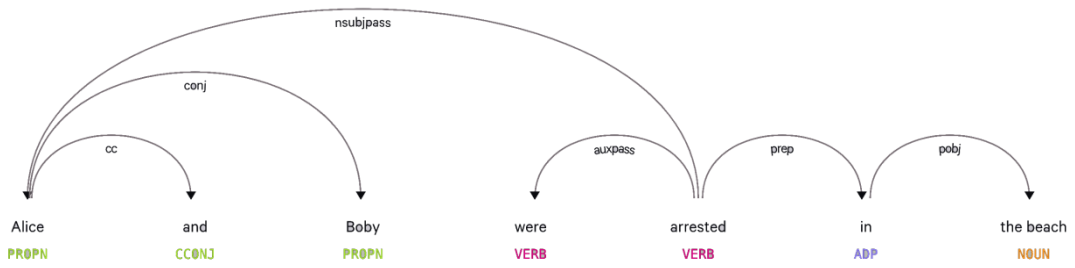


Figure 3: Syntactic Analysis by spaCy. The node labels (associated with the sentence tokens, e.g., 'VERB') give the part-of-speech tags, and the edge labels (associated with the tokens relationships, e.g., 'conj') are universal dependence labels.

2.2. Open-Sesame: the Supportive Semantic-Role-Labeling Method

Open-Sesame [6] is a state-of-the-art method for frame semantic parsing. This method is based on a segmental recurrent neural network [7], that supports its aim argument identification. It does not rely on syntactic representations during the testing phase, only during training. This way, this system presents itself as a cheaper alternative — regarding computational resources and human effort — to develop the syntactic parsers, while stays a competitive approach to the traditional pipeline that we follow in our work.

2.3. Framenet

We provide in this section a more detailed overview of FrameNet that suffices for the purpose of this paper. For a rigorous and comprehensive description of the FrameNet project, we refer the reader to Fillmore et al. [3]. In this work, we use FrameNet version 1.5.

FrameNet is an interconnected network of frames which provides the grounding for a cross-domain semantic representation. In this context, frames represent concepts like Arrest, Coming to Believe and Event. Those concepts also describe semantic roles that entities might have related to those concepts. For instance, some of the semantic roles described in the frame Arrest are Authority, Suspect and Place. Those semantic roles are called frame elements. Each frame

element occurring in a frame has its definition, written in a human-friendly form. Those definitions usually carry an example sentence where the frame elements are annotated as well as the frame itself. This way, we have both frame annotations, also called targets, and frame-element annotations together. For simplicity, we are going to refer to frame-element annotations just as annotations for the rest of the paper.

2.4. The Semantic Role Labeling (SRL) task from FrameNet's point of view

Here, we revisit the semantic role labeling (SRL) task, focusing on how FrameNet supports it as a resource. In doing so, we prepare for our specific research problem of augmenting FrameNet's semi-structured data in the next section.

In FrameNet, the sentences are annotated by humans. The general task of automatically generating those annotations is called frame-semantic parsing, which has SRL as one of its three components. Given a sentence, (i) target identification is the task of finding which token in the sentence should be matched to a frame; (ii) frame identification means to take a given token and assign it to a specific frame, and (iii) argument identification (SRL) is the task of matching frame elements that are members of the selected frames to the correct tokens in the sentence.

The SRL task induces our semi-structured data augmentation problem since SRL relies on a good set of annotated sentences as examples.

As discussed in the previous sections, FrameNet is a widely used resource supporting several NLP tasks. However, as a manually-built resource, it is error-prone and incomplete. For instance, fig. 7a shows that the frame coverage in FrameNet, that is, the number of frames that appear in at least one annotated sentence divided by the total number of frames, is only 70%.

In this work, we intend to increase this coverage so that NLP tasks in general — and SRL in particular — benefit from more frame annotations available. If we can achieve some increase in frame annotations coverage, even if it is not very large, it is bound to provide a relevant contribution to the machine reading community. That is because annotated sentences feed in all machine reading pipelines.

3. AUGMENTATION OF FRAMENET EXAMPLES

We start to state the data augmentation problem by introducing an example and follow it with our proposed methodology.

3.1. The Data Augmentation Problem

Consider the sentence "Most of us know where we took a photo but have a harder time remembering the time we took it.", and assume that Create physical artwork be one correct frame identified with this sentence. The annotation of this sentence concerning the structure of frame Create physical artwork is depicted in Fig. 4. There are three frame elements of that frame, namely, Creator, Representation, and Location of representation, which are mapped to subsets of tokens in the sentence.

From a general point of view, the data augmentation problem in this context is to ask how we could create a new annotation of this sentence using the tokens already mapped to frame elements of the frame Create physical artwork. The goal is to use the already marked tokens to annotate the sentence for another frame.

Figure 4: Create physical artwork annotation with respect to the frame Intentionally create.

Now consider Intentionally create, another frame which is related to Create physical artwork by the ‘has sub-frame of’ relation, as shown in Fig. 5. We exploit such inter-frame relations and then model the data augmentation problem accordingly. In our running example, the problem is reduced to whether or not we could build a new annotation of the sentence in terms of the structure of frame Create physical artwork. The new annotation must comprise not only the frame itself using the target token, but also its frame elements, namely, Creator, Created entity, and Place. It is quite intuitive that Creator from Create physical artwork should map to the frame element of same name from Intentionally create. The frame elements Created entity and Place from Create physical artwork should map to Created entity, and Place from Intentionally create, respectively.

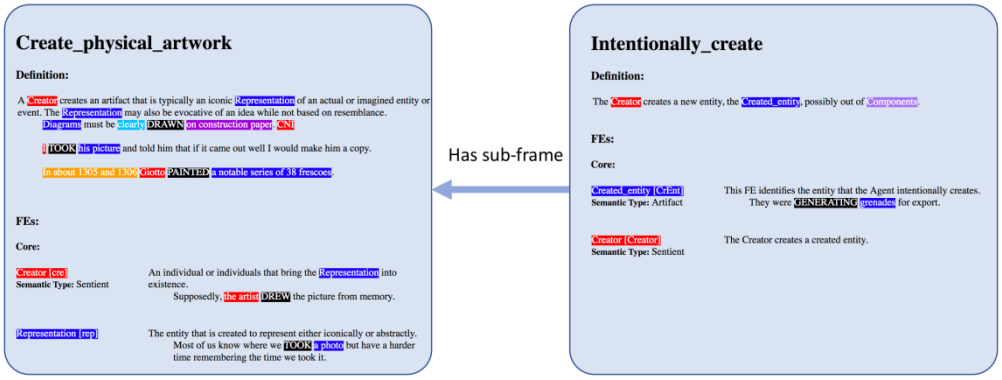


Figure 5: Intentionally create and Create physical artwork frames

3.2. The Notion of Frame Elements Equivalence

Frame elements equivalence is a rather vague concept. We model it in terms of three different notions of equivalence: lexical, semantic, and syntactic. We say that two frame elements from X and Y, respectively, are lexically equivalent if they have the same name. Two frame elements are said syntactically equivalent if there is at least one pair of examples from X and Y where these frame elements appear, and they have the same path of syntactic roles to the target in a syntactic representation. The semantic similarity follows the same concept of the syntactic equivalence, but, instead, we require a path of semantic roles turned into a semantic representation.

Consider the frames X and Y and an annotated sentence x with annotations of frame elements in X. Given that X is related to Y through one of the possible inter-frame relations (e.g., ‘is sub-frame of’), we want to find what annotations we could extend to Y. That is, we want to know if there can be a new annotation of the sentence regarding the frame elements belonging to Y. So, we will say that x is transferable from X to Y if all the frame element annotations in x are transferable to Y. Recall from section 2 that there are two kinds of annotations in an annotated sentence, namely: targets and frame element annotations. The second one we call annotations. An annotation is transferable from X to Y if its frame element is equivalent to one frame element in Y. This assured, we can rewrite the sentence annotation using frame elements of Y, and we can add a new annotation to the sentence.

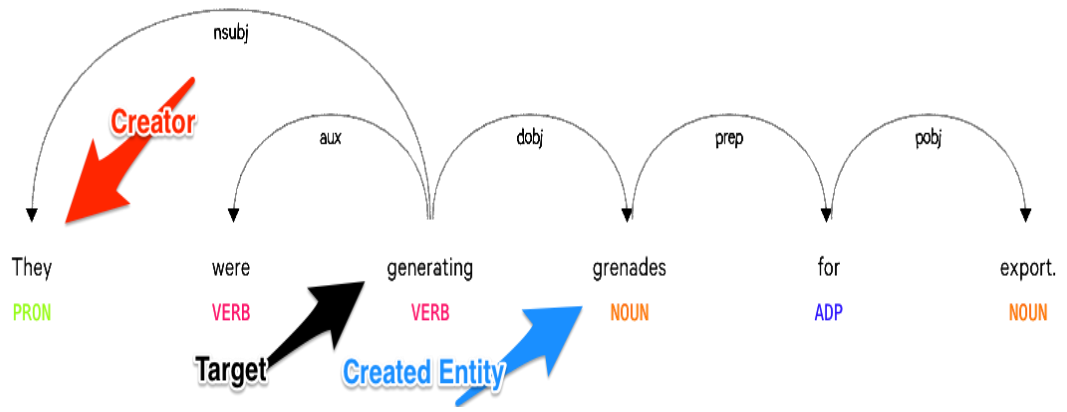
Let us recall the example depicted in figure 5. In order to know if this annotation can be adapted to another frame Create physical artwork, we first have to check if all frame elements of Intentionally create in the annotated sentence are equivalent to some frame element in Create physical artwork. Using the notion of lexical equivalence, we consider Creator to be the same as Creator in Create physical artwork as they both have the same name. Using the syntactic equivalence, we need to check if Created entity is equivalent to Representation. To do that, we take an example of Created entity from Intentionally create and one example of Representation from Create physical artwork and check if the syntactic path to the target is the same, as exhibited in figure 6. Since each frame element in the annotation is equivalent to some frame element in Create physical artwork, we can copy this example to Create physical artwork. If there were any frame elements left that have not an equivalent frame element in Create physical artwork, then the next step would be to check their semantic equivalence the same way we did for the syntactic equivalence.

The same method described before for expanding a frame example is used to expand annotated sentences from the FrameNet Project annotated documents. We show the results of this heuristic on whether we can borrow an annotated sentence in section 4.

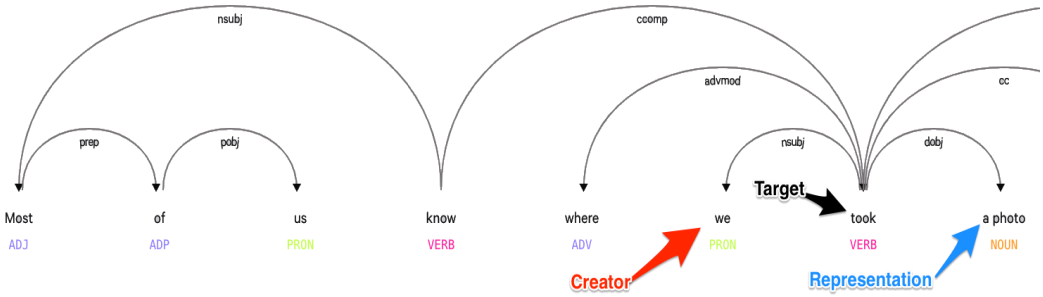
It is clear that 'ways for people with disability to enter the workforce' is not necessarily a piece of physical artwork as this augmented annotation suggests.

3.3. Frame Relations

To elaborate the proposed heuristics, we start by splitting the FrameNet inter-frame relations into two sets: (i) The set of hierarchical relations, depicted in the table 1, are the ones based in the inheritance and part-of concepts, and their reciprocal. (ii) The set of non-hierarchical relations comprises all the other relations and is depicted in table 2. This split is used to evaluate the effect of inheritance on the creation of new annotations. For instance, it is reasonable to think that annotations transferred from the frame Create physical artwork to its parent frame Intentionally create would be correct. Usually, the creation of an artwork is intentional, and all elements from the former frame have a corresponding element in the next frame.



Syntactic representation of example in Intentionally create



Syntactic representation of example in Create physical artwork

Figure 6: Syntactic representation of an example in the frame element descriptions

This way, when we say that the frame 'Coming to believe' inherits from 'Event', it means that 'Coming to Believe' is an 'Event'. And when we say that a 'Halt' is a subframe of 'Motion' it means that the concept 'halt' is part of the concept of 'motion'.

Table 1. Hierarchical relations

Relation	
Inherits from	is a frame of the same kind of the parent
Is Inherited by	the children frames have the same kind
Subframe of	is a part of the parent frame
Has Subframe(s)	is composed by those frames

Table 2. Non-hierarchical relations

Relation	
Perspective on	
Is Perspectivized in	
Uses	might be composed by those frames
Is Used by	might be part of the parent frame
Precedes	
Is Preceded by	
Is Inchoative of	the children are the cause of the root
Is Causative of	the root is the cause of the children
See also	Informational relation.

4. EXPERIMENTS

The purpose of the augmentation method we propose here is to increase the number of available training examples and expand the coverage over less popular frames. This augmentation is particularly useful once we consider the difficulty in manually expanding the FrameNet example set and also the difficulty of adding new documents.

4.1. Data

Our dataset consists of annotated sentences from the collection of annotated documents made available in FrameNet release 1.5. This collection consists of 78 documents annotated by FrameNet's staff; we use the same test set as [6, 8]. Those documents hold together almost 5946 annotated sentences. In those annotated sentences is a total of 23944 frame annotations and

48133 frame element annotations related to those frame annotations. The prefix, that is, the part of the document name before ‘ ’ refers to the source of the document, and the suffix is the document name. In total, there are more than 130000 sentences in the FrameNet project with some kind of annotation. More on the construction of this dataset and FrameNet, in general, is found in [9].

4.2. Evaluation Setting

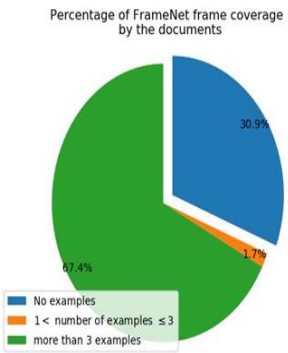
We evaluate the augmentation strategies based on the improvement of the performance of a state-of-the art method in the literature, Open-Sesame. Each one of the multiple training instances is carried out until the same termination criterion is reached, for conformity and ease of comparison, the criterion is the same used in the Open-Sesame paper, we also used the default parameters reported in that paper [6]. This criterion is met when there where no updates in the best loss score reported after 28 validation epochs.

We used the same GloVe embedding [10] and optimized the model using ADAM [11], with a learning rate of 0.0005, and moving average parameter of 0.01. We also set the moving average variance to 0.9999, and we set the parameter (to prevent numerical instability) to 10−8; no learning rate decay is used, as done in the original Open-Sesame paper.

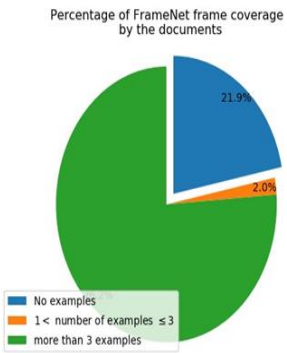
4.3. Results

We evaluated three kinds of augmentation in this project, namely lexical, syntactic, and semantic analysis (described in section 3). The overall gain on number of annotations from each one of those strategies is depicted in figures 7b, 7c, and 7d, respectively. We see a moderate increase of over roughly 13% of the original coverage using the different kinds of augmentations separately depicted in figure 8. This gain indicates that besides the noise addition, the augmentation strategy was beneficial to the semantic-role-labeling task.

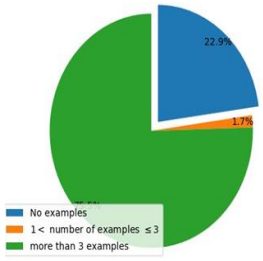
The impact of the augmentation method on the performance of the SRL parser is expressed in table 3. Values in bold are the best values reported. We report precision, recall, and f1-score metrics micro-averaged. Our experimentation shows a small improvement in Open-Sesame’s performance when trained on datasets that undertook the augmentation strategies developed here. This improvement indicates that even with added noise, the use of the augmentation benefited the semantic parser. The annotations from the semantic and syntactic augmentation strategies did not perform better than the lexical strategy. Errors in the logical representations might cause it due to incorrect parsing of the sentences.



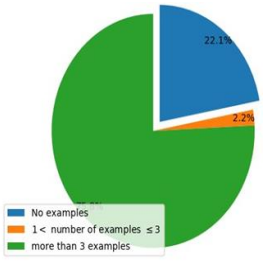
(a) No augmentation



(b) Semantic augmentation



(c) Lexical Augmentation



(d) Syntactic augmentation

Figure 7: Augmentation frame coverage

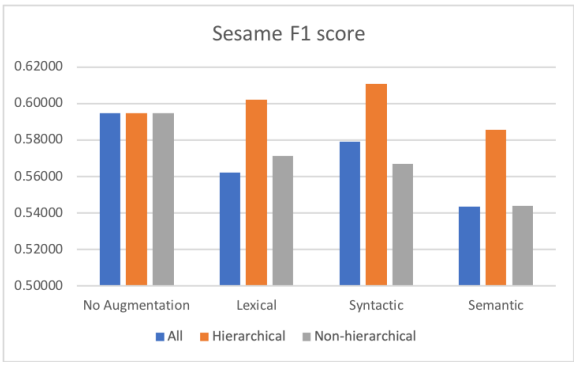


Figure 8: Comparison of Sesame F1 Score

Table 3. Performance of Sesame with the different augmentations

		Precision	Recall	F-1
Semantic	All	0.5946	0.5497	0.5712
	Hierarchical	0.5880	0.5060	0.5439
	Non-hierarchical	0.5975	0.5397	0.5671
Syntactic	All	0.5939	0.5337	0.5622
	Hierarchical	0.6041	0.4939	0.5434
	Non-hierarchical	0.6001	0.5595	0.5791
Lexical	All	0.6083	0.5955	0.6018
	Hierarchical	0.6136	0.5598	0.5854
	Non-hierarchical	0.6374	0.5865	0.6109
	No augmentation	0.5977	0.6030	0.6004

5. RELATED WORK

We considered the three main areas that we have built our contribution upon on, namely: Language resources augmentation, Sentence Representation, and Semantic Role Labeling.

5.1. Language Resources Augmentation

To the best of our knowledge, this is the first work that builds a data augmentation strategy relying only upon the data provided by FrameNet. Other venues of work combine additional language resources with FrameNet to produce SRL parsers. Shi and Mihalcea [12], Giuglea and Moschitti [13], Palmer [14], Laparra and Rigau [15], Tonelli et al. [16], and Green et al. [17] are

examples of work that combine other language resources, such as PropBank [18], VerbNet [19], and WordNet [20] with FrameNet Baker et al. [3], to complement each other or even to generate more frames. It is also possible to combine more than one of those resources; for example, the Predicate Matrix [21] is a new language resource created through the automatic combination of WordNet, Framenet, and Verbnet. Pavlick et al. [22] presents a FrameNet augmentation based on expanding the resources Lexical Units, LUs. They based their augmentation method on automatic paraphrasing using the Paraphrase Database (PPDB) [23] curated by manual crowd sourcing.
The model proposed by Mousselly Sergieh and Gurevych [24] is based on word embedding to identify a mapping between Wikidata relations [25] and FrameNet frames and to annotate the arguments of each relationship with the semantic roles from the second resource. This is an example of a case where FrameNet is used to enrich other resources and is a clear contrast with our work that aims to enhance FrameNet without the use of external corpora, but only on parsing methods. This choice makes this approach flexible and agnostic of external data sources used to train those parsers.

5.2. Logical Form and Sentence Representation

Textual data is found in unstructured ways, as mentioned throughout this paper, and we want to make it as structured as possible, so it is machine-processable. Logical forms can be used to express both the syntactic and semantic aspects of the sentences of a textual document, and much work has been done on building such logical forms.

A usual step is to parse a sentence into a syntactic representation and use this intermediary representation to generate a semantic representation of the meaning covered in the sentence. In particular, [26] devise a system based on the lambda calculus for deriving neo-Davidsonian logical forms from dependency trees. They evaluate the quality of such logical forms derived from the dependency trees of the sentences by feeding those logical forms to a semantic parser. This semantic parser consists of a graph matching algorithm that matches the structure of the logical form to Freebase, a collaboratively created tuple-based knowledge base that later on was used to power Google's Knowledge Graph initiative, [27]. It generates a robust representation of the sentences and can be compared with our current approach in future work. Using this approach as our semantic parser would be a promising comparison since one of their claims is that this representation outperforms a CCG-based representation which composes the Boxer method, used in our work.

Similarly, to our work, [26] creates a new neo-Davidsonian representation of sentences that might improve our current method. [28] combine logical and distributional representations. They use similarity metrics to create weighted rules using Markov Logic Networks [29]. Beltagy et al. [28] show that besides estimating the similarity between sentences, this method can also recognize textual entailment. One can use this textual entailment as another feature for our augmentation purposes.

In the same way, we rely on Boxer to obtain a logic-based parsed output. Previous work has already started from this tool to extract and represent meaning in a structured, machine-processable format from text documents. In particular, [28, 30] combined the parsed logical representation with distributional semantics and Markov Logic Networks. The distributional semantics is used to construct a unified knowledge base from different sources, while MLN is used to perform inference. The neo-Davidsonian representation and MLN are also employed to solve the Science and Math challenge, an NLP competition that aims to produce systems that can answer fifth-grade science exams, as done in [31].

The difficulties of directly applying those methods without any tinkering to our problem are that we calculate if substructures in the sentence are similar, focusing on specific terms. It is not clear

how to apply this concept to most of those methods since they are not concerned with specific terms of the sentence, but the sentence as a whole.

5.3. Semantic Role Labeling

The Semantic Role Labeling, SRL, is the problem of finding semantic roles to entities located in textual documents. SRL is a fruitful area of research containing work that takes advantage of multiple language resources, including FrameNet. The most recent and state-of-the-art approaches are mostly based on statistical methods, in particular, machine learning methods.
The model presented in [4] uses latent variables and semi-supervised learning to improve frame disambiguation for targets unseen at training time. On the other hand, the work shown in [32] consists of a frame identification that is coupled into an argument parsing method to perform FSP. Sling, [33], is a framework for frame-semantic parsing that performs neural-network parsing with bidirectional LSTM input encoding and a transition based recurrent unit. It takes as input only the tokens of the sentence, skipping any previous syntactic or semantic parser. Both methods are machine-learning based.

The semantic parser developed in [13] connects VerbNet and FrameNet by mapping the FrameNet frames to the VerbNet Intersective Levin classes. To further increase the verb coverage, they use the lexicon contained in PropBank and the PropBank semantic annotations to evaluate their system.

6. CONCLUSION

Semantic Role Labeling (SRL) is an essential task towards creating a machine-meaningful representation of textual information. FrameNet is the main supportive resource for this task. However, as a manually-built resource, it is error-prone and incomplete. A large group of frames lacks useful annotations. In this work, we present a data augmentation method for FrameNet documents that increases by over 13% the total number of annotations. As a result, a new dataset is now available for SRL and frame semantic parsing in general. We also show that the annotations generated can improve the performance of a semantic-role-labeling method.

The augmentation methods present in the literature are usually methods for combining FrameNet with other linguistic resources. This work presents an approach to augment the data available in FrameNet using sentence examples in the resource's element descriptions themselves. This way, one can apply our method after (or before) applying some other method present in the literature for a more incisive expansion without necessarily adding redundant information.

A first line of future research is to investigate the impact of this data augmentation in combination with other methods present in the literature. Another possible investigation venture is the exploration of the inter-frame relationships. We suspect that it is possible to further explore the connections amongst frames to infer new relationships amongst frame elements. We also intend to test the method on other electronic (linguistic) resources. For example, WordNet seems a relatively close opportunity for short- to mid-term research.

Semantic Role Labeling (SRL) is an essential task towards creating a machine-meaningful representation of textual information. FrameNet is the primary supportive resource for this task. However, as a manually-built resource, it is error-prone and incomplete. A large group of frames lacks useful annotations. In this work, we present a data augmentation method for FrameNet documents that increases by over 13% the total number of annotations. As a result, a new dataset is now available for SRL and frame semantic parsing in general. We also show that the annotations generated can improve the performance of a semantic-role-labeling method.

The augmentation methods present in the literature are usually methods for combining FrameNet with other linguistic resources. This work presents an approach to augment the data available in FrameNet using sentence examples in the resource's element descriptions themselves. This way, one can apply our method after (or before) applying some other method present in the literature for a more incisive expansion without necessarily adding redundant information.

The first line of future research is to investigate the impact of this data augmentation in combination with other methods present in the literature. Another possible investigation venture is the exploration of inter-frame relationships. We suspect that it is possible to explore the connections amongst frames further to infer new relationships amongst frame elements. We also intend to test the method on other electronic (linguistic) resources. For example, WordNet seems a relatively close opportunity for short- to mid-term research.

REFERENCES

[1] O. Etzioni, M. Banko, M. J. Cafarella, Machine reading, in: Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, AAAI Press, 2006, pp. 1517–1519.

[2] O. Abend, A. Rappoport, The State of the Art in Semantic Representations, Acl 35 (2017) 23–24.

[3] C. F. Baker, C. J. Fillmore, J. B. Lowe, The Berkeley FrameNet Project, in: Proceedings of the 17th International Conference on Computational Linguistics - Volume 1, COLING '98, Association for Computational Linguistics, Stroudsburg, PA, USA, 1998, pp. 86–90. URL: https://doi.org/10.3115/980451.980860. doi:10.3115/980451.980860.

[4] D. Das, D. Chen, A. F. T. Martins, N. Schneider, N. Noah A. Smith, Frame-Semantic Parsing, Computational linguistics 40 (2014) 9 –56.

[5] J. Bos, Wide-coverage semantic analysis with Boxer, in: Proceedings of the 2008 Conference on Semantics in Text Processing, c, Association for Computational Linguistics, Venice, Italy, 2008, pp. 277–286. doi:10.3115/1626481.1626503.

[6] S. Swayamdipta, S. Thomson, C. Dyer, N. A. Smith, Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold, arXiv preprint arXiv:1706.09528 (2017).

[7] L. Kong, C. Dyer, N. A. Smith, Segmental Recurrent Neural Networks, arXiv preprint arXiv:1511.06018 (2015) 1–10. URL: http://arxiv.org/abs/1511.06018. doi:10.21437/Interspeech.2016-40.

[8] D. Das, N. Schneider, D. Chen, N. A. Smith, Probabilistic Frame-Semantic Parsing, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL) 3 (2010) 948–956.

[9] C. F. Baker, C. J. Fillmore, B. Cronin, The Structure of the FrameNet Database, International Journal of Lexicography 16 (2003) 281—296.

[10] J. Pennington, R. Socher, C. Manning, Glove: Global Vectors for Word Representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014) 1532–1543. URL: http://aclweb.org/anthology/D14-1162. doi:10.3115/ v1/D14-1162.

[11] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization (2014). URL: http://arxiv.org/abs/1412.6980.

[12] L. Shi, R. Mihalcea, Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing, Computational Linguistics and Intelligent Text Processing 34 (2005) 100–111. URL: http://link.springer.com/10.1007/978-3-540-30586-6_9. doi:10.1007/978-3-540-30586-6_9.

[13] A.-M. Giuglea, A. Moschitti, Semantic Role Labeling via FrameNet, VerbNet and PropBank, in: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, July, Association for Computational Linguistics, Sydney, Australia, 2006, pp. 929–936. doi:10.3115/1220175.1220292.

[14] M. Palmer, SemLink-Linking PropBank, VerbNet, FrameNet, Technical Report, 2009. URL: http://www.flarenet.eu/sites/default/files/S3_01_Palmer.pdf.

[15] E. Laparra, G. Rigau, Integrating WordNet and FrameNet using a Knowledge-based Word Sense Disambiguation Algorithm, Proceedings of the International Conference RANLP-2009 (2009) 208–213. URL: http://www.aclweb.org/anthology/R09-1039.

[16] S. Tonelli, C. Giuliano, K. Tymoshenko, Wikipedia-based WSD for multilingual frame annotation, Artificial Intelligence 194 (2013) 203–221. URL: http://dx.doi.org/10.1016/j. artint.2012.06.002. doi:10.1016/j.artint.2012.06.002.
[17] R. Green, B. J. Dorr, P. Resnik, Inducing frame semantic verb classes from WordNet and LDOCE, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL'04 (2004) 375–es. URL: http://portal.acm.org/citation.cfm?doid=1218955.1219003. doi:10.3115/1218955.1219003.
[18] P. Kingsbury, M. Palmer, From Treebank to PropBank, LREC (2002) 1989–1993. doi:10.1007/ s13398-014-0173-7.2.
[19] K. Kipper, A. Korhonen, N. Ryant, M. Palmer, A large-scale classification of English verbs, Language Resources and Evaluation 42 (2008) 21–40. doi:10.1007/s10579-007-9048-2.
[20] C. F. Baker, C. Fellbaum, Wordnet and framenet as complementary resources for annotation, in: Proceedings of the Third Linguistic Annotation Workshop, Association for Computational Linguistics, 2009, pp. 125–129.
[21] M. Lopez De Lacalle, E. Laparra, I. Aldabe, G. Rigau, Predicate Matrix: automatically extending the semantic interoperability between predicate resources, Language Resources and Evaluation 50 (2016) 263–289. URL: http://adimen.si.ehu.es/web/PredicateMatrix. doi:10.1007/s10579-016-9348-5.
[22] E. Pavlick, T. Wolfe, P. Rastogi, C. Callison-Burch, M. Dredze, B. Van Durme, FrameNet+: Fast paraphrastic tripling of framenet, ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference 2 (2015) 408–413.
[23] J. Ganitkevitch, B. V. Durme, C. Callison-Burch, PPDB: The Paraphrase Database, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 758–764. URL: https://aclanthology.info/papers/N13-1092/ n13-1092.
[24] H. Mousselly Sergieh, I. Gurevych, Enriching Wikidata with Frame Semantics, in: Proceedings of the 5th Workshop on Automated Knowledge Base Construction, 3, Association for Computational Linguistics, San Diego, CA, 2016, pp. 29–34. URL: http://aclweb.org/anthology/ W16-1306. doi:10.18653/v1/W16-1306.
[25] D. Vrandecic, M. Krotzsch, Wikidata: A Free Collaborative Knowledgebase, Commun. ACM 57 (2014) 78–85. doi:10.1145/2629489.
[26] S. Reddy, O. Tackstr¨ om, M. Collins, T. Kwiatkowski, D. Das, M. Steedman, M. Lapata, Trans-¨ forming Dependency Structures to Logical Forms for Semantic Parsing, Transactions of the ACL 4 (2016) 127–140.
[27] A. Singhal, Introducing the Knowledge Graph: things, not strings, 2012. URL: http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.
[28] I. Beltagy, C. Chau, G. Boleda, D. Garrette, K. Erk, R. J. Mooney, Montague meets markov: Deep semantics with probabilistic logical form, in: Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, ACL, 2013, pp. 11–21.
[29] M. Richardson, P. Domingos, M. Richardson, P. Domingos, Markov logic networks, Machine Learning 62 (2006) 107–136. doi:10.1007/s10994-006-5833-1.
[30] I. Beltagy, S. Roller, P. Cheng, K. Erk, R. J. Mooney, Representing meaning with a combination of logical and distributional models, Computational Linguistics 42 (2016) 763–808.
[31] T. Khot, N. Balasubramanian, E. Gribkoff, A. Sabharwal, P. Clark, O. Etzioni, Exploring markov logic networks for question answering, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, ACL, 2015, pp. 685–694.
[32] K. M. Hermann, D. Das, J. Weston, K. Ganchev, Semantic Frame Identification with Distributed Word Representations, Proceedings of ACL (2014) 1448–1458. URL: http://www.aclweb. org/anthology/P14-1136. doi:10.3115/v1/P14-1136.
[33] M. Ringgaard, R. Gupta, F. C. Pereira, Sling: A framework for frame semantic parsing, arXiv preprint arXiv:1710.07032 (2017).

UNSUPERVISED CLUSTERING FOR DISTORTED IMAGE WITH DENOISING FEATURE LEARNING

Qihao Lin, Jinyu Cai and Genggeng Liu

College of Mathematics and Computer Science, Fuzhou University,
Fuzhou, 350116, China

ABSTRACT

High-dimensional of image data is an obstacle for clustering. One of methods to solve it is feature representation learning. However, if the image is distorted or suffers from the influence of noise, the extraction of effective features may be difficult. In this paper, an end-to-end feature learning model is proposed to extract denoising low-dimensional representations from distorted images, and these denoising features are evaluated by comparing with several feature representation methods in clustering task. First, some related works about classical feature learning are introduced. Then the architecture and working mechanism of denoising feature learning model are presented. As the structural characteristics of this model, it can obtain essential information from image to decrease reconstruction error. When facing with corrupted data, it also runs a robust clustering result. Finally, compared to other unsupervised feature learning methods, extensive experiments demonstrate that the obtained feature representations by proposed model run a competitive clustering performance. The low-dimensional representations can replace the original datasets primely.

KEYWORDS

Unsupervised Learning, Feature Representation, Auto-encoder, Clustering

1. INTRODUCTION

In machine learning and data analysis, difficulties in information processing are caused by large dimensions. Meanwhile, if image datasets are distorted or suffer from noise, the extraction of effective features becomes more difficult. Consequently, learning reusable feature representations from a large number of unlabelled datasets has become a research hotspot. High-dimensional images always need a pre-processing such as dimensionality reduction [1]. Feature representation learning is an effective method [2]. Feature learning can be categorized into supervised based and unsupervised based. Supervised based methods have reached remarkable performance. Linear discriminant analysis (LDA) [3] makes the distance between different types of samples larger, and the distance between similar samples smaller. Locality sensitive discriminant analysis (LSDA) [4] belongs to manifold learning algorithm. Its main idea is to maximize the edge of different classes in each local region. However, as the increasing of data and unmarked label, supervised based methods may have an impact on its accuracy.

The emergence of unsupervised feature learning is better solved 'curse of dimensionality' as well as unmarked labels. Unsupervised feature learning is classified into two parts: linear based and non-linear based. Principal component analysis (PCA) is a statistical method [5]. It uses orthogonal transformation to convert a set of variables that may be related into a set of linearly

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 15-27, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101202

uncorrelated variables. The converted set of variables is called the principal component. Locality preserving projections (LPP) builds a graph on the data set. This graph contains the information of the node's neighbours [6]. The algorithm mainly finds an optimal linear approximation when the high-dimensional data depends on the embedding of the low-dimensional manifold in the surrounding space. PCA and LPP are two linear feature learning algorithms. Neighbourhood preserving embedding selects neighbours to reconstruct linear weights for each point. The core of isometric feature mapping (Isomap) is to find and utilize the characteristics of manifold space, introduce geodesic distance and propose corresponding distance calculation [7]. Locally linear embedding (LLE) is a new feature learning algorithm for non-linear data [8]. It can keep the original manifold structure after dimensionality reduction as far as possible. Isometric projection (IsoP) discovers the in-trinsic geometrical structure of data set [9].

In recent years, auto-encoder (AE) and its family are proposed to realize dimensionality reduction and feature learning. Auto-encoder is an unsupervised learning algorithm (the training example is not marked), which uses back propagation algorithm and makes the target value equal to the input value [10]. It is a neural network which contains three layers. The dimension of hidden layer is much smaller than input layer. Sparse auto-encoder (SAE) limits the number of hidden units to learn more useful features [11]. A neuron is active if its output value is close to 1, otherwise it is not active if its output value is close to 0. Variational auto-encoder (VAE) is an important generation model. It proposes a gradient estimation called stochastic gradient variable bayes [12]. The core of adversarial auto-encoder (AAE) is to use a generator and a discriminator for adversary learning [13]. It's a combination of VAE and adversarial network.

The above methods run a great performance on feature extraction. However, when facing with distorted images, existing unsupervised feature learning methods may be affected. In this paper, an end-to-end feature learning model is proposed to extract denoising low-dimensional representations from distorted image datasets. These denoising features perform well in unsupervised clustering task. As the structural characteristics of this model, it can obtain essential information from image to decrease reconstruction error. Facing corrupted data, it also runs a better result. For evaluating their performance, these features are sent into k-means clustering [14]. Three evaluation metrics are selected for comparison including clustering accuracy (ACC), normalized mutual information (NMI) and adjusted rand index (ARI).

The following parts of this paper are arranged as follows. Some works related to classical feature learning are showed in Section 2. In Section 3, the structure and working mechanism of denoising feature learning model are presented. In Section 4, extensive experiments on eight standard datasets illustrate the effectiveness of presented model. Eventually, this paper is concluded in Section 5.

2. RELATED WORKS

In this section, we introduce several classical feature learning algorithms which are classified into two kinds: unsupervised feature learning and supervised feature learning.

2.1. Unsupervised Feature Learning

The unsupervised feature learning algorithms are categorized into two types: linear and non-linear. The core of linear feature learning algorithms is to obtain a linear mapping relation. Principal component analysis and isometric projection are commonly used linear feature learning algorithms. As for non-linear datasets, linear feature learning algorithms probably meet some

problems. Neighbourhood preserving embedding [15] and isometric feature mapping are two famous non-linear feature learning algorithms.

2.1.1. Isometric Projection

Isometric projection is a linear feature learning algorithm. Nevertheless, isometric projection can handle more complex datasets such as manifold data [16] which is embedded in high-dimensional space. Given a dataset $X \in \mathbb{R}d\times n$, isometric projection finds a mapping function f that makes $yi = f(xi)$ where $\{yi\ \}i=1\ n \in \mathbb{R}k$. Isometric projection defines dM the geodesic distance [17] measure on M which is a non-linear manifold embedded in $\mathbb{R}d$ and dE the standard Euclidean distance. Then optimization objective function is formalized as follows

$$\arg\min_{f} \sum_{i,j}(d_M(x_i,\ x_j) - d_E(f(x_i),\ f(x_j)))^2 \qquad (1)$$

where the mapping function f is to let Euclidean distances can offer an effective approximation to the geodesic distances on M.

2.1.2. Isometric Feature Mapping

Isometric feature mapping is a kind of manifold learning [18] method which is used in feature learning of non-linear data. Isomap algorithm has three steps. First step confirms neighbourhood for each point. There are two ways: k nearest neighbours (k-Isomap) and all points in radius ϵ (ϵIsomap). $d(xi\ ,\ xj)$ represents distance in input space, such that we obtain a weighted graph G. Second, if xi and xj are linked by an edge, initialize shortest path distances $d_G(x_i,\ x_j) = d(x_i,\ x_j)$ or else $d_G\ (x_i,\ x_j) = \infty$. Then $d_G(x_i,\ x_j)$ is constantly replaced by $\min\{d_G(x_i,\ x_j),\ d_G(x_i,\ x_p) + d_G(x_p,\ x_j)\}, p = 1, 2, \dots, N$. N is the number of whole points. Afterwards Isomap creates a matrix DG that consists of the shortest path distances. In the finally step, MDS is applied in DG . Consider the k -dimensional Euclidean space Y that preserves most information of manifolds intrinsic geometry, DY matrix is composed of Euclidean distances $\{d_Y(i,\ j) = \|y_i - y_j\|\}$. Then the cost function is denoted as

$$E = \|\gamma(DG\ ) - \gamma(DY)\|L\ 2 \qquad (2)$$

Where γ indicates an operator that converts distances to inner products.

2.1.3. Principal Component

Analysis Principal component analysis is a statistical method. It uses orthogonal transformation to convert a set of variables that may be related into a set of linearly uncorrelated variables. The converted set of variables is named the principal component. Consider an input image dataset $X = [x_1, \dots\ ,\ x_n]$, we obtain its normalized matrix X'. Covariance matrix C can be presented as follows

$$C = \frac{1}{n} X' X'^T \qquad (3)$$

Then we calculate the eigen values and eigenvectors about covariance matrix C. After that, k eigenvectors corresponding to k largest eigen values are selected. These eigenvectors are utilized to construct the projection matrix W which is ordered by eigen values descend. Finally, low dimensional feature representations $Y \in \mathbb{R}k\times n$ are formalized by: $Y = WX$.

2.2. Supervised Feature Learning

Supervised feature learning algorithms require sufficient labels, nevertheless, they perform a commendable result. In some cases, supervised feature learning algorithms can be improved into semi-supervised and then significantly reduce the need for labels. In this section, we mainly introduce supervised feature learning algorithms linear discriminant analysis, locality sensitive discriminant analysis and semi-supervised algorithm stacked label consistent auto-encoder (SLCA).

2.2.1. Linear Discriminant

Analysis As a classical algorithm in pattern recognition, the basic idea of linear discriminant analysis is to project high-dimensional pattern samples into the optimal discriminant vector space. Given a dataset$\{x_1, \dots, x_n\} \in \mathbb{R}^d$,$\{y_1, \dots, y_n\} \in \mathbb{R}^k$ $(k \ll d)$, attempt to find mapping matrix $A = (a_1, \dots, a_k) \in \mathbb{R}^{d \times k}$ such that $y_i = A^T x_i$. Suppose all samples are sorted into c classes. The objective function is denoted as follows

$$a_{opt} = \arg\max_a \frac{a^T S_b a}{a^T S_w a} \qquad (4)$$

$$S_b = \sum_{i=1}^{c} m_i (u_i - u)(u_i - u)^T \qquad (5)$$

$$S_w = \sum_{i=1}^{c} (\sum_{j=1}^{m_i} (x_j^i - u_i)(x_j^i - u_i)^T) \qquad (6)$$

where Sw is within-class scatter matrix while Sb is between-class scatter matrix. u means the total sample mean vector and mi is the number of data points in i-th class. ui represents the average vector of i-th class. The eigenvectors related to the largest eigen values constitute the basic functions of LDA:

$$S_b a = \lambda S_w a \qquad (7)$$

the aim of LDA is to preserve global class relationship between sample points. And as a classification, it is hoped that the coupling degree between classes is low and the degree of aggregation within classes is high.

2.2.2. Locality Sensitive Discriminant

Analysis Locality sensitive discriminant analysis is a popular data-analytic tool which can discover the local manifold structure. Local structure is more important if lacking of sufficient training samples. LSDA defines a projection by finding the local manifold structure and the projection maximizes the margin between sample points. Given n data points $\{x_1, \dots, x_n\} \in \mathbb{R}^d$, denote $N(x_i) = \{x_i^1, \dots, x_i^k\}$ the k nearest neighbors of x_i and $l(x_i)$ the class label of x_i.For each data point, $N(x_i)$ is divided into two subsets $N_w(x_i)$ and $N_b(x_i)$. $N_w(x_i)$ indicates the neighbours sharing the same label while $N_b(x_i)$ means the neighbours owning different labels

$$N_w(x_i) = \{x_i^j | l(x_i^j) = l(x_i), 1 \le j \le k\}$$

$$N_b(x_i) = \{x_i^j | l(x_i^j) \ne l(x_i), 1 \le j \le k\} \qquad (8)$$

It's obvious that $N_w(x_i) \cap N_b(x_i) = \emptyset$ and $N_w(x_i) \cup N_b(x_i) = N(x_i)$.Then the weight matrices are defined as $W_{w,ij} = 1$ if $x_i \in N_b(x_j)$ or $x_j \in N_b(x_i)$. Let y $= (y_1, \dots, y_m)^T$ be a map, theobjective functions are formalized as

$$\min_W \sum_{ij} (y_i - y_j)^2 W_{w,ij} \qquad (9)$$

$$\max_W \sum_{ij} (y_i - y_j)^2 W_{b,ij} \qquad (10)$$

The objective function (9) attempts to ensure that y_i and y_j are close while x_i and x_j are close and own same label.Maximizing (10) is to ensure that y_i and y_j are far apart if x_i and x_j are close and have different labels.

2.2.3. Stacked Label Consistent Auto-encoder

Stacked label consistent auto-encoder is a semi-supervised method which combines reconstruction and classification [19]. Its architecture is consist of two-layer stacked auto-encoder. Stacked label consistent auto-encoder aims to create a linear map between innermost layer and class labels which constitutes the class label consistency penalty. The optimization objective function is presented as

$$\min_{W_1, W_2, W_1', W_2', D} \|X - W_1' \emptyset(W_2' \emptyset(W_2 \emptyset(W_1 X)))\|_F^2 + \lambda \|L - D\emptyset(W_2 \emptyset(W_1 X))\|_F^2 \qquad (11)$$

Here, X is the input data matrix, D the linear map and L the class labels. Wi and $Wi\ '$ represent the weight between layers. Existing backpropagation techniques can't learn this architecture because there are two outputs. Stacked label consistent auto-encoder solves this problem by the Split Bregman technique. Formulation (11) requires all input samples have corresponding class labels. However, it is difficult to gain all labels and semi-supervision method is allowed. This leads to

$$\min_{W_1, W_2, W_1', W_2', D} \|X - W_1' \emptyset(W_2' \emptyset(W_2 \emptyset(W_1 X)))\|_F^2 + \lambda \|L - D\emptyset(W_2 \emptyset(W_1 X_S))\|_F^2 \qquad (12)$$

where the training data $X = [X_U | X_S]$and the subscripts denote unsupervised or supervised.

3. UNSUPERVISED DENOISING FEATURE LEARNING FOR DISTORTED IMAGE

Facing with distorted images, existing unsupervised feature learning methods may be not robust. For solving unsupervised clustering task of distorted images, an end-to-end feature learning model is presented to extract denoising low-dimensional representations. As the model is based on auto-encoder, next we introduce the structure of auto-encoder. Auto-encoder is a neural network which uses back propagation. Consider an input image $X \in \mathbb{R}^{m \times n}$. Auto-encoder aims to reconstruct a matrix X' which is similar to input data. In this process, auto-encoder network makes a hidden representation $Y \in \mathbb{R}^{d \times n}$ from X ($d \ll$ m).

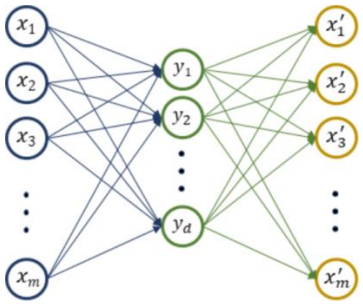


Figure 1. The structure of auto-encoder neural network

As demonstrated in Figure 1, the structure of auto-encoder is split into three parts: input layer, hidden layer and output layer. The process of encoder is denoted as $y = f(x)$and $x' = g(y)$means decoder. The optimization objective function of auto-encoder is represented as

$$\min_{W,b} \Theta(W,b) = \sum_{i=1}^{n} \|x_i - g(f(x_i))\|_2^2 \qquad (13)$$

where W and b mean the weight and the bias of neural network. Predefined activation function usually uses sigmoid function $S(x) = \frac{1}{1+e^{-x}}$.

If input datasets are distorted or suffer from noise, the features obtained by auto-encoder may be affected. Denoising feature learning model aims to enhance the robustness of feature. It is capable of reconstructing clean data from distorted data. Sometimes the reconstructed images could obtain a better performance than original images. Meanwhile it reduces the risk of over-fitting.

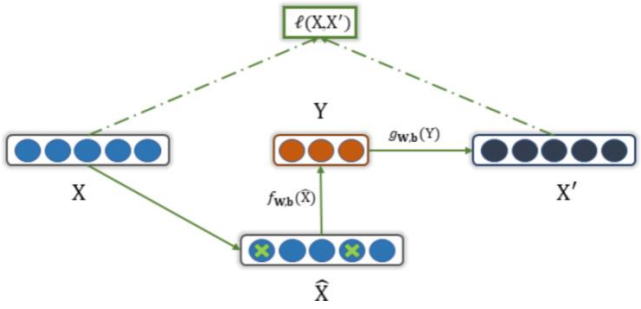


Figure 2. The framework of proposed model

The structure of proposed model is showed in Figure 2. Model accepts distorted data as input and output a clean data. Consider an original image dataset $X = [x_1, \dots, x_n] \in \mathbb{R}^{m \times n}$. As the influence of distortion or noise, corrupted dataset $\widehat{X}$ is formed. In the training process, corrupted dataset can besimulated by adding random zero into input data. Denoising low-dimensional representation is denoted as $Y \in \mathbb{R}^{d \times n}$.Finally, the reconstructed data X' can be presented as

$$Y = f_{W,b}(\hat{X}) = \alpha\left(W^{(1)}\hat{X} + b^{(1)}\right) \qquad (14)$$

$$X' = g_{W,b}(Y) = \alpha\left(W^{(2)}Y + b^{(2)}\right) = \alpha\left(W^{(2)}\alpha\left(W^{(1)}\hat{X} + b^{(1)}\right) + b^{(2)}\right) \qquad (15)$$

where α means activation function. Ordinarily activation function is using logistic sigmoid function $\alpha(x) = \frac{1}{1+e^{-x}}$. Each layer is defined to share the same network parameters. As a consequence, the weight $W^{(1)} = W^{(2)} = W$ and bias $b = [b^{(1)}; b^{(2)}]$. W is a$d \times m$ matrix and b is an-dimensional vector. The optimization objective function of model is presented as

$$\min_{W,b} \Theta(W, b) = L(X, X') = \sum_{i=1}^{n} L(x_i, x_i') = \sum_{i=1}^{n} \|x_i - g(f(\hat{x}_i))\|_2^2 \qquad (16)$$

where L is a cross-entropy loss function. The parameters of model network are denoted as $\theta = \{W, b\}$.They are constantly renovated via iterative descent of L. The detailed steps of denoising feature learning model are summarized as Algorithm 1.

4. EXPERIMENTAL ANALYSIS

In experiment stage, first we introduce eight public image databases. Then three popular clustering evaluation metrics and their working principles are demonstrated in the second section. Except presented model, seven common dimensional reduction algorithms are used to obtain feature representations. Experimental results on eight image datasets are recorded in three tables.

Algorithm 1 Algorithm for presented model

Input: Original data $\boldsymbol{X} = \{\boldsymbol{x_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}}$in $\mathbb{R}^{\boldsymbol{m}}$, the dimension of hidden layer $\boldsymbol{d}$ and learning rate $\boldsymbol{\sigma}$.
Output: Low-dimensional feature representation $\boldsymbol{Y} \in \mathbb{R}^{\boldsymbol{d \times n}}$.
1: Generate corrupted data $\widehat{\boldsymbol{X}} = \{\widehat{\boldsymbol{x}}_{\boldsymbol{i}}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}}$;
2: Initialize weight matrix $\boldsymbol{W^{(1)}} \in \mathbb{R}^{\boldsymbol{d \times m}}$, bias vector $\boldsymbol{b^{(1)}}$and choose an activation function α.
3: **repeat**
4: **for**each point $\widehat{\boldsymbol{x}}_{\boldsymbol{i}}(\boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{n})$**do**
5: Compute $\boldsymbol{y_i}$by Formula (14);
6: Utilize Formula (15) to obtain $\boldsymbol{x_i'}$;
7: Update $\boldsymbol{W}$and $\boldsymbol{b}$by the following Formula $\boldsymbol{W^{(i+1)}} \leftarrow \boldsymbol{W^{(i)}} - \boldsymbol{\alpha}\frac{\boldsymbol{\partial}}{\boldsymbol{\partial W^{(i)}}}\boldsymbol{\Theta(\theta)}$and $\boldsymbol{b^{(i+1)}} \leftarrow \boldsymbol{b^{(i)}} - \boldsymbol{\alpha}\frac{\boldsymbol{\partial}}{\boldsymbol{\partial b^{(i)}}}\boldsymbol{\Theta(\theta)}$with gradient descent method;
8: **end for**
9:**until** convergence
10:**return**$\boldsymbol{Y}$.

4.1. Data Sets

In this section, we will introduce eight public standard datasets. These image datasets are Chars74K, USPS, Yale-B, COIL-20, ORL, CIFAR-10, Fashion-MNIST, SMSHP. The details of them are given in Table I. Specific description of eight image datasets are showed below.

The Chars74K dataset [20] contains two parts: English and Kannada. English symbols have three kinds. First kind contains 7705 characters come from natural images. Second one has 3410 hand drawn characters which use a tablet PC. The last one has 62992 synthesised characters which originate from computer fonts. This dataset is divided into 62 classes (a-z, A-Z, 0-9) and the pixel

size of each image is 32*32. We select a subset of Chars74K dataset. It has 44,044 training images and 8788 test images with 52 classes (a-z, A-Z).

USPS is a handwritten digit image dataset [21]. It owns 9298 handwritten digit images in total. Size of each image is 16*16. USPS is divided into two parts: 7291 training samples and 2007 test samples. The two subsets contain 10 different categories. Label '1' means digit 1 and label '0' represents digit 10.

The extended Yale Face Database B (YaleB) [22] is a face image database. YaleB includes 38 individuals and each individual has 64 images. We resize these image into 32*32 pixels. YaleB is divided to two subsets. Training one has 1928 samples and test one has 486 samples. They contain 38 different classes.

Columbia University Image Library (COIL-20) is an object image dataset. It is gray-scale. COIL20 has 20 objects and each object owns 72 images. They are taken from different angles. The size of these image is 32*32 pixels. COIL-20 contains 1440 samples. Each sample is represented by a 1024-dimensional vector. We divide the dataset into two subsets. First has 1140 training examples and second owns 300 test examples.

Olivetti Research Laboratory (ORL) [23] is a face image dataset. It contains 40 subjects with different ages, sexes and races. There are 10 images in each subject. ORL was made at different times, varying the lighting, facial details (glasses / no glasses) and expressions (smiling / not smiling, open / closed eyes). Each image is resized to 1024-dimensional vector. The dataset has 40 classes in all.

Cifar-10 is a standard color image dataset. It is made up of 60000 images which originate from a larger scale dataset. Cifar-10 contains 10 classes (cat, dog, automobile, bird, airplane, deer, ship, frog, horse, truck). There are 6000 images in each class. The size of image is 32*32. It is split into two subsets. Training samples have 1928 images and test samples own 486 images. They contain 38 different classes.

Fashion-MNIST is a clothing image dataset. It contains 10 classes (bag, coat, trouser, shirt, sandal, T-shirt, dress, pullover, sneaker, ankle boot). Fashion-MNIST includes 60,000 training samples and 10,000 testing samples. The size of each image is 28*28 pixels. Each sample is represented as a 784-dimensional feature vector.

SMSHP (Sebastien Marcel Static Hand Posture) is a hand-posture image dataset [24]. It consists of 5531 images. SMSHP is divided into 6 different types (point, five, v, a, b, c). For simplicity, the size of these hand posture images is denoted as 32*32 pixels. They are split into two subsets. First one has training images and second one owns 1106 test images. Each example is unified as a 1,024-dimensional feature vector.

Table 1. A brief description of the tested datasets.

ID	datasets	# samples	# features	# classes
1	Chars74K	52832	1024	52
2	COIL-20	1440	1024	20
3	USPS	9298	256	10
4	ORL	400	1024	40
5	YaleB	2414	1024	38
6	Cifar-10	60000	3072	10
7	SMSHP	5531	1024	6
8	Fashion-MNIST	70000	784	10

4.2. Parameter Setting

In this paper, learning rate of all methods is set as 0.01. For an objective comparison, we reduce each dataset into k-dimension uniformly. Where the number of hidden layer units k is set as 40. The size of each bach is denoted as 100. And we fix the number of training as 50. For simulating distorted image, we add random noises into those eight image datasets. Ten samples from each processed dataset are demonstrated in Figure 3.

Figure 3. The samples of distorted datasets.

4.3. Evaluation Metrics

In this section, we mainly introduce the evaluation metric of clustering. In the final stage of experiment, the k-means algorithm is used to calculate performance of extracted features. Consider a sample dataset $D = \{x_1, \dots, x_n\}$.The clusters obtained by k-means algorithm for clustering are denoted as $C = \{C_1, \dots, C_n\}$.Then the square error can be computed by

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - u_i\|_2^2 \qquad (17)$$

where $u_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$.The core of algorithm is to minimize E.

Clustering accuracy is an important reference index of clustering performance [25]. It is used to compare predicted labels with real labels provided by data. The value of clustering accuracy can be presented as

$$ACC = \frac{\sum_{i=1}^{n} \delta(s_i, map(r_i))}{n} \qquad (18)$$

where r_i and s_i represent predicted label and real label separately. The number of data is set as n. $\delta(x, y) = 1$ if $x = y$, otherwise$\delta(x, y) = 0$.Normalized mutual information can be used to measure the similarity of clustering results [26]. Consider a mutual information $I = (\Omega; C)$.It represents the increase of category information Ω by giving cluster information C. Then normalized mutual information is presented as

$$NMI = \frac{I(\Omega; C)}{(H(\Omega) + H(C))/2} \qquad (19)$$

where H means entropy. Adjusted Rand index is a functionto calculate the distribution similarity of two labels [27]. Thisfunction has no requirements for the definition form of label.

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}} \qquad (20)$$

where $ARI \in [-1, 1]$.The higher the value, the more consistent the clustering results with the real situation.

4.4. Experimental Results

In this section, comprehensive experiments are presented. In addition to denoising feature learning model, we run several classical feature learning algorithms for comparison. These methods include PCA, NPE, LPP, Isomap, LLE, IsoP and auto-encoder. For assessing the performance of feature representation, we choose k-means algorithm to make a clustering. Three popular evaluation metrics ACC, NMI and ARI are used for revealing an intuitive result. Meanwhile the original data without dimensionality reduction is also sent into k-means as the baseline. Finally, all the experimental results on eight processed image datasets are displayed in three tables.

From the Table 2 - 3, it is intuitive that the low-dimensional feature representations extracted by denoising model run a better performance. In Table 2, the clustering accuracy of denoising model ranks first on seven image datasets except USPS. In Yale-B dataset, the feature representations extracted via denoising model perform a favorable result compared with the baseline. In Table 3, the performance of denoising model reach first on six datasets. On Chars74K image dataset, normalized mutual information of denoising model is 60.8% while baseline is 45.3%, with a greater improvement. On CIFAR-10, denoising model reaches second best result which is close to locality preserving projections. In Table 4, the adjusted rand index of denoising model ranks first on six datasets while classic unsupervised feature learning algorithms also perform well. Especially on the extended Yale Face Database, the adjusted rand index of denoising model reaches a bigger improvement.

Table 2. Clustering accuracy (mean% + std%) with different unsupervised feature learning methods

dataset/method	Chars74K	USPS	Yale-B	COIL-20	ORL	CIFAR-10	F-MNIST	SMSHP
Baseline	31.8±0.9	65.6±1.8	11.3±0.4	56.1±2.7	63.6±1.8	23.9±0.3	54.4±1.2	38.7±1.5
PCA	34.5±1.6	68.2±2.3	12.2±0.3	65.7±1.4	64.8±2.6	24.5±1.2	59.3±0.7	36.5±1.1
NPE	35.6±0.8	**71.3±1.8**	28.2±1.5	61.3±2.3	72.7±1.4	21.3±0.5	52.4±0.9	36.8±1.2
LPP	33.2±1.5	68.5±1.4	30.3±1.4	66.7±0.5	68.6±2.5	24.2±0.8	58.1±3.3	38.3±0.9
Isomap	24.6±0.3	67.3±2.6	32.6±2.1	68.6±1.7	59.8±2.6	23.9±1.2	57.2±2.9	34.5±0.4
LLE	28.9±1.2	64.2±1.5	27.4±1.7	60.2±1.1	53.6±1.7	25.4±2.3	53.7±1.8	33.6±1.5
IsoP	34.3±2.5	70.2±0.9	25.3±0.8	65.3±2.8	62.4±2.1	26.3±1.6	54.2±1.3	35.2±1.2
Auto-encoder	32.6±1.2	67.4±2.4	19.7±0.4	59.9±1.3	65.6±3.5	29.7±0.9	58.2±2.6	34.3±0.8
Ours	**37.2±0.6**	69.5±0.8	**33.8±1.2**	**70.4±2.2**	**74.5±2.3**	**32.5±1.1**	**61.5±1.4**	**39.9±0.6**

Table 3. Normalized mutual information (mean% + std%) with different unsupervised feature learning methods.

dataset/method	Chars74K	USPS	Yale-B	COIL-20	ORL	CIFAR-10	F-MNIST	SMSHP
Baseline	45.3±0.7	63.6±1.6	12.8±0.5	74.9±1.8	73.4±2.2	8.6±0.6	54.5±1.4	7.1±0.8
PCA	50.6±2.3	61.0±0.2	14.3±0.4	76.7±2.3	76.8±1.5	8.2±0.4	53.2±1.6	8.4±0.6
NPE	55.5±0.8	63.2±0.8	37.6±1.6	74.6±0.6	80.2±3.2	8.5±0.3	52.1±0.7	11.9±0.8
LPP	53.4±0.9	67.6±1.9	35.4±0.3	76.8±1.4	77.9±1.8	**9.8±0.6**	58.7±2.2	9.3±1.2
Isomap	45.3±1.2	65.9±2.2	36.8±2.0	73.5±2.1	74.1±2.3	9.2±0.1	54.6±0.8	8.8±0.7
LLE	52.7±1.6	63.2±2.8	23.5±1.4	72.3±4.2	73.8±2.7	7.6±0.2	53.4±0.6	**12.0±0.5**
IsoP	49.6±2.3	58.3±1.5	29.3±2.3	78.5±3.6	75.7±1.3	8.2±0.7	52.6±0.9	9.2±0.3
Auto-encoder	49.2±1.5	56.7±2.4	22.6±1.2	75.2±1.9	77.3±2.6	7.4±0.3	56.2±1.8	8.3±0.5
Ours	**60.8±2.2**	**68.1±2.3**	**39.6±0.8**	**79.1±2.3**	**82.4±2.4**	9.6±0.2	**62.5±1.3**	10.1±0.6

Table 4. Adjusted rand index (mean% + std%) with different unsupervised feature learning methods.

dataset/method	Chars74K	USPS	Yale-B	COIL-20	ORL	CIFAR-10	F-MNIST	SMSHP
Baseline	20.8±0.9	53.6±2.2	2.3±0.2	50.7±3.6	48.4±2.1	4.5±0.6	38.6±0.9	4.4±0.2
PCA	23.5±0.4	60.2±0.6	3.5±0.4	55.3±1.8	49.6±2.0	5.1±0.6	41.2±0.7	6.1±0.3
NPE	24.3±2.2	56.6±0.7	13.6±0.3	53.3±2.5	53.6±1.4	5.7±0.4	33.7±1.3	5.4±0.1
LPP	20.7±0.8	62.2±1.3	12.4±0.5	61.8±3.2	46.8±2.3	4.2±0.1	42.3±2.5	4.8±0.9
Isomap	21.1±0.6	62.7±2.4	13.8±0.2	60.5±1.3	38.9±0.6	4.7±0.6	40.7±1.8	3.7±0.2
LLE	19.3±0.4	56.7±3.2	9.2±0.8	51.4±2.6	45.6±0.7	6.1±0.2	**46.3±0.7**	4.5±0.3
IsoP	22.5±0.3	58.2±2.8	11.8±0.7	**64.3±3.2**	40.4±1.5	6.7±0.5	43.6±1.2	5.1±0.2
Auto-encoder	23.2±0.6	61.8±3.5	6.9±0.2	52.5±2.4	42.3±2.3	7.3±0.8	42.1±0.5	7.3±0.6
Ours	**26.4±0.5**	**65.9±2.3**	**14.6±0.4**	54.9±2.1	**55.7±1.6**	**7.5±0.4**	44.7±0.3	**8.8±0.4**

5. CONCLUSION

In this paper, facing the problem regard to high-dimensional of distorted images, an end-to-end denoising feature learning model was proposed to obtain high robust feature representations. Then the extracted features were evaluated by *k*-means clustering. Compared to other unsupervised feature learning methods, extensive experiments on eight processed image datasets demonstrated that denoising model ran a competitive performance. The low-dimensional representation could replace the original dataset primely. But in the experiment, it was obvious that larger dimensions and categories caused a bad influence on clustering performance. In the future work, we will be concerned with the image datasets which own many categories. We may add semi-supervised training to attempt a better result.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants No. 61877010 and No. 11501114, and the Fujian Natural Science Funds under Grant No. 2019J01243. Genggeng Liu is the corresponding author of the article.

REFERENCES

[1] S. Wang, W. Zhu, Sparse graph embedding unsupervised feature selection, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48 (3), pp. 329–341, 2018.
[2] S. Wang, W. Pedrycz, Q. Zhu, W. Zhu, Subspace learning for unsupervised feature selection via matrix factorization, Pattern Recognition, 48 (1), pp. 10–19, 2015.
[3] S. Wang, J. Lu, X. Gu, H. Du, J. Yang, Semi-supervised linear discriminant analysis for dimension reduction and classification, Pattern Recognition, 57 (C), pp. 179–189, 2016.
[4] D. Cai, X. He, K. Zhou, J. Han, H. Bao, Locality sensitive discriminant analysis, In Proceedings of the 20th International Joint Conference on Artifical Intelligence, pp. 708–713, 2007.
[5] H. Hotelling, Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology, 24 (6), pp. 417–520, 1933.
[6] X. He, Locality preserving projections, Advances in Neural Information Processing Systems, 16 (1), pp. 186–197, 2003.
[7] Z. Huang, X. Xu, L. Zuo, Reinforcement learning with automatic basis construction based on isometric feature mapping, Information Sciences, 286, pp. 209–227, 2014.
[8] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science, 290 (5500), pp. 2323–2326, 2000.
[9] C. Deng, X. He, J. Han, Isometric projection, in: National Conference on Artificial Intelligence, pp. 528–533, 2007.
[10] Y. Wang, H. Yao, S. Zhao, Auto-encoder based dimensionality reduction, Neurocomputing, 184 (C), pp. 232–242, 2016.

[11] J. Deng, Z. Zhang, E. Marchi, B. Schuller, Sparse autoencoder-based feature transfer learning for speech emotion recognition, in: Affective Computing and Intelligent Interaction, pp. 511–516, 2013.
[12] J. Walker, C. Doersch, A. Gupta, M. Hebert, An uncertain future: forecasting from static images using variational autoencoders, in: European Conference on Computer Vision, Springer, pp. 835–851, 2016.
[13] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434.
[14] P. Fränti, S. Sieranoja, K-means properties on six clustering benchmark datasets, Applied Intelligence, 48 (12), pp. 4743–4759, 2018.
[15] X. He, D. Cai, S. Yan, H. J. Zhang, Neighborhood preserving embedding, in: Tenth IEEE International Conference on Computer Vision, pp. 1208–1213, 2005.
[16] A. N. Gorban, B. Kgl, D. C. Wunsch, A. Y. Zinovyev, Principal manifolds for data visualization and dimension reduction, Springer Berlin Heidelberg, 2008.
[17] C. Varini, A. Degenhard, T. W. Nattkemper, Isolle: lle with geodesic distance, Neurocomputing, 69 (13), pp. 1768–1771, 2006.
[18] B. Raytchev, I. Yoda, K. Sakaue, Head pose estimation by nonlinear manifold learning, in: IEEE Proceedings of the 17th International Conference on Pattern Recognition, Vol. 4, pp. 462–466, 2004.
[19] A. Gogna, A. Majumdar, R. Ward, Semi-supervised stacked label consistent autoencoder for reconstruction and analysis of biomedical signals, IEEE Transactions on Biomedical Engineering, 64 (9), pp. 2196–2205, 2016.
[20] C. Yi, X. Yang, Y. Tian, Feature representations for scene text character recognition: A comparative study, in: 12th International Conference on Document Analysis and Recognition, IEEE, pp. 907–911, 2013.
[21] K. Proedrou, I. Nouretdinov, V. Vovk, A. Gammerman, Transductive confidence machines for pattern recognition, in: European Conference on Machine Learning, Springer, pp. 381–390, 2002.
[22] A. S. Georghiades, P. N. Belhumeur, D. J. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23 (6), pp. 643–660, 2001.
[23] G. Guo, S. Z. Li, K. Chan, Face recognition by support vector machines, in: Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 196–201, 2000.
[24] W. Guo, G. Chen, Human action recognition via multi-task learning base on spatial–temporal feature, Information Sciences, 320, pp. 418–428, 2015.
[25] K. A. Nazeer, M. Sebastian, Improving the accuracy and efficiency of the k-means clustering algorithm, in: Proceedings of the World Congress on Engineering, Vol. 1, pp. 1–3, 2009.
[26] Z. F. Knops, J. A. Maintz, M. A. Viergever, J. P. Pluim, Normalized mutual information based registration using k-means clustering and shading correction, Medical Image Analysis, 10 (3), pp. 432–439, 2006.
[27] J. M. Santos, M. Embrechts, On the use of the adjusted rand index as a metric for evaluating supervised classification, in: International Conference on Artificial Neural Networks, Springer, pp. 175–184, 2009.

AUTHORS

Qihao Lin

Qihao Lin received the B.E. degree in Network Engineering from Fuzhou University, Fuzhou, China, in 2018. He is currently pursuing the M.S. degree with the College of Mathematics and Computer Science, Fuzhou University. His research interests include machine learning and computer vision.

Jinyu Cai

Jinyu Cai received the B.S. degree in computer science and technology from Fuzhou University, Fuzhou, China, in 2018. He is currently pursuing the M.S. degree with the College of Mathematics and Computer Science, Fuzhou University. His research interests include machine learning, computer vision and pattern recognition.

Genggeng Liu

Genggeng Liu received the B.S. degree in computer science and the Ph.D. degree in applied mathematics from Fuzhou University, Fuzhou, China, in 2009 and 2015, respectively. He is currently an Associate Professor with the College of Mathematics and Computer Science, Fuzhou University. His contributions have been published in IEEE Transactions on Cybernetics, IEEE Transactions on Industrial Informatics, ACM Transactions on Design Automation of Electronic Systems, etc. His research interests include computational intelligence and its application.

A GAME-BASED INTERACTIVE TRAINING SYSTEM FOR IMPROVING THE ENGAGEMENT OF HAZARD LEARNING AND TRAINING

David Hu[1], Jack Gao[2], Yu Sun[3], Fangyan Zhang[4]

[1]Portola High School, Irvine, CA 92618, USA
[2]Sage Hill School, Irvine, CA 92618, USA
[3]California State Polytechnic University, Pomona, CA, 91768, USA
[4]ASML, San Jose, CA, 95131, USA

ABSTRACT

Natural Disasters, events that are frequently occurring around the world, taking away homes and innocent lives within minutes or even seconds with the only goal of destruction. Only when we see it on the news of fatalities around the globe when we realize how fragile human life is. The most concerning problem is that current disaster training procedures are not sufficient in preparing the general public in the case of natural disasters such as earthquakes which can wipe out thousands of homes and cause massive casualties if not properly prepared for. To address this situation, the first prototype that we came up with was the Safety Lifetime earthquake simulation game. We believe simulation-based learning would be better at covering more information as well as making the lessons more memorable. As a prototype, Safety Lifetime only contains the simulation of a real earthquake, and lessons to guide the user through what to do during an earthquake, what items to collect, what is the safest sheltering location, and more. In order to verify the effectiveness of the training system, we performed a small-scale user study. 10 users are divided into two groups. Group A is given the booklet earthquake educational materials, while Group B is provided with the game system. Each group spends 10 minutes to learn the content, followed by finishing a quiz. The result shows that the average score of the Group A is 8.5/10, while the average score of the Group B is 9.3/10.

KEYWORDS

Disaster training, earthquake, simulation-based learning, training system.

1. INTRODUCTION

There have been cases of extreme natural disasters that have happened throughout the years. Here we take Earthquake as an example as they are the most prominent threat in California. At least 369 people die - most in and around Mexico City - during a magnitude 7.1 earthquake in 2017. in 2018, more than 460 people are killed after a 6.9 magnitude earthquake hit the Indonesian island of Lombok [3]. It leveled homes, mosques, and businesses, displacing some 350,000 people. In the United States, the infamous San Francisco Earthquake in 1908 caused more than 3000 death Only when we see it on the news of fatalities around the globe when we realize how fragile human life is. This year started with dramatic events such as kobe's unfortunate death, the coronavirus, and the rumors of a world war which really makes you wonder how life could be gone in an instant, which will change everything. Natural disasters, like earthquakes, are sudden and almost unpredictable, so it is good to always take precautions and be prepared for when these events strike. However, the central problem of this is the lack of natural disaster survival

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 29-36, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101203

education. As reported in a FEMA survey in 2015, nearly 60 percent of American Adults have not participated in natural disaster training [1]. This is even more detrimental when considering that 80 percent of Americans live in counties that have been hit by a weather-related disaster since 2007 as reported by the Washington Post in 2018 [2]. The people in high rate earthquake countries like China and Indonesia almost have no access to education about these natural disasters, which is the reason why this simulation/program would serve as a protection and education for the safety of these individuals [14].

Natural disasters are extremely dangerous and have devastating effects, especially on those who are inadequately prepared [4] [5]. Currently, for most adults, earthquake and natural disaster training systems present the survival information in a word-based format such as websites and booklets or in a video and lecture-based format as seen on YouTube and television. The first practical problem with current disaster training systems is the lack of coverage as few American would actively seek out the information online if it is not directly presented to them. In addition, it must be recognized that text-based learning is not effective at conveying large quantities of information in a concise or enjoyable way. Though readily accessible, the long articles and texts on the many websites are often complemented by very few pictures in between and may discourage learners to read through information in its entirety. This will lead to a less comprehensive and potentially flawed understanding of the earthquake survival process. As a result, the learner may struggle to recall much of the information presented due to the lack of constant learner engagement. Though video-based learning addresses some of the flaws mentioned above, it is still problematic as most videos only explain a part of earthquake survival but do not present the information regarding earthquake survival in a holistic manner. In addition, the lack of learner engagement remains problematic with video-based learning as well. The reality is that people are not educated to face any of these devastating events, nor have an understanding of what to do during situations of a disaster under the current system.

In this paper, we utilized Unity and visual studios to create a game like simulation on earthquake and added in the lessons, with audio recording and quizzes at the end for the users to take so that they have understood the material they have learned in each lesson [6] [7] [11]. Our goal in creating this project is to introduce earthquake survival knowledge in a more concise and interesting way by presenting them in the form of a game. In order to verify the effectiveness of our simulation-based training system, we performed a small-scale user study. We selected 20 users of similar educational background. The 20 users are divided into two groups. Group A is given booklets of earthquake educational materials, while Group B is provided with the game system, Safety Lifetime. Each group spends 10 minutes to learn the content and is asked to finish a quiz at the end of the 10-minute studying period. The contents of the quiz are all covered in the booklets and the simulation, and none of the questions overlap with the questions that were part of the simulation. The Quiz taking environments are the same, and all communications or outside interference have been monitored and deterred. The result shows that the average score of Group A is 8.5/10, while the average score of Group B is 9.3/10. This 0.8 difference is quite significant as this is only a ten-question quiz, and there is an eight percent increase in two groups of people with very similar educational backgrounds. However, this result cannot definitively prove the simulation's effectiveness as slight variations in the subjects' inherent earthquake survival knowledge can have a noticeable impact in a small subject pool. The subjects from Group B may have another advantage. The quizzes within the simulation may have better prepared them for the test quiz given to them at the end of their 10-minute studying period. Nonetheless, the simulation was able to achieve our initial goal for the project. It seemed to be effective at educating the test groups as it presents a more engaging game-based learning method. Its true effectiveness can be proven upon more comprehensive testing with bigger subject groups chosen specifically for their similar level of earthquake survival knowledge before the experiment.

The rest of the paper is organized as follows: Section 2 explains the challenges we faced in designing our prototype; Section 3 further elaborates on the simulation's design process and creation as well as giving further details regarding specific components of the simulation; Section 4 explains the details regarding our experimentation; Section 5 presents the related works on this topic, some of which inspired the creation of this project. Finally, Section 6 gives the concluding remarks, as well as elaborating on future developments of this project.

2. CHALLENGES

There are a couple of challenges in the project. They will be discussed one by one in this section.

2.1. Challenge 1: The Differences in User Platforms.

In order to further spread awareness and educate people on natural disasters survival, it is important to make our simulation and game accessible to as many people as possible. It is difficult to make the simulation and game compatible with the vast variety of platforms that the user may wish to use. Some of the simulations and games may need to be reprogrammed to be compatible to different platforms but this remains an issue we will focus on later in the project's development.

2.2. Challenge 2: The Difficulty to Give Important Additional Information to Users.

The purpose of our simulation and game is to present more interactive and educational lessons on natural disaster survival. However, immediately we realized that California does a fairly good job on public education regarding earthquake survival, our primary objective of interest. Everyone in California and most likely in the United States will have some basic knowledge on the concept of duck and cover. Though there is some additional information involved with earthquake survival, many of them are merely suggestions that will not significantly enhance the users' chance of survival. Inclusion of those insignificant details will only distract the user of the essential ideas behind earthquake survival. In short, to develop our intended lessons, we aim to add important additional information to enhance the prevalent lessons on Earthquake survival while simultaneously avoid overloading the users with information. We realized that current Earthquake lessons only focus on the things to do during an earthquake and not so much before or after. The methods of Earthquake preparations are often tedious and difficult for users to remember from traditional lessons. As a result, we chose to develop a three-staged lesson with additional emphasis on the essential safety procedures before and after the Earthquake. In doing so, we hope to present a more complete and interactive lesson of the necessary procedures in Earthquake survival.

2.3. Challenge 3: The Difficulty to Set Up Complex Branching Pathways in the Game.

Currently in the game, one must complete all the tasks in the game in order to proceed to the other lessons. While this serves our purpose in reinforcing the lessons we gave previously before the simulation, this does not seem to be an accurate reflection of the sequence of real-life events. For example, not bringing bandages in the survival kit does not prevent the earthquake from happening as it does in the simulation when it prevents the lesson's progression. However, in real life, the lack of bandages increases one's risk for infections after the Earthquake which the simulation does not show. In conclusion, while our first prototype gives the user a better knowledge in the complete procedure of earthquake survival, it does not adequately show the reason behind the actions in procedure. Thus, this problem may negatively affect the

effectiveness of the simulations in making the lessons more memorable. This is a problem we will later address as we introduce more complexities into our prototype.

3. SOLUTION

3.1. Overview of the Solution

Safety Lifetime is a game-based learning system created through Unity using a C# Script [12] [13]. Our method is inspired by the introduction of simulation-based learning in colleges and medical programs as they have shown to be more effective than more traditional styles of learning. In this simulation, we ask the learner to apply basic earthquake survival knowledge they learned into situations of realistic scenarios during a real earthquake. The first lesson and simulation would be what to do before an earthquake happens, and what items to pick up that will be best in that situation. The second lesson focuses on during the earthquake, including where the best places for hiding are, and the best shelters. The third will be around after the disastrous event, what to do and where the safest place is. In doing so, we also hope to address the lack of consistent learner engagement seen in current earthquake survival training programs, and therefore make the lessons more interesting and memorable.

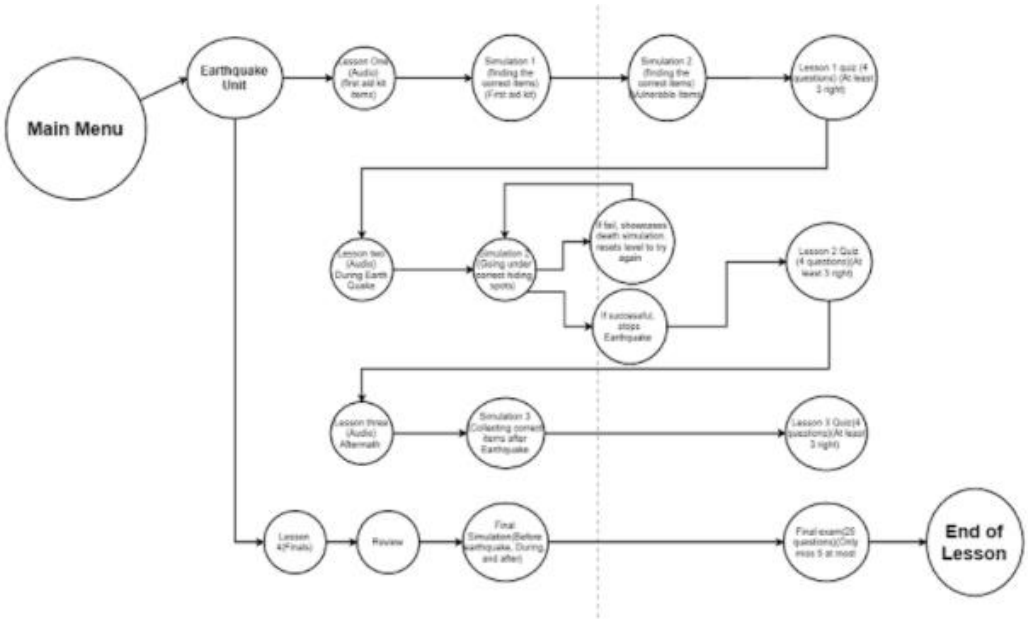


Figure 1: components of the Safety Lifetime system

As a demonstration of the components, we should look to this flow diagram that we have created. Upon the main menu, there will be a button for the learner to select the earthquake unit if you wish to precede. We broke the entire lesson into three smaller lessons: lesson one for earthquake preparation, lesson two for safety during an earthquake, and lesson three for what to do after an earthquake. The lessons must be learned in order, and upon completion, the learner will have the option to take lesson four which is a compilation of all the previous lessons if the learner wants to review the material in a do to prepare for an earthquake. After the video, lesson one of the simulations will run and require the learner to pick up items essential to earthquake preparation. Once the first-person character comes within the sphere collider, the learner can hit P to collect the item. The collection of all six items trigger the next event which will be a six-problem quiz. If the learner gets less than five wrong on the quiz, it will trigger the event for the video of lesson

two where the learner will learn about duck-and-cover. The simulation for lesson two will start after the learner finishes the video of lesson two and require the learner to hide in a spot marked by a small green ball under a table for duck-and-cover. We used the Unity animator to rotate the head of the first-person character at various angles to create the earthquake affect, and we used the same script as the previous lesson to trigger the quiz event. Once the first-person character comes within the collider of the green ball, it will automatically collect the ball and thus trigger the next event. We added a few seconds lag to more realistically simulate the earthquake, and the quiz will start in three seconds after initial collection of the green ball. The completion of the quiz will take the learner to the video of lesson three which will cover the safety procedure after an earthquake. Similarly, after the end of the video, the learner will be required to apply their knowledge in a simulation where they have to first contact local authorities and move to a safe clearing out of the house. The completion of the simulation will take the learner to the final quiz. After completion of the final quiz, the lesson will end.

3.2. Implementation

In Safety Lifetime, when a character is moving out of the house, there is an object outside that represents the action of contacting emergency services, and for it to be activated the character needs to be near or at one place with the project, and press key (p) to activate the action. However, it is hard to assign the letter to represent the action of picking up and using the item. The way I solved the problem is using the concept called collider in unity, assigning the player and the sign as colliders. When they collide with each other, it generates an event which would show that the player has activated the action at the item's spot. Another challenge was I did not know how to create an easier way for the quiz to move onto the next scene after the user answers 5 questions right. There are multiple quizzes, which means I have to write a function multiple time. I solved that problem by creating one function called correct *choice()* which transitions into the next scene when called to run.

```
// PlayerMovement
private void PlayerMovement( float horizontal, float vertical )
{
   bool grounded = controller.isGrounded;
   Vector3 moveDirection = myTransform.forward * vertical;
    moveDirection += myTransform.right * horizontal;
    moveDirection.y = -10f;
    if(jump)
    {
            jump = false;
            moveDirection.y = 25f;
            isPorjectileCube = !isPorjectileCube;
    }
    if (grounded) {
            moveDirection *= 7f;
            controller.Move( moveDirection * Time.fixedDeltaTime );
   }
    if (!prevGrounded && grounded ) {
            moveDirection.y = 0f;
            prevGrounded = grounded;
    }
}
```

The two parameters (horizontal and vertical) represent how hard the user presses the button, vertical and horizontal for the character's movement in the game. W, A, S, D are represented for movement and space for jump. The Boolean value of bool grounded determines if the user-controlled character is on the ground or if the character is in the air. The "myTransform.forward" and the "transform.right" uses the physics equation of the vector function(a value, and a direction) to move the player from the position it is in, which ultimately performs the action using the "controller.Move" function. If the user is jumping, proved by boolean value true or false, the mathematical equation of multiplying the position of y to be a fixed value of 25 by adding 25.If user is grounded, proved again by boolean, the movement is generated to be larger and multiplied speed by 7 (moveDirection *= 7f) to move faster on the ground. There are two logical parameters, prevGrounded and grounded. If prevGrounded and now grounded, it means the user was in the air, and the value for ground movements are added after landing. This is the overall process of movement in one frame of the program.

```
public void correctChoice()
{
  correct += 1;
  if (correct >= 5)
  {
      UnityEngine.SceneManagement.SceneManager.LoadScene(NextScene);
   }
   questionPanels[questionNumber].SetActive(false);
   questionNumber += 1;
   if(questionNumber<questionPanels.Length){
     questionPanels[questionNumber].SetActive(true);
   }
}
```

This abstraction I chose is the program for true or false answers on a quiz. If the user gets one question right, he gets 1 point. Using a loop and creating a function called *"correctChoice()"* made it easier. The program also determines whether the user has picked enough correct answers and achieved a score of at least 5 questions right. As the user passes the score, the next scene and simulation loads, whereas if the user gets less than 5 questions correct, it starts over at question one of the quizzes. It would be complex to have this code in multiple places, since the quiz and questions appear multiple times after each lesson and simulation. Now, if the function is called, it takes care of most of the quiz program and achieves the goal of counting points and transitioning to the next scene.

4. EXPERIMENTS

To verify the effectiveness of our simulation, we performed a small-scale user study in which we selected 20 users of very similar educational backgrounds. Group A is given booklets of earthquake educational materials, while Group B is provided with the game system, Safety Lifetime. Each group was given a 10-minute studying period, and were asked to take a quiz regarding earthquake safety. The contents of the quiz are all covered in the booklets and the simulation, and none of the questions overlap with the questions that were part of the simulation. The Quiz taking environments are the same, and all communications or outside interference have been monitored and deterred. The results were then compared to see if there is a distinguishable difference between the two learning systems.

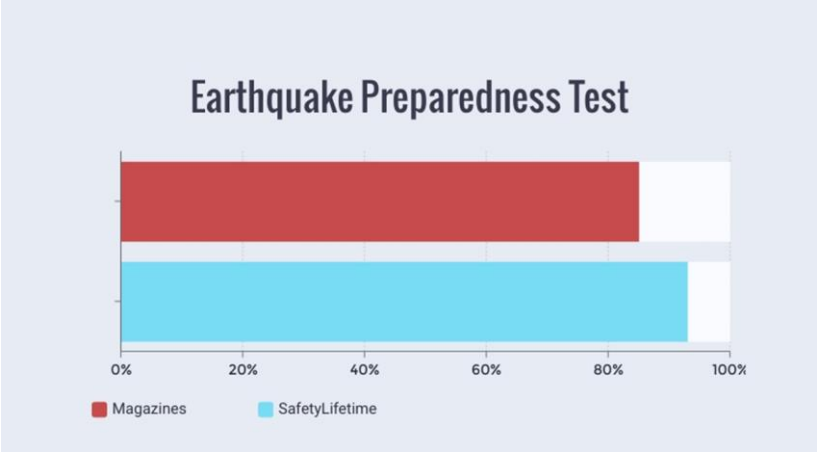


Figure 2: Earthquake Preparedness Test of the Safety Lifetime system

The graph above shows the averaged quiz scores of Group A and B. As can be seen, there is an 8% difference in their quiz scores with the scores from Group B being slightly higher than Group A. Though this experience cannot definitively prove the simulation's effectiveness as slight variations in the subjects' inherent earthquake survival knowledge can have a noticeable impact in a small subject pool, it does seem to suggest a correlation since a 0.8-point score difference is quite significant for a 10-question quiz.

5. RELATED WORK

Dr. Fatimah Lateef explained that the introduction and use of simulation-based learning can be very valuable in replacing and amplifying real experiences in medical education and that "Teamwork training conducted in the simulated environment may also offer an additive benefit to the traditional didactic instruction, enhance performance, and possibly also reduce errors." [8] In this project, we also attempted to utilize simulation-based training to enhance the quality of earthquake survival education. Steadman, Randolph H utilized a randomized control trial to show that full-sale simulation-based learning is more effective than problem-based learning for teaching fourth year medical students regarding acute care assessment and management skills [9]. In this work, we also show similar improvements in the quiz score of learners using our simulation compared to traditional style learning. Cant, Robyn P., and Simon J. Cooper reported in their paper that their review of quantitative evidence for medium to high fidelity simulation using manikins in nursing, in comparison to other educational strategies concluded that simulation based learning is an effective teaching and learning method when best practice guidelines are adhered to and may have advantages over other teaching methods [10]. Though this report examines mostly physical simulations, the advantages of physical simulations are comparable to the digital simulation we utilized in our project.

6. CONCLUSION AND FUTURE WORK

In this project, we created a simulation which we named Safety Lifetime in hopes for finding a more effective method to educate learners on earthquake survival. In order to test the effectiveness of the new training system, we performed a small-scale user study in which we asked two groups of ten users with similar educational backgrounds to study and take a 10-question quiz in a similar and highly controlled environment. Group A was given traditional earthquake survival education material while the other was given the Safety Lifetime simulation. The result shows that the average score of Group A is 8.5/10, while the average score of Group B

is 9.3/10. This 0.8 difference is quite significant as this is only a ten-question quiz though it cannot definitely prove the effectiveness of the simulation as a result of the small subject pool which could be affected by the differences between the users' inherent earthquake survival knowledge before the experiment.

In this project, we are limited by time and resources, the graphics of the simulation and the control of the first-person view can be refined with more high-resolution assets from the Unity Asset store. In the simulation itself, we also did not include different branching pathways and different ending within the simulation due to its relative complexity. The simulation's compatibility with IOS platforms can also be optimized. In addition, what was stated in the methodology section was the coding process of Unity where we figured out the variables for each handlers and abstractions. Many algorithms were used with one following another. We also had to use parameters to build a scene and objects in the simulation. By linking each scene from algorithms and displays altogether, the program is created to run through a series of events.

The simulation is just a prototype. Our goal is to expand our prototype to present information regarding other natural disaster and accident preparation for the learners in a more effective and entertaining way to educate them about their personal safety.

REFERENCES

[1] Sixty Percent of Americans Not Practicing for Disaster: FEMA Urges Everyone to Prepare by Participating in National PrepareAthon! Day on April 30 | FEMA.gov, 28 Apr. 2015, www.fema.gov/news-release/2015/04/28/sixty-percent-americans-not-practicing-disaster-fema-urges-everyone-prepare.

[2] Grieser, Justin. Report: 243 Million Americans Affected by Weather Disasters since 2007. 9 Apr. 2013, www.washingtonpost.com/news/capital-weather-gang/wp/2013/04/09/report-243-million-americans-affected-by-weather-disasters-since-2007/.

[3] "History of Deadly Earthquakes." BBC News, BBC, 19 Aug. 2018, www.bbc.com/news/world-12717980.

[4] Abbott, Patrick L. Natural disasters. New York: McGraw-Hill, 2008.

[5] Alexander, David. Natural disasters. Routledge, 2018.

[6] Smith, Danial. "Earthquake Rebuild: A Game for the Stealth Learning of Middle School Math." (2014).

[7] Parisi, Tony. Learning virtual reality: developing immersive experiences and applications for desktop, web, and mobile. " O'Reilly Media, Inc.", 2015.

[8] Lateef, Fatimah. "Simulation-based learning: Just like the real thing." Journal of Emergencies, Trauma and Shock 3.4 (2010): 348.

[9] Steadman, Randolph H., et al. "Simulation-based training is superior to problem-based learning for the acquisition of critical assessment and management skills." Critical care medicine 34.1 (2006): 151-157.

[10] Cant, Robyn P., and Simon J. Cooper. "Simulation-based learning in nurse education: systematic review." Journal of advanced nursing 66.1 (2010): 3-15.

[11] Perreault, Gregory, Mimi Wiggins Perreault, and Matthew Van Dyke. "The Power of Digital Games in Disaster Preparation and Post-Disaster Resilience." 2017 International Communication Association Conference. 2017.

[12] Norton, Terry. Learning C# by developing games with unity 3D. Packt Publishing Ltd, 2013.

[13] Murray, Jeff W. C# game programming cookbook for Unity 3D. CRC Press, 2014.

[14] Frankenberg, Elizabeth, et al. Education, Vulnerability, and Resilience after a Natural Disaster. 2013, www.ncbi.nlm.nih.gov/pmc/articles/PMC4144011/.

AN EMPIRICAL STUDY WITH A LOW-COST STRATEGY FOR IMPROVING THE ENERGY DISAGGREGATION VIA QUESTIONNAIRE SURVEY

Chun-peng Chang, Wen-Jen Ho, Yung-chieh Hung,
Kuei-Chun Chiang and Bill Zhao

Institute for Information Industry, Taipei, Taiwan

ABSTRACT

Based on neural network and machine learning, we apply the energy disaggregation for both classification (prediction on usage time) and estimation (prediction on usage amount) on 150 AMI (Advanced Metering Infrastructure) smart meters and a small amount of HEMS (Home Energy Management System) smart plugs in a community in New Taipei City, Taiwan. The aim of this paper is to clarify how we lower the cost, obtain the model of appliance usage from only a small portion of households, improve it with simple questionnaire, and generalize it for prediction on collective households. Our investigation demonstrates the benefits and various possibilities for power suppliers and the government, and won the Elite Award in the Presidential Hackathon 2020, Taiwan.

KEYWORDS

Energy Disaggregation, Non-intrusive Load Monitoring, Deep Learning, Autoencoder

1. INTRODUCTION

The big data of electricity sales services will be used to provide users with more various value-added applications and power suppliers with business opportunities. Energy disaggregation, or so called NILM (Non-intrusive Load Monitoring), is a particular study field in the electricity industry, and has huge potential to benefit targets mentioned above. It was developed by George W. Hart [1] in the 80s, to infer the individual states of the appliances from the aggregated meter measuring the voltage and the current from outside the houses. This is exactly the literal meaning of "Non-intrusive" in NILM. Nowadays, it is not only a theoretical study, but also a practical strategy going to start in many countries.

A recent research work of Kelly and Knottenbelt [2] have demonstrated the possibility of utilizing deep learning, which leads successful progress in many fields, such as image recognition, into the region of NILM. Hereafter, many researches in energy disaggregation [3] was developed quickly. These researches, however, are not suitable for numerous households outside the laboratory due to both the price and the privacy. Expensive meters with high sampling rate are needed for every appliance inside the house, and hence not applicable for a generalization to the whole city or the whole country. On the other hand, AMI, the cheap smart meters outside the house with low-sampling rate of 1mHz (sampling period of 15min) [4][5], are quite suitable. And more and more countries regard AMI as fundamental infrastructure. Our

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 37-43, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101204

study is based on Ming-Hsuan-Huang-Cheng, a real community in New Taipei City, Taiwan. This community is of 150 households, in which all are with AMI outside their houses. Moreover, within this community we have collected 20 volunteer households and deployed smart plugs of HEMS inside each of their house, for up to five appliances (air conditioner, refrigerator, washing machine, bottle warmer, and television) and the total power. We have collected the HEMS data of these volunteers for 1 year so far. This study focuses on the period of June 2020, for both AMI and HEMS data, and make classification and estimation on the 150 AMI households with their AMI meters only.

For the reference, we provided a visual example of AMI and HEMS data, as figure 1 and 2.

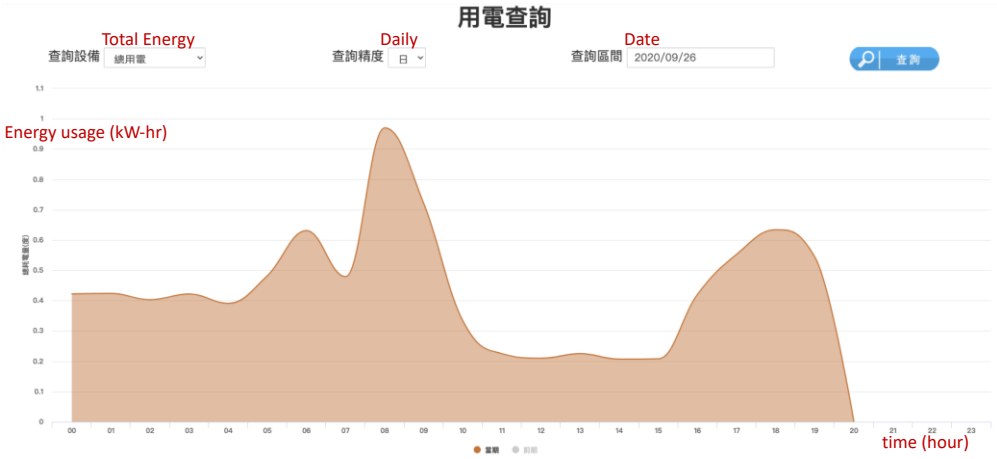


Figure 1: A visual example of total energy consumption in AMI data.

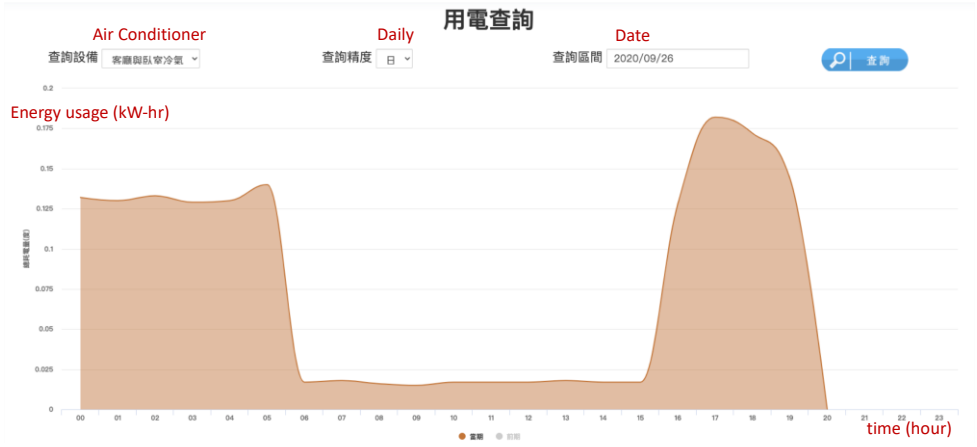


Figure 2: A visual example of energy consumption of the air conditioner in HEMS data.

2. METHOD

This study is composed of 2 parts: NILM estimation and classification.

2.1. NILM estimation

The first part, estimation, is to estimate the portion of energy consumed owing to each appliance. The estimation strategy is inspired by our previous research “An Analysis of Semi-Supervised

Learning Approaches in Low-Rate Energy Disaggregation" [5] and imitate a semi-supervised learning framework similar to it. As Figure 3, we perform sparse auto-encoder for the feature extraction on the daily time series of total power of both AMI and HEMS, and cluster these features with K-means clustering so that each HEMS feature is correspond to some ones of AMI nearby in the feature space. Through clustering we may naturally assume that the usage behaviors in the same cluster are similar, so we assign appliance consumption of HEMS as labels to the total power data of AMI in the same cluster, which lack these labels originally. This process is the unsupervised learning stage to obtain the feature extraction models. Hereafter, the sample range for subsequent supervised training is enlarged from 20 HEMS households to 150 AMI households. The main weakness of supervised learning on few samples is overcome.

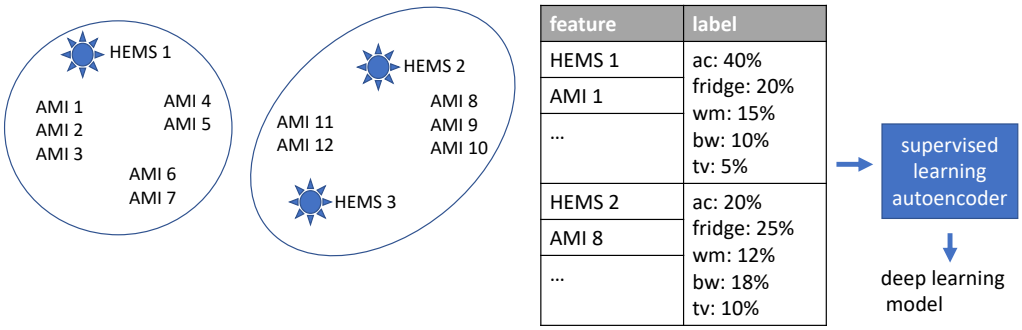


Figure 3: training process for the estimation model. By clustering features in the feature space, we assume features in the same cluster should have similar appliance consumption.

2.2. NILM classification

On the other hand, the second part of this study, classification, is to determine whether each appliance is turned on in a period of time. We divide a day into 3 sections: morning 7:00~12:00, afternoon 12:00~22:00, and night 22:00~7:00 as the classification labels. This partition is designed to match the behavior of common households. We build a multi-label binary classifier to classify the daily usage data. For refrigerators, which are supposed be always turned on, we focus on the period when they are heavily used.

3. EXPERIMENTS

We divide 20 HEMS volunteer households for 5-fold cross validation. In each turn 16 households are for the training data and 4 ones are for the validation. The HEMS data of these 16 households, which is of the sampling period 3 min, is down-sampled to 15 min as the AMI data behaves and used for model training for the 5 appliances as mentioned. The difference of total power between HEMS meters and AMI meters have been calibrated and corrected. The model training process is basically based on the repository 'NeuralNILM' composed by Kelly et al at GitHub (https://github.com/JackKelly/neuralnilm).

3.1. Questionnaire

After the naïve model training of NILM estimation and classification, we have made a simple questionnaire on Google Form about the usage periods. 95% of the AMI users have answered this questionnaire sheet. We did not investigate about the refrigerators since it should be always turned on in the public awareness.

There are 2 issues to be concerned relating to behavioral science. First, it is impossible to know the daily behavior via questionnaire. These answers are to consider the "conjecture" from people toward themselves. People have their different standards on how often they use the appliances when they mark the period as "often used." Second, the answer may be wrong if people are ignorant about how they use their appliances. Therefore, this questionnaire should not be used naively as validation of the model training.

To make a sanity checking, we have compared the questionnaire with the data of HEMS to obtain the precision, as Table 2. We have designed a threshold of days for the classification: an appliance will be marked as "often used" in some period if it is turned on within that period in more days than the threshold days. For example, if someone watches television in 16 nights, then he is marked as "often uses television in the night." We adjust the threshold so that the sum of the precision of the prediction and the questionnaire obtains the maximum.

Table 1: An example of the questionnaire

Columns	**Answer**
User	cpchang@iii.org.tw
Appliances	air conditioner, refrigerator, washing machine, bottle warmer
usage of air conditioners	morning night
usage of refrigerators	night
usage of washing machines	afternoon, night
usage of bottle warmers	morning afternoon, night
usage of televisions	none

Table 2: Comparison between the questionnaire and the prediction toward NILM classification

Appliances	**Precision of model prediction**	**Precision of questionnaire**	**Threshold days of "Often used"**
television	0.73	0.57	15
air conditioner	0.62	0.73	10
bottle warmer	0.61	0.79	15
washing machine	0.89	0.67	6

And we may observe that the questionnaire behaves better than the model prediction for the air conditioners and the bottle warmers. So, our next step is to tune the models for these 2 appliances.

3.2. Tune models via Questionnaire

As Figure 4, we choose to tune our model for AMI classification of the air conditioners and the bottle warmers. First, we train the model of classification from the HEMS and the AMI features. To make up the labels corresponding to AMI features, we impose the data of bottle warmers and the air conditioners from the questionnaire so that these features and labels can be used to train model as well. The obtained classifier model is used to tune the estimation percentage by encouraging or suppressing the ratio weight if it is determined to be turned on or off in this period, respectively. Specifically, if the television is to be determined on in some period with the NILM classification, then we raise the percentage estimation of television by 25% in that period. On the other hand, we lower the percentage estimation of television by 25% if it is determined to be off in that period.

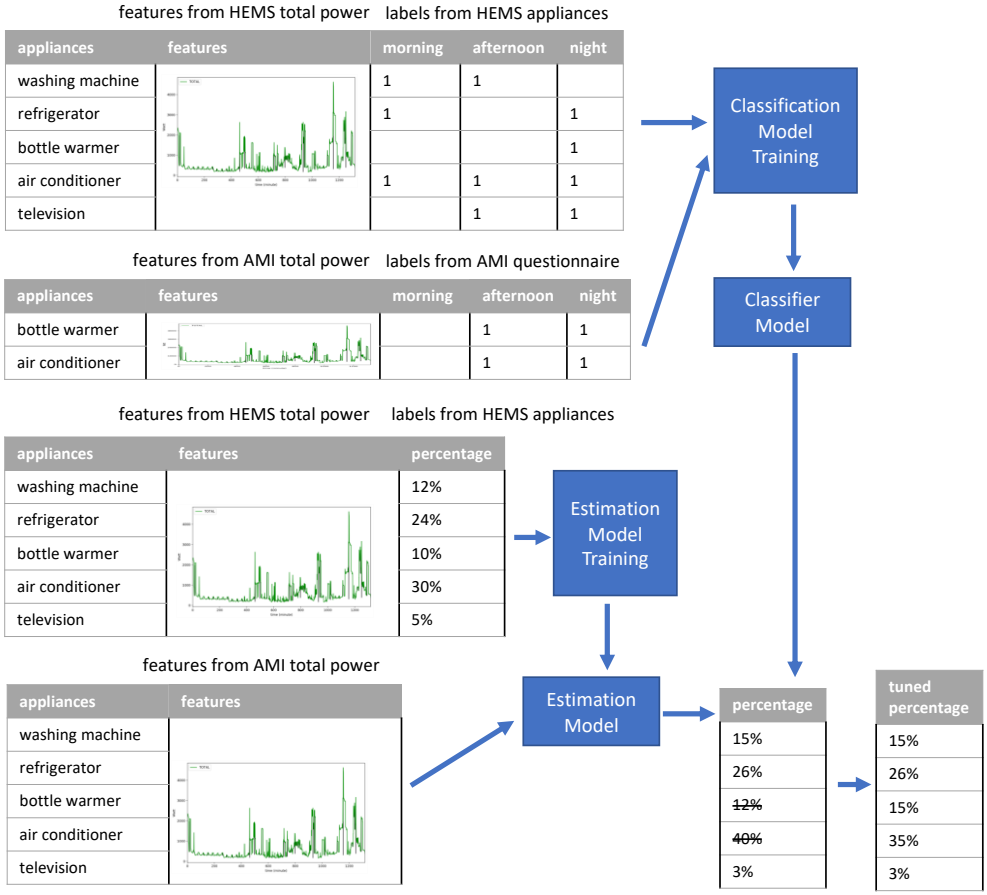


Figure 4: Questionnaire-tuned models for both classification and estimation.

Table 3 shows the prediction result tuned via questionnaire. We impose REITE (relative error in total energy) and the precision to be the measure for estimation and classification, respectively,

$$REITE = \frac{|predicted\ power\ consumption - actual\ power\ consumption|}{\max(predicted\ power\ consumption, actual\ power\ consumption)},$$

$$Precision = \frac{Number\ of\ periods\ in\ which\ this\ appliance\ is\ actually\ turned\ on}{Number\ of\ periods\ in\ which\ this\ appliance\ is\ predicted\ to\ be\ turned\ on}.$$

This improved result implies this low-cost-questionnaire-tuned strategy works.

Table 3: Prediction result

Appliances	**Classification Precision**	**Estimation Relative Error in Total Energy**
refrigerator	0.89	0.06
air conditioner	0.73	0.09
air conditioner	0.62	0.73
bottle warmer	0.61	0.79
washing machine	0.89	0.67

4. CONCLUSIONS

For our empirical study, as Figure 5, we have investigated the NILM estimation and the classification on 150 AMI households in the same community by their smart meters. And we deployed a small amount HEMS smart plugs for the supervised model training. To improve the models, we have made free questionnaire and investigated the reliability of it. We have extracted the reliable part of this questionnaire to modify our own models, and demonstrated this low-priced way works. This strategy costs few compared with thoroughly deployed smart plugs, so we believe it is an efficient way for power suppliers and the government. In September, 2020, we promoted our research in the Presidential Hackathon 2020, Taiwan. And we have won the elite award.

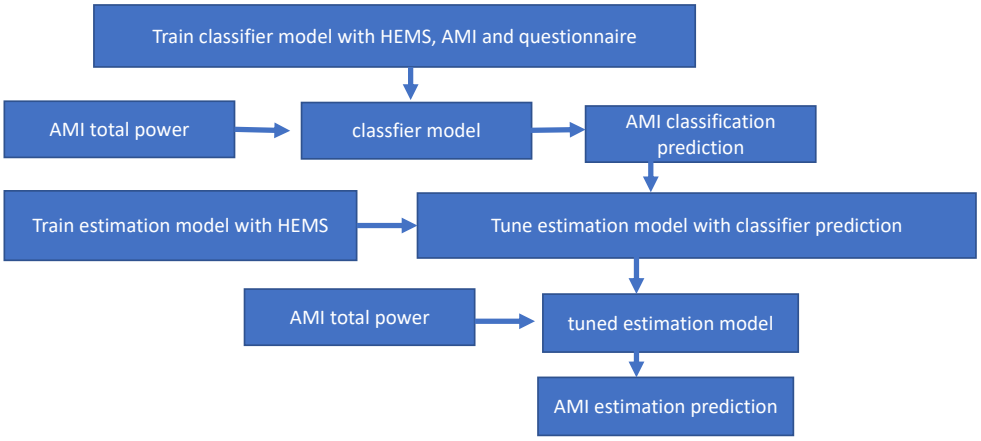


Figure 5: flow chart of questionnaire-tuned AMI classification and estimation

ACKNOWLEDGEMENTS

This work was supported by the Bureau of Energy, Ministry of Economic Affairs, Taiwan under the Grant No. 109-D0101-2.

REFERENCES

[1] G. W. Hart, “Nonintrusive Appliance Load Monitoring,” Proc. IEEE, vol. 80, no. 12, pp. 1870–1891, 1992.
[2] J. Kelly and W. Knottenbelt, “Neural NILM: Deep neural networks applied to energy disaggregation,” BuildSys 2015 - Proc. 2nd ACM Int. Conf. Embed. Syst. Energy-Efficient Built, pp. 55–64, 2015.
[3] R. Bonfigli, A. Felicetti, E. Principi, M. Fagiani, S. Squartini, and F. Piazza, “Denoising autoencoders for Non-Intrusive Load Monitoring: Improvements and comparative evaluation,” Energy Build., vol. 158, pp. 1461–1474, 2018.
[4] K. Basu, A. Hably, V. Debusschere, S. Bacha, J. Dirven, and A. Ovalle, “A comparative study of low sampling non-intrusive load disaggregation,” Oct. 2016.
[5] F. Y. Chang and W. J. Ho, “An Analysis of Semi-Supervised Learning Approaches in Low-Rate Energy Disaggregation,” in Proceedings - 2019 3rd International Conference on Smart Grid and Smart Cities, ICSGSC 2019, 2019, pp. 145–150.

AUTHORS

Chun-peng Chang obtained the Ph.D. degree of physics from National Tsing Hua University, Taiwan in 2014. Now he is a data scientist in the Institute for the Information Industry, Taiwan. His research is the analysis and prediction on energy consumption and generation via machine learning and deep learning.

MATHEMATICAL SIMULATION OF PACKAGE DELIVERY OPTIMIZATION USING A COMBINATION OF CARRIERS

Valentyn M. Yanchuk, Andrii G. Tkachuk, Dmitry S. Antoniuk, Tetiana A. Vakaliuk, and Anna A. Humeniuk

Zhytomyr Polytechnic State University, Zhytomyr 10005, Ukraine

ABSTRACT

A variety of goods and services in the contemporary world requires permanent improvement of services e-commerce platform performance. Modern society is so deeply integrated with mail deliveries, purchasing of goods and services online, that makes competition between service and good providers a key selection factor. As long as logistic, timely, and cost-effective delivery plays important part authors decided to analyze possible ways of improvements in the current field, especially for regions distantly located from popular distribution centers. Considering both: fast and lazy delivery the factor of costs is playing an important role for each end-user. Given work proposes a simulation that analyses the current cost of delivery for e-commerce orders in the context of delivery by the Supplier Fleet, World-Wide delivery service fleet, and possible vendor drop-ship and checks of the alternative ways can be used to minimize the costs. The main object of investigation is focused around mid and small businesses living far from big distribution centers (except edge cases like lighthouses, edge rocks with very limited accessibility) but actively using e-commerce solutions for daily activities fulfillment. Authors analyzed and proposed a solution for the problem of cost optimization for packages delivery for long-distance deliveries using a combination of paths delivered by supplier fleets, worldwide and local carriers. Data models and Add-ons of contemporary Enterprise Resource Planning systems were used, and additional development is proposed in the perspective of the flow selection change. The experiment is based on data sources of the United States companies using a wide range of carriers for delivery services and uses the data sources of the real companies; however, it applies repetitive simulations to analyze variances in obtained solutions.

KEYWORDS

Simulation, Customer Behaviour, Optimization, E-commerce.

1. INTRODUCTION

1.1. Formulation of the Problem.

It is very hard to imagine the contemporary world without planned deliveries, goods, and services ordered online or without scheduled charges for services, etc.

The delivery of e-commerce products has reached unexpected heights in the last few years. Most deliveries made with e-commerce consist of parcels, small packages, and food containers. Forrester Analytics builds the trends that the share of online retail will continue to grow steadily in the next years in the US [1]. Deliveries may have a variety of options like collection points, pickup locations, or direct delivery to the customer location.

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 45-55, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101205

From the perspective of distribution policy, there are two main types of deliveries from the perspective of the full path and last-mile logistics:

Fast delivery (within the same day or next day early morning), where the deadline for delivery passes very fast;

Lazy delivery (optional days for delivery) when the customer has to adapt to specific delivery days which is not necessary the next day or not even the nearest day or the week. Some Lazy deliveries can deliver on Wednesdays or Mondays and Fridays depending on the capacity of the delivery provider.

So, the workload of the online shop team is entirely packed. Every day the web-shop or any other online service operational employee collects items for an order then carefully packs them and sends via delivery service to the end-users. The great ease of this brings the online rate shopping tools that can calculate the costs precisely. With the variety of different vendor's carriers and modes of delivery, the end-user may choose between cost speed and flexibility.

Business to Business (B2B) E-commerce solutions have even more carriers and delivery modes due to the fact they are making delivery of smaller and bigger parcels sometimes even renting the place in a big car so-called Less than Truck Load (LTL) or Greater than Truck Load GTL services. Business to Consumer (B2C) segment has standardized.

At the market of the US, there are a lot of big players like FedEx, UPS, USPS, which makes the majority of deliveries for domestic, interstate, and international deliveries. All these carriers have web APIs that enable quick rate shopping for many e-commerce platforms [2]. At the same time, these big players are not picking bulky deliveries, which are greater than 65 kg, which can be an issue for car parts deliveries, etc. All these constraints are less impacting when you have specified add-ons that use online services helping rate-shop any basket or order and give almost an immediate result with the delivery rate per several shipping options. Nowadays this becomes rather standard to use online APIs that help quicker rate-shop the customers' basket and indicate the price. For better and precise calculation, the basket composition should also keep the delivery address, which should be validated by the delivery service to guarantee the parcel delivery. Dimensional packaging (width × height × depth) are optional but more important for bigger deliveries, which may be reviewed in other manuscripts, due to the different nature of the study.
Below is the simplified scheme of web-shop interaction with an online application integrated service, outlined by authors based on the online experience with the majority of online platforms.

(Step 1) the client forms basket at the online service.

(Step 2a) the basket is getting totals and tax calculated along with the total weight of the delivery (required).

(Step 2b) the dimensional information per product should be indicated per item for the possible use of packaging software calculating the costs for dimensional delivery (optional).

(Step 3) the Shipping origin (the depot or warehouse address) and Delivery address (in full, including Zip-code, city, state, country, Street, and Street 2 addresses should be provided (required).

(Step 4) Online API returns the calculated costs that can be added to the order total and prepared for payment with the full variety of methods and options for delivery.

(Step 5) Online ordering service forms the order for fulfilment and further this order is going for package. Some APIs provide reservation of the tracking number for the delivery, as long as this is finalized and will proceed to the carrier for delivery.

(Step 6) Delivery is scheduled and final delivery time is indicated to the client.

There is specifically highlighted the delivery address and shipping origin address, as they are the key factor for delivery costs calculation and dictate the distance or zone of delivery that plays an important role for the carriers.

Step 3 is an important part of the e-commerce flow, as this information should be dully validated for proper calculation. Besides, the Delivery Address should be recognized by the carrier to validate if there is a delivery to that address. In addition, here is another problem in the investigation: the rural addresses are hardly recognized by carriers, even if they are fully registered addresses with appropriate geolocation. Most carriers are covering specific network locations and points of delivery where the API can calculate the costs. Even the phone call to the carrier does not help much if the operator is seeing the same situation in his system.

There are always debates around the rural delivery areas [3], far distant location with limited delivery services [4] and the current research will not make the edge cases better, and however, the rural areas will certainly be considered. Unlike B2C, which is always dealing with different locations nowadays there are well developed rural zones and distant locations, where many farmers, smaller businesses concentrate their main locations, as these locations real estate and the land is cheaper. However, this does not resolve the problem of delivery, which becomes more and more problematic.

Several works already considered this problem [5] and many times the authors tried to streamline the delivery paths and wanted to go beyond their possibilities. For re-sellers it is important to save on costs as much as possible and keep the very good service. One of the costs, where the re-seller still loses is the transportation to the re-sellers' depots or packing points and then a calculation of the delivery should go further.

As known, the last-mile [6] delivery is currently regarded as one of the most expensive and least efficient portions of the entire supply chain.

This idea was observed in detail by Reyes and Taniguchi [7-8], based on the generalized vehicle routing problem (vendors' truck fleet in the given case) with time windows that have been explored. The study of Reyes, which also considered the earlier publications of urban and rural areas [9] proved that fleet of trucks type of delivery, could reduce total distances up to 40% for an application in the city of Atlanta. This research intends to build on the current state of the art by integrating the notion of travel time uncertainty.

As a very simple and direct solution the fleet of trucks for vendors cruising around the United States and delivery items, but the costs for truck maintenance and payments for the use of the service increases annually.

1.2. Analysis of Recent Research and Publications.

The solution proposed is aimed mainly to optimize the e-commerce processes, help vendors and customers to get their orders in the best and cost-effective way. These intentions are highlighted in various publications directed to technical, economical sciences and a great deal of them still

lies in the aspect of logistics optimization as well as forecasts of the upcoming infrastructural changes.

To cover this multidisciplinary approach let us disclose the existing relationships between e-commerce investigations made for shipments deliveries [3], including domains of domestic deliveries [4], rural areas deliveries [6-8], customer purchasing habits [11], simulation of deliveries in e-commerce systems.

The authors highlighted the approach of integration of delivery [2] were combined with the e-commerce solution with API services provided at the existing market. Overlapping of that work with publications of Routhier (2013) and Morganti and Dablanc (2014) uncovered the city and outside city delivery approach covering the transportation perspective and possible ways of further optimizations in that domain. Authors constantly suggested considering the direct and combined approaches of using the transportation system to optimize the time of delivery, however, the time is not always leading to a cost-effective solution.

Uwe Clausen, Christian Geiger (2016) in their hand-on testing of the last mile concept tried to cover the problems of building the optimal logistic way for larger vehicles using creating the Urban Consolidation Centers. However, this covers only the part of approaches that contemporary e-commerce systems have to consider, and smaller and medium enterprises may not arrange such centers on their own. Besides the approaches reviewed for Europe are not always easily applicable to the US with the higher distances and wider distribution of centers.
Reichheld and Schefter (2000), Abrham et al. (2015), Zelazny (2017) or Ehrenberger et al. (2015) observe that there is a significant relationship between long-term growth of companies' profitability and customer purchase intention [13] however that indicated a good insight that analyzing mid-size companies and trace the turnover and orders circulation it will be easier to identify the dependencies between the options people usually chooses and possible shipment options the current vendors can offer.

For testing data generators many researchers applied Monte Carlo methods, among them, are Sakas, Vlachos, (2014) to create simulations when the mean and variances are known – for our research we can use it to easier generate the experimental data for the current research. The authors decided to apply the Monte Carlo method based on recent researches described in this work as this suits best the nature of the research performed and reflects statistical inference regarding the original sampling domain. For the first step of the simulation, we have taken the original entries of orders accumulated in the test ERP and a very similar distribution will be generated with the Monte Carlo method. Comparison of Monte Carlo with the "what if" method gives better confidence intervals, that will give a better approximation.

2. THE METHOD

In a simplified view, the conceptual framework of the interaction of the re-sellers' eCommerce solution with vendors and delivery networks can be presented as in Figure 1.

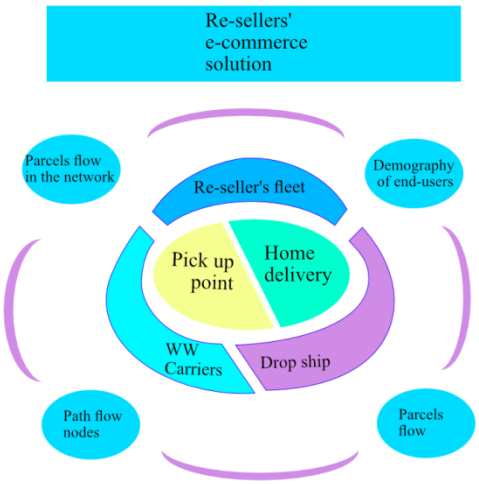


Figure. 1. The conceptual framework for a delivery network, from an operators' point of view. Adapted from [4]

So, in the current investigation, let us generate hundreds of orders based on real addresses of user form the Web-shop database and attempt to:

- Rate-shop those with World Wide Carriers (WW carriers) like UPS and FedEx with the basket compositions to export from sample shops.
- Select Pick-Up delivery from the nearest location (according to [10-11] it still gives a lot of benefits from an economic point of view).
- Use the truck fleet of the vendor for distant deliveries with the truck cost taken per ride. Allocate an option to deliver via partner network, which will be the fleet of the supplier, which anyways deliver the parcels to the online sellers. Drop-ship service is also having costs, but these costs are covered with the Dealer Discount rate, which can be still considered as a benefit. For the given task and simplicity let us set up the discount on delivery costs to 5%.

For a given investigation of the products selected is not that important, as at Step 2a important is the weight of the basket assembled. To mitigate the risk that clients will not be satisfied with generated basket fetching their order – authors filtered out orders that are having too high prices for delivery since that is obvious the customers may give up delivering them.

Dimensional values for products are not important, the focus is at the selection of the delivery per weights and the range of weight will be between 60, which still falls into International deliveries and already interested in LTL service providers and are both good for Own and Supplier's truck fleets.

For the experiment taken 4 US companies where the vendor provides the goods for re-seller, who sell the products to the customer. Having statistical information on these companies the authors can compare the results of simulations with orders that were found and compared in the system via scanning the database and comparing the costs incurred to date. The companies are selected in pares to investigate 2 fleets of re-sellers versus the fleet of vendors for delivery of drop-ship orders.

The analysis of the results will be presented in the context of how much costs the customer saves with the time window not more than 1 day longer than the vendor's truck fleet, which is a guaranteed way of delivery.

3. RESULTS

Using the Monte Carlo method there were generated n orders in 4 networks located in California and distributed across 4 networks S1, R2, R3, and S4 that have the following distribution of clients per with urban and rural zones and percentage of network with the distant rural zone (see the Table 1).

Table 1. Distribution of the network with rural and urban zones

Netw	Classification	Urban zone clients	Rural zone clients	Orders generated for a week	% of Rural distant zone orders
S1	Supplier of R2	210	380	1200	30
R2	Re-seller of S1 and S4	250	320	1800	25
R3	Re-seller of S4	200	180	1300	45
S4	Supplier of R1 and R2	320	260	1500	26

There is no overlapping of clients between networks; however, dependencies of Supplier and Re-seller are given in the column Network Dependencies and orders distributions.

It has been performed 8 simulations (by 2 runs for each type of simulation to calculate the deviation) using Test Dynamics NAV (Enterprise Resource Planning system) custom developed add-on feeding the system with Delivery addresses of a given distribution of Rural Distant zone orders and called the Online and network services for the price calculation. A custom add-on is developed based on Retail Add-on that can trace the system of discounts and particularly evaluate the behavior of the customer. The ease of use of this Add-on got rid of behavioral model validation, as Retail Add-on collects the customer orders in the context of basket composition and the selected shipping method.

Method Monte-Carlo is used to check if the number of the rural distant zone is more than 10% deviation from setting up original parameters as recommended in [11].

When the Monte-Carlo method was applied, the following distribution of the carriers was made across networks (see Table2).

Comparison of orders generated by the Monte-Carlo method shown in the last column of the table.

Table 2. Distribution of the selection for Far rural zone

Carriers	Average price per order with the same weight	S1	R2	R3	S4	% of deviation
FedEx	120 USD	219	73	41	-	12
Vendor Truck	80 USD	301	523	162	400	18
Supplier Truck	75 USD	490	18	212	5	20
PickUP location	20 USD*	130	186	395	265	38
Total	-	1140	800	810	670	-

*The price per pick-up appeared due to the fact of overdue for taking orders from location center
Including into simulation the factor of cancellation of pick-up locations (distributors may potentially cancel delivery to a specific location) and see how the average price impacts the customer if they keep selecting the rest of vendors. The first run shows that customers start selecting either Supplier truck or FedEx, especially if the weight combination comes closer to the FedEx edge weight of 60 kg (See Table 3).

Table 3. Second simulation results overview

Carriers	Average price per order with the same weight	S1	R2	R3	S4	% of deviation
FedEx	100 USD	393	57	432	260	22
Vendor Truck	80 USD	17	48	65	24	18
Supplier Truck	75 USD	730	695	497	386	15
Total	-	1140	800	810	670	

Results were reviewed with the descent analysis of the simulation results and, due to the fact the ERP generation add-on is used, the dependency of the Vendor Truck, which delivers to the pick-up location of Supplier organization, that chain was filtered.

To avoid gaps in the behavioral model it was decided to export data from Google Analytics Extended (supplied by Google Tag Manager) choice preferences versus a list of available vendors and repeat the simulation. It was found that the generation add-on is trying to mimic the web-shop user's behavior in case of absence of Pickup location following the tendency of the user's choice.

As an additional limitation, it was decided to exclude FedEx. At this moment of simulation, it became easier to trace that customers' orders became in the status of Dropped, as some users indeed were leaving web-shops not finding the guaranteed delivery provider. According to [12] the variety of delivery methods for the B2C segment should have a wider range of prices for delivery, then the choice of the customer will be either in favor or cheapest or most reliable carrier (personal user preference). However, B2B users in this simulation behaved exactly like B2C customers, preferring to abandon the basket or order and leave, rather than select the cheapest option. However, another tendency was noticed, that abandoned baskets were noticed from the users, who reached to the limit of the price they are ready to pay; thus, the factor of the economy played an important role. The last attempt of simulation taken with the assumption that the customer may use the Vendor's truck for the same price as the supplier, but the additional day may be added to the total route time (see Table IV).

Table 4. Forth simulation results overview

Carriers	Average price per order with the same weight	S1	R2	R3	S4	% of deviation
Vendor Truck	75 USD	621	729	549	493	20
Supplier Truck	75 USD	519	71	261	177	22
Total	-	1140	800	810	670	

As it is seen from the results the clients behaved more reluctant to the delay in 1 day, however, they selected the home delivery via the Vendor's truck, rather than the supplier. I assume this happened due to the price change and most probably, no competence with the supplier appeared a key factor for Vendors' truck network. This also proves the statements described in [12] that users ordering bigger deliveries are usually more reluctant to have guaranteed delivery, rather than

have it fast, but the tendency is equivalent for both: B2B and B2C segments as it is also indicated in [13].

The current investigation is complete, however during the investigation, it was identified, that simulation of orders creation with the only ERP data without analytical data will only indicate the quantitative part, without the qualitative part that can be filled with additional statistical data per customer on personal preferences of choices in different cases of orders created in the live systems and presence or absence of the carrier in certain circumstances. Thus, adding Google Analytics Extended data on user choice allowed us to complete the investigation.

Google Analytics covered the validation of Users Acceptance Testing (UAT) and reflection of their behaviour for generated orders. Additionally, the use of Google Analytics given an insight into the comparison of generated orders and directed to UAT for different groups of users during the application testing and run-down tests performed in all 4 companies.

Table 5, indicating the coverage of orders by comparison of generated orders and directed to UAT with orders recently submitted in the system (completed orders) will serve as the validation baseline for the approach suggested.

Table 5. Coverage of the real orders for Far rural zones* that correspond to generated orders per simulation

Carriers	**Average price per order with the same weight**	**S1, %**	**R2, %**	**R3, %**	**S4, %**
FedEx	120 USD	92	92	86	75
Vendor Truck	80 USD	90	87	95	65
Supplier Truck	75 USD	89	68	73	85
PickUP location	20 USD*	76	64	86	92
Average	-	87.65	77.75	85	79.25

* Far Rural zones are distantly located areas having lower delivery capacities, mostly implemented by private delivery networks and lower flows of international carriers.
As we see, generated orders have rather high coverage from the data perspective and lie in the frame of deviations we calculated recently (Table 6.).

Table 6. Coverage of the real orders for Domestic zones that correspond to generated orders per simulation

Carriers	**Average price per order with the same weight**	**S1, %**	**R2, %**	**R3, %**	**S4, %**
FedEx	120 USD	97	92	95	97
Vendor Truck	80 USD	90	92	90	96
Supplier Truck	75 USD	95	88	89	87
PickUP location	20 USD*	90	89	86	92
Average	-	93	90.25	90	93

A higher level of coverage for orders generated for domestic zones is relatively easy to explain from the perspective of a higher number of orders generated for domestic zones.

4. CONCLUSIONS

The series of simulations performed given a possibility to evaluate the behaviour of users in case of absence of usual carriers and taking the decision of cheaper solution It also demonstrates options the web-shop owner may offer to the mid and small-size businesses alternative delivery services, where the drop-ship of vendor delivers to the final destination or drop-point with minimal costs if the additional discount is given to keep the order fulfillment. As a continuation of this work can be developed and on-line web-service to support delivery network visualizing deliveries of vendors and re-sellers, where all packages can be loaded into the truck, correct the route, and deliver goods and services with Vendor's fleet, as long as the route and discount permits.

To apply the current solution to the real industrial case will involve the expansion of the Vendor and Re-seller networks sharing the same data model for orders and street validation. In the case of contemporary ERP systems use that should not be a very difficult problem, however, it may involve additional 3-rd party services.

The paper highlighted the combination of approaches of drop-ship deliveries and selection of different routes for total costs optimization and indicated that behaviour of the client is not always driven by economic circumstances, but also by habits and tendencies of customers, so well described in [7].

For future investigations, authors will involve Google Tag Manager data, collected for customers, as the import of options for checkout selection helped discover how customers may potentially abandon basket if they do not see the option of the habitual carrier or the price for the order versus delivery cost exceeds a specific limit.

REFERENCES

[1] Forrester Analytics (2018). Forrester Online Retail Forecast, 2018 to 2023 (US).

[2] Yanchuk V. Gumenyuk A., Tkachuk A integrated add-ons for shipment providers and their connection to the e-commerce solution – Proceedings of II International scientific-practical conference "Computer technologies: innovations, problems, and solutions". – Zhytomyr, – 2017, 17-18

[3] Routhier, J. L. (2013). French cities' urban freight surveys. City logistics research: A transatlantic perspective. Conference proceedings 50 Summary of the First EU-US Transportation Research Symposium. (pp.9–14). Washington, DC: Transportation Research Board of the National Academies. doi:10.1108/IJPDLM-01-2016-0008

[4] Morganti, E., Dablanc, L., Fortin, F., 2014. Final deliveries for online shopping: The deployment of pickup point networks in urban and suburban areas. Research in Transportation Business & Management, 11, 23-31. doi: 10.1016/j.rtbm.2014.03.002

[5] Motte-Baumvol, B., et al. / Asian Transport Studies, 4(3) (2017). doi: 10.11175/eastsats.4.585

[6] Song, L., Cherrett, T., McLeod, F., Wei, G., 2009. Addressing the last mile problem. Transport impacts of collection and delivery points. Transportation Research Record: Journal of the Transportation Research Board, 2097, 9-18. doi: 10.3141/2097-02

[7] Reyes, D., Savelsbergh, M., Toriello, A., "Vehicle routing with roaming delivery locations," Transp. Res. Part C Emerg. Technol., vol. 80, pp. 71–91, 2017. doi: 10.1016/j.trc.2017.04.003

[8] Visser, J., Nemoto, T., & Browne, M. (2013). Home delivery and the Impacts on the urban freight transport: A review. Urban areas recent advances in city logistics: Proceedings of the VII international conference on city logistics, Bali, Indonesia, June 17–19 (pp. 14–31). doi: 10.1016/j.sbspro.2014.01.1452

[9] E. Taniguchi, R.G. Thompson, T. Yamada, ed by E. Taniguchi, T. Fang Fwa and R.G Thompson (CRC Press, 2014), p. 1

[10] Weltevreden, J.W, 2008. B2c e-commerce logistics: the rise of collection-and-delivery points in The Netherlands. International Journal of Retail & Distribution Management, 36, 8, 638- 660. doi: 10.1108/09590550810883487
[11] Schewel, L., & Schipper, L. (2012). Shop 'till we drop: A historical and political analysis of retail goods movement in the United States. Environmental Sciences Technology, 46–18, 9813–9821. doi: 10.1021/es301960f
[12] Accenture (2015). Adding Value to Parcel Delivery. www.accenture.com Accessed on 9 Jan 2017.
[13] Roudposhti, V.M., Nilashi, M., Mardani, A., Streimikiene, D., Samad, S., & Ibrahim, O. (2018). A new model for customer purchase
[14] Intention in e-commerce recommendation agents. Journal of International Studies, 11(4), 237-253. Doi:10.14254/2071-8330.2018/11-4/17

AUTHORS

Valentyn Yanchuk, Associate Professor at the Department of Automation and Computer-Integrated Technologies named after Prof. B.B. Samotokin, Zhytomyr Polytechnic State University, Zhytomyr, Ukraine. Valentyn Yanchuk, born in 1975, received a Candidate of Technical Sciences degree (Ph.D.) from the Pukhov Institute for Modelling in Energy Engineering, National Academy of Sciences of Ukraine (IMPE) in 2002. Since 1997, he has been working in the field of Software Engineering and Information Technologies at the Zhytomyr Polytechnic State University, where he is currently working at the position of Associate Professor and in the practical field of IT business. His research interests include software engineering, business analysis, computer-based modeling, E-Commerce. He has published several papers and proceedings in the journals and material of conferences.

WWW: https://ztu.edu.ua
E-mail: v.yanchuk@gmail.com

Assoc. Prof. Andrii Tkachuk, Head of the Department of Automation and Computer-Integrated Technologies named after prof. B.B. Samotokin, Zhytomyr Polytechnic State University, Zhytomyr, Ukraine. Andrii Tkachuk, born in 1989, received a Candidate of Technical Sciences degree from the National Technical University of Ukraine "Kyiv Polytechnic Institute", Ukraine, in 2014. Since 2012, he has been working in the field of information technologies, automation, and robotics at the Zhytomyr Polytechnic State University. Andrii Tkachuk is an Expert of the Scientific Council of the Ministry of Education and Science of Ukraine, a Board member of NGO “Youth integration center”. His research interests include information technologies, automated aviation gravimetric systems, mobile robotics, armament stabilization systems. He has published several papers in international journals, is a reviewer of The scientific journal Aviation which is included in the Scopus database.

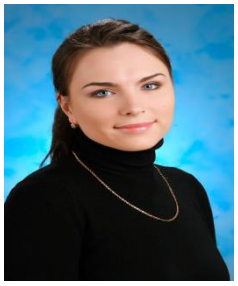
Www: https://ztu.edu.ua/ua/structure/faculties/fikt/kakt.php
e-mail: andru_tkachuk@ukr.net

Dmytro Antoniuk, Assistant Professor of the Department of Software Engineering, Zhytomyr Polytechnic State University, Zhytomyr, Ukraine. Dmytro Antoniuk, born in 1981, received a Candidate of Pedagogical Sciences degree (Ph.D.) from the Institute of Information Technologies and Learning Tools, Ukraine, in 2018. Since 2003, he has been working in the field of Software Engineering and Information Technologies at the Zhytomyr Polytechnic State University, where he is currently an Assistant Professor of the Department of Software Engineering and in the practical field of IT business. His research interests include software engineering, business in IT, computer-based business-simulation, economic and financial literacy of technical professionals. He has published several papers and proceedings in the journals and material of conferences.

WWW: https://ztu.edu.ua
E-mail: Dmitry_antonyuk@yahoo.com

Dr. Tetiana Vakaliuk, professor of the Department of Software Engineering, Zhytomyr Polytechnic State University, Zhytomyr, Ukraine. Tetiana Vakaliuk, born in 1983, received a Candidate of Pedagogical Sciences degree from the National Pedagogical Dragomanov University, Ukraine, in 2013, and a Doctor of Pedagogical Sciences degree from the Institute of Information Technologies and Learning Tools of the National Academy of Sciences of Ukraine, in 2019. Since 2019, she has been working in the field of information technologies at the Zhytomyr Polytechnic State University. Her research interests include information technologies, ICT in Education, Cloud technologies. She has published several papers in international journals, is a member of editorial boards of Information Technologies and Learning Tools, Zhytomyr Ivan Franko State University Journal: Pedagogical Sciences, Collection of Scientific Papers of Uman State Pedagogical University.

WWW: https://sites.google.com/view/neota
e-mail: tetianavakaliuk@gmail.com



Anna Humeniuk Associate Professor at the Department of Automation and Computer-Integrated Technologies named after Prof. B.B. Samotokin, Zhytomyr Polytechnic State University, Zhytomyr, Ukraine. Anna Humeniuk, born in 1986, received a Candidate of Technical Sciences degree (Ph.D.) from the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" in 2011. Her scientific interests include automated measuring systems, gravimetry, flexible manufacturing system design automation. She has published several papers and proceedings in the journals and material of conferences.

WWW: https://ztu.edu.ua
E-mail: gum_ann@ukr.net

AN INTERNET-OF-THINGS APPLICATION TO ASSIST THE DETECTION OF FALLING TO THE GROUND

Yifei Yu[1], Yu Sun[2], Fangyan Zhang[3]

[1]Sage Hill School, Newport Coast, CA, 92657
[2]California State Polytechnic University, Pomona, CA, 91768
[3]ASML, San Jose, CA, 95131

ABSTRACT

As people get old, the risk of them falling increases; the fall will impact senior citizens more negatively than younger people. My grandmother once fell and hit her when she was alone at home, and she instantly became unconscious. Frequently, senior citizens are unable to help themselves after they fall, even if they remain conscious. However, there isn't a product that senior citizens can use to notify their relatives right away if they fall, and this leads to the question of how we can bring immediate aid to all senior citizens after they fall. This paper brings forward the product and software that can solve this problem. The product is a small wristband that detects any falls or collisions and notifies relatives right away. The software is an accompanying app that shows the data recorded from those falls or collisions, specifically designed for family members to keep track of their elders. We applied our application during our test sessions and conducted a qualitative evaluation of the approach. The results show that this experiment is a great solution to our problem, but with a few limitations and weaknesses.

KEYWORDS

Detection of falling, wristband, iOS, Android

1. INTRODUCTION

Falls are dangerous and costly. As people get old, the risk of falling increases; the fall will impact senior citizens more negatively than younger people. My grandmother once fell and hit her when she was alone at home, and she instantly became unconscious. No one in the family knew about this until someone arrived home and found that she fainted. This was a severe example, but senior citizens are unable to help themselves after they fall, even if they remain conscious. For instance, they might be unable to stand back up or call out loud. According to the Centers for Disease Control and Prevention (CDC), millions of older people—those 65 and older—fall every year, and one out of five falls causes a severe injury such as broken bones or a head injury [1]. Without immediate treatment of the patient, these injuries will only get worse and may even lead to death. For this reason, I created a small wristband that can detect an individual's falls and notify his or her relatives on the spot.

My mission is to help senior citizens receive the aid instantly after they fall, especially when no one is around. I designed a small, foldable wristband that can be worn around wrists or put into the pockets. It contains an accelerometer that detects a collision and records its magnitude, a GPS that marks the location, a clock that records the time of the fall, and a SIM card that can send these statistics to the phone and notify family members. Accompanying the wristband is an

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 57-65, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101206

application compatible with IOS and Android devices. It contains a built-in Google Map that shows the location of the last fall and its magnitude and time. Furthermore, aside from showing the latest data, it stores all of the past data and presents them on a Full Statistics screen. In the event of a fall, the accompanying app will notify the family members and show the fall data.

In this way, I strive to bring aid to senior citizens quickly after their accidents in the short run and aspire to reduce the number of unprotected accidents in the long term. My mission is to prevent senior citizens from being unattended after they fall by providing them immediate aid.

The market for fall detection and faster medical help is growing quickly due to the influx of new technologies [2]. Less than ten companies are developing or researching development of this kind of product (elderly fall sensors/help) [14][15]. According to some market research, I discover that some of the competing companies in the elderly care industry use techniques and systems that have been proposed to act as fall alert systems [3] [4]. These companies also sell fall alert systems to the elderly, in the form of bracelets that act as fall detectors. In the case of a fall, the bracelet contains a button that allows the user to call for help. Other companies offer their product with a similar function, but it involves the user physically typing a text message to the company's emergency help center to receive immediate aid. However, these proposals assume that the elderly are still conscious and able to move after they fall, which is rarely the case in practice. Furthermore, the products competing companies offer are not very user friendly due to their complicated user-interface, whether it is on the bracelet itself or on the accompanying phone application. The last problem is the battery life. According to my research, a company called LifeFone actually has very similar functions to the product I am developing [5]. However, their device (bracelet) only has a battery life of 5 days.

We have developed innovative measures to ensure the safety of senior citizens: an alert system that can bring aid to senior citizens in a quick and cost-saving way. The alert system has two components: hardware and software.

The goal of the application was to help senior citizens in case they fall and are unable to help themselves by immediately notifying family members. It should prevent the senior citizens from suffering too long from their accidents by bringing help to them right away. Requirements or Criteria included simplicity, user-friendliness in both APP and product (wristband), precision in location, magnitude, and time detected by the product (wristband) and its durability. The constraint included weak signal detection of the product (wristband) at some locations and its durability.

Our experiment includes the following materials: Particle Electron 3G-U260; Ublox SARAU260: GPS that pinpoints the location of the wristband with accelerometer attached to detect collisions or any forceful physical impacts; Google Maps: a map that shows the location of the wristband; Arduino: computing platform to program the wristband; and Thunkable: computing platform to code mobile APP[6] [7].

In our experiment, we first attached the Ublox SARA-U260 (GPS and accelerometer) onto the Particle Electron 3G-U260 (asset tracker), which made the hardware or wristband. Then, we linked the hardware to the Particle Console Web IDE to program. Afterwards, we coded the hardware in Arduino so that it showed the A, G, and Accel variables. The [A] variable shows a forceful physical impact upon the hardware. Such impact is above the safety threshold and would be considered a fall or collision, [G] variable shows the location of the fall. It provides the latitude and longitude of the GPS, and the [Accel] variable shows any physical impact that is below the safety threshold of the [A] value. Finally, we programmed the APP using Thunkable to show the location, time, and magnitude of the collisions. It includes a map (Google Maps) on

the top half of the screen and most recent fall statistics. The past falls are recorded into the cloud server.

Two experiments have been conducted to verify the following two aspects of the system. In Experiment 1, we tested the accuracy of the falling detection using different algorithms. The core falling algorithms rely on the accelerometers embedded in the system. We have tested the different machine learning algorithms to classify the falling status based on the sampled accelerometers values, such as SVM, RandomForest, and Decision Trees [9][10][11]. It turns out that RandomForest has the highest accuracy of 94.5%.

In Experiment 2, we tested the accuracy of the falling detection using different parameters. We tested the training dataset collected using different combinations of the accelerometer values such as (x), (y), (z), (x, y), (x,y,z). Although most of the accuracy is similar to each other, the 3value combination (x,y,z) is on average higher than the rest.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the experiment, including weak signals, the need to refresh data, and the lack of a button; Section 3 focuses on the details of our work including screenshots of the code and corresponding explanations; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

2.1. Challenge 1: Weak Signal

Weak signals or connections have always been part of a challenge in this project. During testing sessions of the prototype (the bracelet), the data of the prototype sometimes doesn't show up on our computer screens. In addition, there have sometimes been flickering green lights on the prototype device, showing that our connection is weak. As a result, when we test the prototype by slamming it, the magnitude, location, and time of the collision doesn't show up on our computer screens, suggesting that the data is not received. This problem can be especially dangerous in the real world. If the user falls and no fall data is tracked, the user will not receive any aid and will suffer the consequences of the fall. Therefore, it is necessary for us to improve the signal and connection between the hardware (bracelet) and the API, which store and transfer the data, so that the high accuracy can protect our users.

2.2. Challenge 2: Need to Refresh Data

We test our prototype by slamming it against the table to resemble the user falling to the ground in real life. During our testing sessions, we come to realize that we cannot slam the prototype too often. In other words, if we slam it once at 1:00, we cannot slam it again 5 seconds later. That is because the accelerometer within the hardware is slow to record the data of the second fall. As a result, immediately slamming the prototype again after the first slam will not yield any data. What is even worse is that when we test it again minutes after the first test, the data still may not show up. As a result, we have to manually press the button on the side of the Particle Electron to refresh, and that's when it starts picking up data from our tests again. However, the user shouldn't have to press the refresh button at all in real life, because the device should be able to refresh itself and record all the necessary data.

2.3. Challenge 3: Lack of a Button

Our prototype, most of the time, is able to record the relevant data of the fall, including the magnitude, time, and location. Assuming that all of these features are recorded 100% of the time, the device still hasn't reached its full potential to help the user. It lacks a button. If the user hasn't fallen, but isn't feeling well, there should be a button he or she can press to immediately acquire help.

Standard A4 (210mm x 297mm) portrait page set-up should be used. The left, right, top and bottom margins should be 30mm. Do not use any headers, footers or footnotes. No page numbers. Single column. All main text paragraphs, including the abstract, must be fully (left and right) justified. All text, including title, authors, headings, captions and body, will be Times New Roman font.

3. SOLUTION

The alert system has two components: hardware and software. The hardware is a foldable wristband that can be worn on the wrist or put into the pocket. Inside, there is an accelerometer that detects a collision and records its magnitude, a GPS that marks the location of the fall, a clock that records the time of the fall, and a transmitter that can send these statistics to the phone and notify family members via the app. A safety threshold is programmed into the accelerometer to ensure that not every movement is considered a collision. The safety threshold, in other words, is an experimentally-tested number (16,000) that distinguishes the difference between a regular movement and an accident. Any recorded magnitude below the threshold (15,700, for example) is considered safe, and the device will report it to the app. On the contrary, any recorded magnitude above it (18,000, for example) is a fall, and the device will record its data along with the other relevant information. The corresponding phone app will notify family members immediately.

The software of the system is the accompanying app, a free download application that is compatible with IOS and Android devices. Family members can download it to track the elderly's locations and falls. Opening its home screen, the top half of the user interface is a built-in Google Map that shows the location of the latest fall and its magnitude and time. Aside from illustrating the latest data, it stores the past data on a Full Statistics screen.

The hardware and software work together to create an efficient system. During a collision, the hardware detects all the information and sends it to the app. The app will send notifications to the family members' phones through vibration and on-screen texts. We will sell our alert systems through online sales from selling platforms, and through direct sales with senior care centers. We believe that developing a cheap and efficient alert system will minimize the dangerous implications of a serious fall.

In the case of a fall, the bracelet will detect the location, time, and magnitude of the fall. All three pieces of the data will be transferred onto the app, which notifies family members and shows all the data transferred. This app has already been published and families can download it from both iOS and Android stores to keep track of their elders. Looking at the app, the top-half of the screen is a built-in Google Map which shows the location, bottom-left corner is the time, and bottom-right corner is the magnitude. For our purposes, we designed the magnitude from a scale from 1-5. 1 being the lightest fall and 5 being the heaviest.

Because we initialized accelThreshold, we can use it as a reference for reading accelerometer data. We set “if” statement at the beginning so that the device will only publish data if t.readXYZmagnitude() is greater than our defined safety threshold of 16,000. Any value greater than that threshold will show, because it is considered a fall; any value less than the threshold will not show, because they are regular movements the user makes, such as walking or running, which report a safe magnitude. The “if” statement that follows checks if transmittingData is true, and then publishes using the Particle.publish(“<readable name>”, <value>, 60, PRIVATE) function to output the value given a readable name. Then, the “if” statements underneath check if the GPS has a fixed point, and will proceed to transmit data if the fix is confirmed. Upon transmitting, the process of token delimiting occurs inside of the “for” loop and “while” loop and separates the Latitude and Longitude into two separate values for the gps_coords[] array.
Particle.publish(“<readable name>”, <value>, 60, PRIVATE) is run once again to publish the separated gps values.

```
void loop() {

    if (t.readXYZmagnitude() > accelThreshold) {
        // Create a nice string with commas between x,y,z
        String pubAccel_x = String::format("%d", t.readX());
        String pubAccel_y = String::format("%d", t.readY());
        String pubAccel_z = String::format("%d",t.readZ());
        String A_mag= String::format("%d", t.readXYZmagnitude());

        // Send that acceleration to the serial port where it can be read by USB
        Serial.println(pubAccel_x);
        Serial.println(pubAccel_y);
        Serial.println(pubAccel_z);
        Serial.println(t.readXYZmagnitude());

            // If it's set to transmit AND it's been at least delayMinutes since the last one...
        if (transmittingData) {
            lastPublish = millis();
            Particle.publish("A_mag", A_mag, 60, PRIVATE);
            Serial.println(A_mag);
            }
    }

    t.updateGPS();


    // GPS requires a "fix" on the satellites to give good data,
    // so we should only publish data if there's a fix
    if (t.gpsFix()) {
        // Only publish if we're in transmittingData mode 1;
        if (transmittingData) {
            // Short publish names save data!
            string = t.readLatLon();
            for(int j = 0; j < strlen(string); j++){
                LL[j] = string[j];
            }
            char * LatLon = strtok(LL, ",");
            int i = 0;
            while( LatLon != NULL ) {
                gps_coord[i] = LatLon;
                LatLon = strtok(NULL, delim);
                i++;
            }

            Particle.publish("G_lat", gps_coord[0], 60, PRIVATE);
            Particle.publish("G_lon", gps_coord[1], 60, PRIVATE);
        }
        // but always report the data over serial for local development
        //Serial.println(gps_coord[0]);
        //Serial.println(gps_coord[1]);
    }
```

Figure 1: key code in implementation

Figure 2: result from change in magnitude

The visual above depicts the change in magnitude with respect to time detected by the device during test falls.

Figure 3: result from change in latitude and longitude

The two visuals above show the change in latitude and longitude detected by the device in respect to time during a test fall. Both data are screenshots from ThingSpeak, our API.

4. EXPERIMENT

Our experiment includes the following materials: Particle Electron 3G-U260; Ublox SARAU260: GPS that pinpoints the location of the wristband with accelerometer attached to detect collisions or any forceful physical impacts; Google Maps: a map that shows the location of the wristband; Arduino: computing platform to program the wristband; and Thunkable: computing platform to code mobile APP.

In our experiment, we first attached the Ublox SARA-U260 (GPS and accelerometer) onto the Particle Electron 3G-U260 (asset tracker), which made the hardware or wristband. Then, we linked the hardware to the Particle Console Web IDE to program. Afterwards, we coded the hardware in Arduino so that it showed the A, G, and Accel variables. The [A] variable shows a forceful physical impact upon the hardware. Such impact is above the safety threshold and would be considered a fall or collision, [G] variable shows the location of the fall. It provides the latitude and longitude of the GPS, and the [Accel] variable shows any physical impact that is below the safety threshold of the [A] value. Finally, we programmed the APP using Thunkable to show the location, time, and magnitude of the collisions. It includes a map (Google Maps) on the top half of the screen and most recent fall statistics. The past falls are recorded into the cloud server.

Two experiments have been conducted to verify the following two aspects of the system. In Experiment 1, we tested the accuracy of the falling detection using different algorithms. The core falling algorithms rely on the accelerometers embedded in the system. We have tested the different machine learning algorithms to classify the falling status based on the sampled accelerometers values, such as SVM, RandomForest, and DecisionTrees. It turns out that RandomForest has the highest accuracy of 94.5%. In Experiment 2, we tested the accuracy of the falling detection using different parameters. We tested the training dataset collected using different combinations of the accelerometer values such as (x), (y), (z), (x, y), (x,y,z). Although most of the accuracy is similar to each other, the 3-value combination (x,y,z) is on average higher than the rest.

5. RELATED WORK

The first work is done by a company called Guardly [12]. It created a simple app that allows users to communicate whether or not they require assistance. They have also created a hardware device, but it has no fall detection. Our project differs from Guardly in that our app is more integrated to the users' needs. For example, our hardware alone is able to detect the fall of its users and automatically contacts family members for assistance. Furthermore, it automatically provides the fall statistics, including the location, time, and magnitude. Guardly, in contrast, requires users to manually type on the app to let people know that they need help. Although they have a good customer service center that receives users' call for assistance, their app adds inconvenience for the users.

Guardly is strong in that it has a clean-looking app. In other words, it has little to no bugs and sends users' message to their customer service every time. This is our weakness because sometimes our app glitches, so pressing a button may not do anything; the app will not receive the command or signal. Guardly is weak in that their concept is too simple: it contains only an app and a simple device. The device is almost useless as it performs the same function as the app, which is for users to call for help when they are not feeling well or when they fall. This field is our strength because our hardware can automatically detect users' fall and contact help (family members) for the users.

The second project/company we researched is Kitestring [8]. The company's concept is also very simple, as it sends automated SMS to the user's device (also a bracelet) and checks if they respond. If the user doesn't, it tells personalized emergency contacts. Our work is different from that of Kitestring in that we have both hardware and software, while Kitestring only has a hardware. Kitestring also requires users to manually respond to texts while our work does not require anything from the users. Compared with Kitestring, our strength is that we have both an app and hardware; it requires no customer service center which Kitestring has. In other words, Kitestring requires users to manually input a list of emergency contacts, and its customer service will send automated SMS to check on the user. If the user doesn't respond, it will notify the user's emergency contacts. It is so complicated. Our work, on the contrary, is very simplistic and user friendly.

Guardly's work has a portable fall detector that includes a GPS with Wifi, and can be worn as a necklace [13]. It also has a button that users can press in order to contact emergency help. Out of all works, Guardly has the most similar work to ours. However, one difference is that it also has a customer service center, which the device calls in case the user needs emergency assistance. For us, our device calls family members rather than the customer service. Since Guardly's work is very similar to ours, its strength is that it has a very strong and well functioning app and device. Because Guardly is already an established company, they do a very good job with minimizing glitches, so all data and messages are successfully transmitted between the user to

the API and to the customer service. This is our weakness because our system is full of glitches that need to be fixed in the meantime.

6. CONCLUSION AND FUTURE WORK

To solve the problem of senior citizens being unattended in the case of a fall, we developed a system that involved both hardware and software. The hardware was a bracelet for senior citizens that tracks the magnitude, time, and location of the fall. These data would be recorded on an API, which would also be transferred to the accompanying phone app that would show the data for family members. To test the accuracy of the alert system, especially the hardware, we ran two experiments. In the first experiment, we tested the accuracy of the falling detection using different algorithms SVM, RandomForest, and DecisionTrees. It turned out that RandomForest had the highest accuracy of 94.5%. In the second experiment, we tested the accuracy of the falling detection using different parameters. We tested the training dataset collected using different combinations of the accelerometer values. Although most of the accuracy is similar to each other, the 3-value combination is on average higher than the rest.

Currently, we haven't achieved the maximum amount of accuracy in our system due to weak connections between the device and API, and a weak refreshing system. Both of these limitations contribute to a less accurate system because data is not recorded at times.

The current practicality of the system is not very high due to the size of the prototype. Because the prototype is very big and bulky, it will definitely be too big for anyone to wear around their wrist once it's incorporated into a bracelet. Therefore, it is necessary to minimize the size of the physical components that record the data, so the bracelet will be small and convenient for all users. To maximize the usefulness the bracelet can bring for its users, a button must be added. Even if the user hasn't fallen, he or she can still press the help button on the bracelet if the user isn't feeling well, and will receive immediate help.

In the future, we plan to improve the accuracy of the system by boosting the signal connections between the device and API, strengthen the practicality by decreasing the size of the bracelet, and enhance the optimization by adding a "help" button. We strongly believe that by following these measures, our alert system will be able to reach its full potential.

REFERENCES

[1] Important Facts about Falls. 10 Feb. 2017, www.cdc.gov/homeandrecreationalsafety/falls/adultfalls.html. Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.
[2] Mubashir, Muhammad, Ling Shao, and Luke Seed. "A survey on fall detection: Principles and approaches." Neurocomputing 100 (2013): 144-152.
[3] Wu, Falin, et al. "Development of a wearable-sensor-based fall detection system." International journal of telemedicine and applications 2015 (2015).
[4] Mubashir, Muhammad, Ling Shao, and Luke Seed. "A survey on fall detection: Principles and approaches." Neurocomputing 100 (2013): 144-152.
[5] Matias, Igor, Nuno Pombo, and Nuno M. Garcia. "Towards a Fully Automated Bracelet for Health Emergency Solution." IoTBDS. 2018.
[6] Arduino, Store Arduino. "Arduino." Arduino LLC (2015).
[7] Levy, Paul Blain. "Thunkable implies central." (2020).
[8] Kitestring, www.kitestring.io/.
[9] Do, Thanh-Nghi, et al. "Classifying very-high-dimensional data with random forests of oblique decision trees." Advances in knowledge discovery and management. Springer, Berlin, Heidelberg, 2010. 39-55.

[10] Burrell, Jenna. "How the machine ‘thinks’: Understanding opacity in machine learning algorithms." Big Data & Society 3.1 (2016): 2053951715622512.
[11] Michie, Donald, David J. Spiegelhalter, and C. C. Taylor. "Machine learning." Neural and Statistical Classification 13.1994 (1994): 1-298.
[12] Beta, Brodie. “Guardly: An IOS App That May Save Your Life.” The Next Web, 7 Apr. 2011, thenextweb.com/apps/2011/04/08/guardly-an-ios-app-that-may-save-your-life/.
[13] IMANI, ANITA. "Design and development of a user interface for a mobile personal indoor navigation assistant for the elderly." (2014).
[14] El-Bendary, Nashwa& Tan, Qing & Pivot, Frederique & Lam, Anthony. (2013). Fall detection and prevention for the elderly: A review of trends and challenges. International Journal on Smart Sensing and Intelligent Systems. 6. 1230-1266. 10.21307/ijssis-2017-588.
[15] Chaudhuri, Shomir et al. “Fall detection devices and their use with older adults: a systematic review.” Journal of geriatric physical therapy (2001) vol. 37,4 (2014): 178-96. doi:10.1519/JPT.0b013e3182abe779

CRICTRS: EMBEDDINGS BASED STATISTICAL AND SEMI SUPERVISED CRICKET TEAM RECOMMENDATION SYSTEM

Prazwal Chhabra, Rizwan Ali and Vikram Pudi

International Institute of Information Technology, Hyderabad, India

ABSTRACT

Team Recommendation has always been a challenging aspect in team sports. Such systems aim to recommend a player combination best suited against the opposition players, resulting in an optimal outcome. In this paper, we propose a semi-supervised statistical approach to build a team recommendation system for cricket by modelling players into embeddings. To build these embeddings, we design a qualitative and quantitative rating system which considers the strength of opposition also for evaluating player's performance. The embeddings obtained, describes the strengths and weaknesses of the players based on past performances of the player. We also embark on a critical aspect of team composition, which includes the number of batsmen and bowlers in the team. The team composition changes over time, depending on different factors which are tough to predict, so we take this input from the user and use the player embeddings to decide the best possible team combination with the given team composition.

KEYWORDS

Cricket Analytics, Data Mining and Data Analytics

1. INTRODUCTION

The advent of statistical modelling has contributed significantly to the success of teams and players in different sports. Different methods have been developed to evaluate player performances in different sports, but team sports pose a different challenge, as comparing two players of same nature and getting a suitable team against an opposition, is difficult. For example, in cricket [1], a team sport which is discrete in nature, comparing two players of same nature (comparing a batsman with another batsman, or a bowler with another bowler) in same and different teams is a complex task. Often, the players are compared based on their quantitative aspects like high scores, wickets taken and career averages (number of runs scored or conceded per dismissal) and teams are decided based on them only. The quantitative factors provide insights but miss some important aspects: Quality of Runs Scored: Two players who played against different oppositions (which are ranked differently) and performed similarly, will have similar statistics. In the mentioned case, the player who scored against better opposition, should be rated better. Quality of Dismissals: Dismissals of batsman with higher career average should be rated more than dismissals of batsman with lower career averages.

This paper tries to keep above two important aspects in mind and build a rating system called 'Quality Index of Player (ϕPlayer) which includes qualitative and quantitative aspects of player performance. Later, using ϕPlayer, we represent players as embeddings, to build a "Semi-Supervised Team Recommendation System". The embeddings obtained, describe the strengths and weaknesses of the players based on theirpast performances. While drafting a recommender

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 67-77, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101207

system, factors like overall complexity and set of parameters to be considered, are a major factor and a system with high complexity won't be much useful for instantaneous results. If all the possible valid team combinations are considered from a pool of players, the complexity of that would still be polynomial, but very high. Proper selection of parameters along with considering orderings following some constraints can be useful for instantaneous results and can also be used for in-match results when match is not going as predicted. The method, although proposed for cricket, can also be extended to other sports with some modifications.

2. RELATED WORK

In literature, the player rating methods like A. Ramalingam [2], MG Jhawar et al. [3], S. Akhtar et al. [4], Margaret I. Johnston et al. [5] are quantitative in nature and give high weight to batsman with more batting averages (runs scored per wicket) and bowlers with lower bowling averages (runs conceded per wicket). Some studies include graphical representations to compare players (A. Kimber [6] proposed a graphical method to compares bowlers). Q. Zhou et al. [7] explains how team recommendations should work, considering the aspect of expanding teams and substituting team members. L. Li, H. Tong et al. [8] also explains team member replacement, considering skill and structure matching.

Also, the existing work on "Team Recommendation for Cricket" mainly rank players on statistical measures or some techniques like clustering, etc. Prakash, C. Deep [9] ranks players using a Clustering Algorithm based on different batting and bowling parameters. In S.B. Jayanth et al. [10], K-Means and SVM with RBF Kernel was used to recommend teams for 2011 World Cup. F. Ahmed et al. [11] maximizes the overall batting and bowling strength of a team by optimizing a Multi Objective Problem. NSGA-II algorithm was used to optimize the overall batting and bowling strength of a team and find team members in it.

3. PROPOSED SOLUTION

In this paper, we propose a qualitative and quantitative rating mechanism called ’Quality Index of Player ϕPlayer' which is used to build player embeddings. CRICTRS, a semi supervised team recommendation system, uses the player embeddings and recommends a team based on opponent's strengths and weaknesses. The system utilizes the weakness of the opponent and finds a similar player in our team to recommend against the opponent for each player in opposition. This process is done considering every player in the opposition. For representational purpose, a bipartite graph can be used, with opposition being on one side, and our players on the other.

3.1. Dataset

Cricsheet dataset [12] contains data of over 1400 international ODI matches, played between 2005 to 2019. For each match, ball by ball data is available, with following features: 'Inning', 'Over', 'Team Batting', 'Batsman', 'Non-Striker', 'Bowler', 'Runs-Scored', 'Extras', 'Wicket' and 'Dismissed Batsman'. Along with this, details like competing teams, date of match, venue of match, match and toss result are also available. Cricinfo [13] was used to validate the information across each match.

3.2. Player Rating System for Cricket

Improving upon existing methods we try to build a method that considers quality of runs and wickets while rating the players. A brief overview of CRICTRS is shown in Fig. 1.

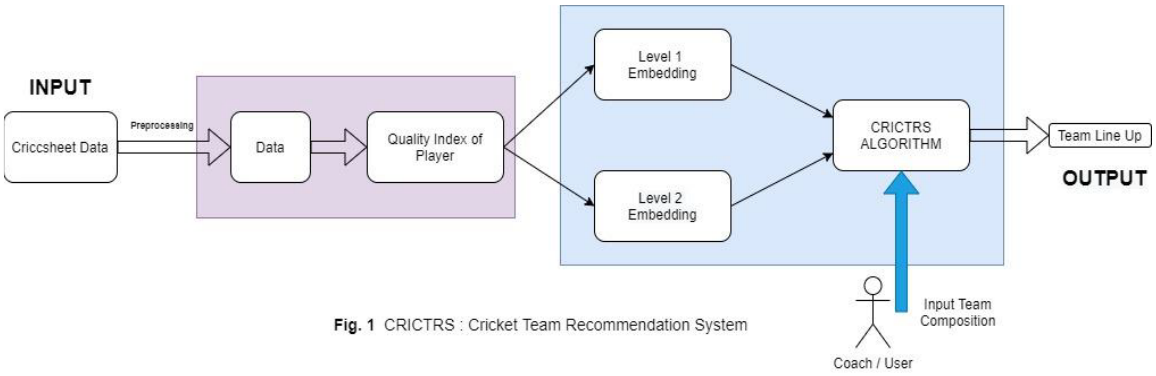


Fig. 1 CRICTRS : Cricket Team Recommendation System

3.2.1. Modelling a Match

We take the idea from [2] and model each delivery as a Bernoulli trial. The two possible outcomes for each Bernoulli trial or a delivery are 'r' runs scored or a wicket, where 'r' is defined as average runs scored by a batsman per ball. To evaluate a batsman's individual performance in a team, we evaluate the performance of a team that contains 11 replicas of same player and calculate expected score by that team with 10 wickets in hand. Thus, a team with 11 replicas of batsman, on average, will score 300*r runs in a match, when the team does not lose any wicket in a 50 over (300 balls) ODI Match. But if a team loses 'w' wickets, where w<10, then the team will score (300-w) *r runs in that match. In case of an all-out, when team loses all the wickets, average runs scored will be (b-10) *r, where b is the number of balls the team faced in that match. Similarly, a bowler can be evaluated by replacing 'r' as average run conceded per ball and evaluating expected runs conceded by a team of 11 replicas of the bowler. Using above, the expected outcome of match can be written as: -

$$\text{r} = \frac{runs}{ball} \quad , \quad avg = \frac{runs}{wicket}$$

$$\text{P(dismissal)} = \text{1-p} = \frac{1}{(balls\_per\_wicket)} = \frac{r}{avg}$$

$$E(runs) = Eall\_\_{out}(runs) + Eall\_\_{out}(runs)$$

$$E_{all-out}(runs) = \sum_{b=10}^{300} r*(b-10)*\left(\binom{b-1}{9}*p^{(b-1)-9}*(1-p)^{9}\right)*(1-p)$$

$$E_{\overline{all-out}}(runs) = \sum_{w=0}^{9} r*(300-w)*\binom{300}{w}*p^{300-w}(1-p)^{w}$$

$$E(runs^2) = \sum_{b=10}^{300} r*(b-10)*\left(\frac{b-1}{9}*p^{(b-1)-9}*(1-p)^{9}\right)*(1-p) + \sum_{w=0}^{9} r*(300-w)*\frac{300}{w}*p^{300-w}(1-p)^{w}$$

$$\sigma_{runs} = \sqrt{E(runs^2) - \left(E(runs)\right)^2}$$

3.2.2. Quality of Runs and Dismissals

The approach in [2] is completely quantitative and misses an important aspect of quality of opposition. We replace the quantity metrics of 'r' and 'avg' used in [2] with our quality and quantity-based metric which is re-evaluated as follows: -

$$C_{batsman} = \text{Career Averarge of Batsman}$$

$$C_{bowler} = \text{Career Averarge of Bowler}$$

Quality Metrics of Batsman

$$\text{Quality of Dismissal } (\boldsymbol{\phi}\text{Dismisaal}) = \frac{\mathbf{Cbatsman}}{\mathbf{Cbowler}}$$

$$\text{Quality of Run Scored}(\boldsymbol{\phi}\text{run}) = \frac{\mathbf{Cbatsman}}{\mathbf{Cbowler}}$$

Quality Metrics of Batsman

$$\text{Quality of Dismissal } (\boldsymbol{\phi}\text{Dismisaal}) = \frac{\mathbf{Cbatsman}}{\mathbf{Cbowler}}$$

$$\text{Total runs conceded=Runs Conceded+Extras}$$

$$\text{Quality of Run Scored}(\boldsymbol{\phi}\text{run}) = \frac{\mathbf{Cbowler}}{\mathbf{Cbatsman}}$$

With this method we can consider some important aspects of the match, that are difficult to capture otherwise. These include: -

1) Dismissals of top order batsman matter more and as usually top order batsman have higher career average, thus a bowler who takes wickets of in form high average batsman is more rewarding than a bowler who takes wickets of tail-enders.

2) Extras were completely ignored by all previous metrics. Here if a bowler bowls more extras, then he might be under pressure, thus extras are also an important metric while considering bowlers. A bowler who gives away more extras provides greater risk to the team by giving away runs.

After re-evaluating the 'r' and 'avg' of players using above, we finally calculate the player rating represented by '**Quality Index of Player**' (ϕ_{Player}) which is evaluated as:

$$\text{Quality Index of Player } (\phi_{\text{Player}}) = \frac{E(runs) - \phi avg}{\sigma \text{runs}}.$$

On evaluating the results, we observe the following: -

1) In [2] spinners and in general bowlers who bowled in the middle overs of the innings had higher rating than the bowlers who bowled in the powerplay and death overs, but our method regularises this as shown by the above examples. In above table, Harbhajan Singh and Yuvraj Singh had significantly higher ratings as compared to others by [2], but our method regularises the rating, and no such disparity is there.

2) Also, there is a difference in rating by [2] and ϕPlayer and some players are given higher rating by our method as compared to [2]. We believe that this is because these players performed better against strong oppositions, which should be valued more and [2] did not include this aspect of performance while rating the player.

3) We compared performance of different players over the years and computed the rating at different stages of their career. Figure 2 shows rating of Virat Kohli's performance over years. The Quality Index and rating by [2] were normalised and plotted. Our method gave higher rating to his performance in 2012 as compared to 2016, which [2] rates as highest. On a closer look we see that the bowlers he faced in 2012 included experienced players like Lasith Malinga, Nuwan Kulasekara, Umar Gul, Saeed Ajmal, Clinton McKay, etc. While the bowlers he faced in 2016 included players like Jimmy Neesham, Josh Hazzlewood, Mitchell Santner, etc who were in the early stages of their career in 2016. Thus, we believe that the performance of Virat Kohli in 2012 should be rated more than 2016, as captured by our rating method also. A similar analysis is done for comparing Pat Cummins' performance over the years. We can see difference in ratings by [2] and our system, which is due to his performance against different oppositions.

Table 1. Rating of Batsmen

Batsman	Rating By [2]	$\phi Player$
Virender Sehwag	2.05	8.87
Sachin Tendulkar	4.88	11.45
Gautam Gambhir	3.47	12.95
Yuvraj Singh	2.66	7.08
Mahendra Singh Dhoni	6.80	9.50
Yusuf Pathan	1.23	3.07

Table 2. Rating of Bowlers

Bowler	Rating By [2]	$\phi Player$
Zaheer Khan	1.94	3.77
Praveen Kumar	1.75	3.40
Ashish Nehra	1.49	3.04
Harbhajan Singh	4.90	3.15
Yusuf Pathan	0.68	2.04
Yuvraj Singh	2.36	3.09

Fig 2: Performance comparison of Virat Kohli

Fig 3: Performance comparison of Pat Cummins

3.3. Semi-Supervised Team Recommendation System

ϕPlayer proposed above helps us rate batsman and bowlers in cricket. But using rating directly to recommend a team can yield bad results as it does not capture the weaknesses and strengths of players, which are an important factor while forming a team against any opposition. Thus, we use

an embedding based approach where we model each player as an embedding by comparing their performance against other players they face. The embeddings derived, capture the strengths and weaknesses of the players.

3.3.1. Player Representation: 2 Level Embeddings

We represent all the players as vectors. Every batsman is assigned a number $\in \{1, 2\ldots$ NBatsman$\}$, where NBatsman denotes the total number of batsmen. Similarly, every bowler is assigned a number $\in \{1, 2\ldots$ NBowler$\}$, where NBowler denotes the total number of bowlers.

Level 1 Embeddings: For a batsman, the Level 1 embedding is a vector, where index 'i' of the embedding gives ϕbatsman against the bowler assigned the number i. Similarly, for a bowler, the Level 1 embedding is also a vector, where index 'i' of the embedding gives ϕbowler against the batsman assigned the number i. The

ϕPlayer at each index is calculated using batsman versus bowler data corresponding to that index, extracted from different matches in [12]. Also, all those indices in the embeddings are set to 0, for which versus data is not available, as those players did not face each other.

Level 2 Embeddings: The Level 2 embeddings are derived from Level 1 embeddings. For every batsman we compare ϕPlayer over his career against all the bowlers with the values at every index in the level 1 embedding representing batsman's performance against those bowlers. If the players have faced each other and there is a significant negative deviation in the performance of the batsman i.e. there is a drop in ϕPlayer for batsman, then the index is set to 0, otherwise it's set to 1. Similar process is followed for the bowlers and embeddings are constructed for bowlers also by comparing their performance against different batsmen with their overall performance. The level 2 embeddings, in a way, represent a batsman in terms of the bowlers he dominates, and the bowlers in terms of the batsmen they dominate.

In cricket, batsmen have weaknesses against bowlers. For example, a batsman may struggle against spin bowlers and play well against medium/fast bowlers. Thus, our embeddings capture the strengths and weaknesses of the players. Both levels of embeddings provide different views of player's strengths and weaknesses. Using the derived embeddings, we propose a 'Semi Supervised Recommender System' to suggest a team based on the given input and the opponent. We formulate our problem as: -

Suggesting a Team Given the player embeddings and team composition i.e. the number of players of each type (Batsman, Bowler and Wicket-Keeper), along with potential list of players in opposition, suggest a team of players best suited to face any combination of players of opposition.

3.3.2. Getting the Right Team Combination

Every team has a critical aspect of team composition, which includes the number of batsmen and bowlers in the team. The team composition changes over time depending upon the playing conditions, venue, opponent team, etc. No team combination is suited for all the conditions. The dataset [12] , only gives us ball by ball proceedings of the match, with no information regarding playing conditions. O. Alkan et al.

[14] builds a team based on the roles required for the opportunity and assigning the right person to fulfil it. In cricket, team composition decides the roles of different players. Thus, we propose a 'Semi Supervised' approach for our recommender system, where the coaches input desired team

composition. The team composition comprises of the number of batsmen, bowlers, and all-rounders in the team. Also, the coach inputs a set of players among which the team selection is to be made. Thus, our semi supervised recommender system combines the experience of a coach with our model to generate an optimum team combination.

3.3.3. Constraints on the Solution

The output is a set of (Player, Role), where Role is either batsman, bowler, all-rounder, or wicketkeeper. If it is being used for suggesting a team, then the following conditions should be satisfied: -

- Number of elements in the set >= 11
- Number of Players with the role wicketkeeper should be >= 1
- Number of Players with the role bowler and all-rounder should be >= 5

4. ALGORITHM FOR TEAM RECOMMENDATION

This section explains the algorithm used for CRICTRS in detail. Team configuration is taken as input for recommending players. Players are listed out using Level 2 embeddings against whom the opposing players are weak, from all countries. After that, for each opposition player, we check if there is a significant similarity in the Level 1 player embedding of a player from our team and any player in the above list of players against whom the opposition players have weaknesses. If so, then we can say that the player will outperform that opposition player. This approach is derived from collaborative filtering (explained in Y. Koren et al. [15]) as it uses an embedding based approach, although they were built without the use of supervised learning. Rather, we use a completely statistical approach to derive the embeddings.

We first adapt our solution to satisfy the constraints. Hence, the required number of wicketkeepers are selected first. Afterwards, the batsmen and bowlers are selected. Selection of these players depends on the opposition. For batsmen and wicketkeepers, the bowlers in opposition are considered, while bowlers are selected considering the batsmen in the opposition team.

4.1. Bipartite Graph Representation

A representation in the form of a bipartite graph can be constructed, where batsman is on one side, and bowlers are on the other, and vice versa. This representation is constructed based on the weakness of the opposition's players. A player of our team and the opposing team is connected by an edge, if the similarity between one of the players whom the player in the opposing team is weak, with our player, is below a threshold. Thus, we can construct such graph for batsmen and bowlers of our team, with bowlers and batsmen on the other side of the graph, respectively. Level 2 embeddings are used to construct the bipartite graph.

4.2. Ordering Players from Bipartite Graphs

Each edge on the bipartite graph has an edge weight equal to ϕ_{Player} against each other from Level 1 embeddings. For each player of our team, we compute δ, which is defined as $\delta = \frac{Mean\ of\ Edge\ Weights}{Standard\ Deviation\ of\ Edge\ Weights}$ using the edges connected to the player node. The significance of selecting δ as the deciding parameter is that a player with high variance would have a lot of difference in the ϕ_{Player} against the player he/she dominates, thus providing greater risk to the team. Hence, stability across the players dominated is also considered in our algorithm. We sort

the players in decreasing order of δ and make the selection.

4.3. Selection of Players from the Orderings

We construct the orderings of batsman, bowlers, and wicketkeepers. Wicketkeepers are selected first, as our recommendation should satisfy the constraints mentioned in section 3.3.3. While selecting the required number of batsmen, we check if they can bowl too and if so, whether they can bowl better than they can bat. So, we pick players with role batting all-rounder or bowling all-rounder. Similarly, bowlers and bowling all-rounders are picked.

5. EXPERIMENTS

CRICTRS involves a comprehensive embedding-based approach, where we represent a player as an embedding and select a team against the opposition using it. Different experiments were done to ensure the quality of individual components and a validation of overall results was also done.

5.1. Validating the Player Embeddings

The derived embeddings were validated by analysing the clusters of formed by them. Two of the obtained clusters from above are shown in table 3 and table 4.

Table 3. Cluster of similar Batsmen

Alastair Cook
Marcus Trescothik
Nathan Astle
Hashim Amla
Virat Kohli
David Warner

Table 4. Cluster of similar Bowlers

Saad Nasim
Ajit Agarkar
Khaled Mahmud
Rubel Hossain
Umesh Yadav
CR Braithwaite
Azhar Ali

As we can observe from the above, the players in the same cluster, either have same playing style or they have dominated over similar kind of opposition. This validates our idea of modelling the players as embeddings, to capture the strengths, weaknesses, and other traits of the player, which is difficult to capture from statistics only.

5.2. Team Recommendation for Different Matches

We used CRICTRS to get teams for different matches. We obtained recommended team for South Africa, for the Australia v/s South Africa match on Oct 2, 2016. Using the Level 1 and Level 2 embeddings we obtained the bipartite graphs as shown in figure-4 and figure-5. The team recommended by CRICTRS, with a team composition of 5 batsman, 4 bowlers, 1 wicketkeeper and 1 bowling all-rounder is shown in table 5. On comparing with the team that played in the actual match, we see that CRICTRS suggests Hashim Amla in place of Behardien, while rest of the team is same. In the actual match, Behardien did not perform well. Also, when Hashim Amla played in the next match on Oct 5, 2016 against Australia then he performed significantly better than Behardien. Similarly, Indian team for the India v/s Pakistan match on 4th June 2017 was derived. As the two cricketing nations, do not play much cricket against each other, the versus statistical data is not available. Thus, CRICTRS provides an efficient team recommendation mechanism in such case by identifying similarities from the players opposition has already faced. The team recommended by CRICTRS, with a team composition of 3 batsman, 4 bowlers, 1 wicketkeeper, 1 batting all-rounder and 1 bowling all-rounder is shown in table 6.

Table 5. Recommended Players for Australia v/s South Africa

Batsman	Bowler	Wicketkeeper	Batting All-Rounder	Bowling All-Rounder
Hashim Amla	Imran Tahir	Q De Kock	None	Wayne Parnell
Faf du Plesis	Kagiso Rabada			
David Miller	Dale Steyn			
JP Duminy	Andile Phelukwayo			
Rilee Rossouw				

Table 6. Recommended Players for India v/s Pakistan

Batsman	Bowler	Wicketkeeper	Batting All-Rounder	Bowling All-Rounder
Rohit Sharma	Umesh Yadav	MS Dhoni	Kedhar Jadhav	Ravindra Jadeja
Virat Kohli	Jasprit Bumrah			Hardik Pandya
Shikhar Dhawan	R. Ashwin			
	B. Kumar			

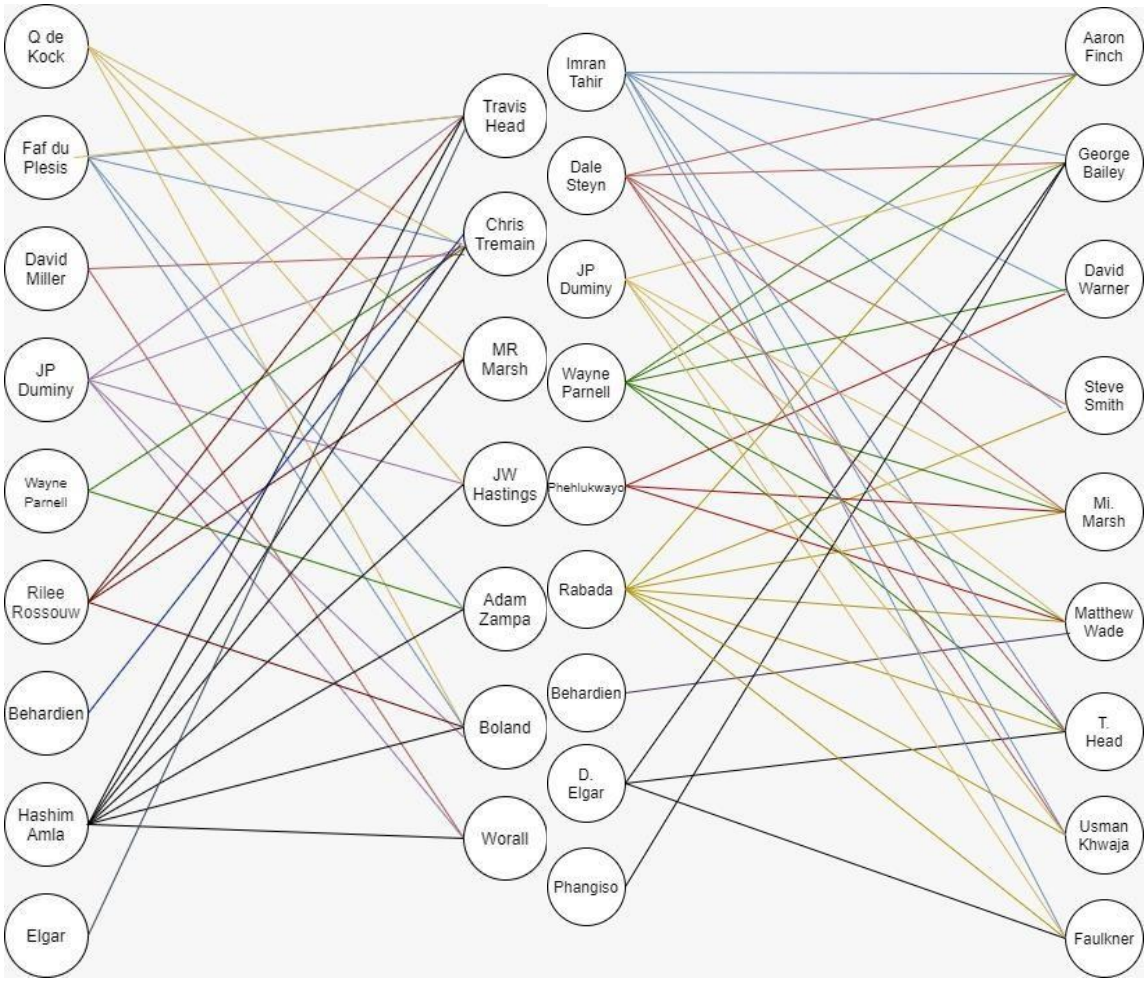


Fig 4: Bipartite Graph Representation of Batsman Bowler of South Africa against Australia

Fig 5: Bipartite Graph Representation of of South Africa against Australia

5.2.1. Team Line-Up similarity for ICC CWC - 2019

Comparisons were done between CRICTRS team recommendation results and actual team line-ups for ICC Cricket World Cup, 2019. The tournament had 48 matches in total, out of which 3 were abandoned due to rain, without a bowl bowled. For the other 45 matches, which had a winner, we compared the similarity in the line-ups generated by CRICTRS and the team line-ups in the actual match. The team composition for each team was kept similar to that in the actual match. The results obtained are shown in table 7: -

Table 7. Team Line-up similarity for ICC CWC-2019

	Team Line-up Similarity
Winning Team	82.47%
Losing Team	74.36

The low team line-up similarity with losing teams suggest that, CRICTRS recommends a different line-up for losing teams. Also, high similarity with winning teams validates that CRICTRS's recommended team line-up have high winning chances. Thus, above statistics prove useful in validating CRICTRS as a cricket team recommendation system.

6. CONCLUSION

CRICTRS provides a method to use historical data of different matches and recommend team based on opposition's strengths and weaknesses. CRICTRS models players as embeddings using the data and identifies weaknesses, strengths, and other traits of players. Cricket is an ever-evolving sport, with introduction of new rules, regulations, and technology. According to a new set of rules, the ICC allows players, who suffer concussions during a match, to be replaced in their team's playing XI [16]. However, the regulations emphasize on a 'like-for-like' replacement for the concussed player and that part remains to be under dark clouds [17]. Our embedding based approach can be used here, in finding and validating a 'like-for-like' replacement by checking the similarity in player embedding of the injured player and the replacement. In our analysis, we also tried to include domestic circuit players by using VORP (Value Over Replacement Player)[18] theory from baseball and consider a run scored by a domestic level batsman against another bowler at domestic level, as 0.8 runs scored for the batsman and 1.2 runs conceded for the bowler. Thus, players across various levels can be compared with this and a uniform attribute is created.

Thus, CRICTRS can prove to be useful team recommendation tool and can help coaches and team management to decide their team line-ups against an opposition. Also, CRICTRS can also be extended to build a team recommendation system for other sports by modelling the sport game as a Bernoulli trial with appropriate outcomes. The ϕPlayer can be evaluated using the data, and the obtained rating system can be used to derive the embeddings in a similar manner.

7. ACKNOWLEDGEMENTS

Prazwal Chhabra and Rizwan Ali contributed equally to this paper.

REFERENCES

[1] Video Introducing the basics of What is Cricket? https://www.icc-cricket.com/about/development/what-is- cricket (accessed July 24, 2020).

[2] A. Ramalingam, "Bernoulli runs using 'book cricket' to evaluate cricketers," Reson, vol. 17, no. 5, pp. 441–453, May 2012, doi: 10.1007/s12045-012-0047-2.

[3] MG Jhawar, and Vikram Pudi, "Honest Mirror: Quantitative Assessment of Player Performance in an ODI Cricket Match." in MLSA@ PKDD/ECML, 2017, pp. 62-72. Available: http://ceur-ws.org/Vol-1971/paper-08.pdf

[4] S. Akhtar, P. Scarf, and Z. Rasool, "Rating players in test match cricket," Journal of the Operational Research Society, vol. 66, no. 4, pp. 684–695, Apr. 2015, doi: 10.1057/jors.2014.30.

[5] Margaret I. Johnston, Stephen R. Clarke, and David H. Noble. "Assessing player performance in one-day cricket using dynamic programming." 1993.

[6] A. Kimber, "A Graphical Display for Comparing Bowlers in Cricket," Teaching Statistics, vol. 15, no. 3, pp. 84– 86, Sep. 1993, doi: 10.1111/j.1467-9639.1993.tb00664.x.

[7] Q. Zhou, L. Li, N. Cao, N. Buchler, and H. Tong, "E xtra," presented at the RecSys '18: Twelfth ACM Conference on Recommender Systems, Sep. 2018, doi: 10.1145/3240323.3241610.

[8] L. Li, H. Tong, N. Cao, K. Ehrlich, Y.-R. Lin, and N. Buchler, "Replacing the Irreplaceable," presented at the the 24th International Conference, 2015, doi: 10.1145/2736277.2741132.

[9] Prakash, C. Deep, C. Patvardhan, and C. Vasantha Lakshmi. "AI Methodology for Automated Selection of Playing XI in IPL Cricket.", 2017. Available: http://www.ijetsr.com/images/short_pdf/1497897339_6- cdac191_etsr.pdf

[10] S. B. Jayanth, A. Anthony, G. Abhilasha, N. Shaik, and G. Srinivasa, "A team recommendation system and outcome prediction for the game of cricket," JSA, vol. 4, no. 4, pp. 263–273, Nov. 2018, doi: 10.3233/JSA-170196.

[11] F. Ahmed, A. Jindal, and K. Deb, "Cricket Team Selection Using Evolutionary Multi-objective Optimization," in Swarm, Evolutionary, and Memetic Computing, Springer Berlin Heidelberg, 2011, pp. 71–78.

[12] Cricsheet, 2019. [Online], Available: https://cricsheet.org/downloads/odis.zip.

[13] ESPNcricinfo. https://www.espncricinfo.com/ (accessed June 1, 2020).

[14] O. Alkan, E. M. Daly, and I. Vejsbjerg, "Opportunity Team Builder for Sales Teams," presented at the 23rd International Conference, 2018, doi: 10.1145/3172944.3172968.

[15] Y. Koren and R. Bell, "Advances in Collaborative Filtering," in Recommender Systems Handbook, Springer US, 2010, pp. 145–186.

[16] ICC Test-Match-Playing-Conditions. https://icc-static- files.s3.amazonaws.com/ICC/document/2019/09/02/9182955c-04a4-4fa0-a2b6-f5908f02a51d/ICC-Test-Match-Playing-Conditions-Final-1-September-2019.pdf

[17] ICC guidelines regarding 'like-for-like' replacement. https://www.cricket.com.au/news/concussion-substitutes- like-for-like-replacement-icc-test-championship-ashes-geoff-allardice/2019-07-30

[18] Wikipedia Contributors." Value over replacement player." Wikipedia.com https://en.wikipedia.org/wiki/Value_over_replacement_player (accessed June 1, 2020).

ANALYZING AND FILTERING FOOD ITEMS IN RESTAURANT REVIEWS: SENTIMENT ANALYSIS AND WEB SCRAPING

Nina Luo[1], Caroline Kwan[2], Yu Sun[3], Fangyan Zhang[4]

[1]Webb School of California, Claremont, CA 91711, USA
[2]Westridge School, Pasadena, CA, 91105, USA
[3]California State Polytechnic University, Pomona, CA, 91768
[4]ASML, San Jose, CA, 95131, USA

ABSTRACT

Online reviews now influence many purchasing decisions. However, the length and significance of these reviews vary, especially when reviewers have different criteria for making their assessments. In this paper, we present an efficient method for analyzing restaurant reviews on the popular review site known as Yelp. We have created an application that uses web scraping, natural language processing, and a blacklist to recommend customer favorite dishes from restaurants. To test the app, we have conducted a qualitative evaluation of the approach. Through analyzing two different ways to obtain Yelp reviews and evaluating our word filtering process, we have concluded that an average of 51% of nonfood words are filtered out by the blacklist we made. We provide further details of its deployment, user interface design, and comparison to the opinion mining field, which utilizes similar tools to make financial market predictions based on the perceived public opinion on social media.

KEYWORDS

Web scraping, natural language processing, flutter, iOS, android

1. INTRODUCTION

Online user reviews provide feedback about a person's experience with a business [1]. These reviews are often written by customers and generally include an overall score based on a customer's experience and expectations with a service or product. As individual user reviews often provide assessments based on different criteria, it is necessary to piece together a wide range of customer preferences and impressions on different aspects of a service or product to get a complete and accurate picture. This way, we are more well informed and aware of the contrasting opinions regarding the business.

The internet not only provides easy access to large amounts of information but also serves as a global forum and publishing platform [2]. With the rise of fast-paced modern technologies and social networks, immediate satisfaction is granted to users because information can quickly be found and read. As a result, more customers are using the internet to evaluate and comment on services and products [3]. Online website reviews give customers the opportunity to share their experiences and express their observations. Through the first-hand perspective of a customer, user reviews detail the positives and negatives of an experience. Positive reviews are shown to increase revenue, improve a company's reputation, and boost customer loyalty. While positiv e reviews strengthen the reliability and image of a service or product, negative reviews allow for

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 79-88, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101208

improvement and strengthen credibility. When a business replies to a negative review, the business acknowledges the complaint and shows that its values customer feedback. This interaction provides an opportunity for a company to demonstrate its principles and reduces the impact of negative reviews.

Online user reviews are important as they can influence purchasing decisions and determine the reputation of a business [15]. Furthermore, these reviews encourage business and customer relationships by allowing businesses to respond to customers and establish trust through customer engagement. While reviews provide insight about a service or product, some reviews can be extensive or resemble a rant. Although websites have implemented systems to filter out unhelpful reviews, not many reviews are removed and most reviews are published onto the internet for users to read. However, small business websites tend to lack users in the first place and therefore have a low amount of reviews. A limited amount of reviews can mislead users as the few reviews dominate the narrative of a service or product.

Luckily, there are websites dedicated to reviewing restaurants and businesses [14]. A popular review site known as Yelp consolidates business information and allows customers to easily write and read reviews. Large and small businesses alike are given a platform to gain recognition and listen to customer feedback. An automated software finds effective reviews and recommends highly rated restaurants and businesses. As Yelp's automated software regularly tweaks itself and gathers more reviews, the recommended reviews change over time. In order to simplify and ease the task of reading tons of reviews, Yelp implements features known as "Review Highlights" and "Popular Dishes". The "Review Highlights" section points out popular words written in reviews and is a quick way for users to discover the main components of a restaurant or business. While the highlighted words usually refer to a location, item, or name, some words can be irrelevant or unhelpful to users. Another feature is the "Popular Dishes" section. As the name suggests, this section is a gathering of reviews and photos of popular foods. However, this feature does not show up on all restaurant pages, and the dishes are automatically selected by Yelp's machine learning algorithm. Although Yelp uses algorithms to find popular words and dishes in reviews, its sorting process is not always relevant, and the features do not apply to all businesses. To improve their algorithms, Yelp has recently partnered with GrubHub to provide full restaurant menus.

Since a majority of users seek restaurant information from Yelp, this project focuses on facilitating the process of sorting information in restaurant reviews. The approach presented in this paper focuses on recommending highly rated food dishes based on Yelp user reviews aided by web scraping and sentiment analysis. We have developed Dishcovery, a mobile application that allows users to discover the best tasting, most popular foods in nearby restaurants.

Dishcovery uses machine learning to compile Yelp reviews and construct a sentiment analysis rating of food items in restaurants. Unlike Yelp which only provides an overall 5-star rating system for reviews, Dishcovery ranks individual food dishes in restaurants on a precise decimal scale from negative one to one. Using Yelp as a resource to provide restaurant reviews, Dishcovery searches for food items and ranks the dishes based on the positive and negative comments made in reviews. The mobile application is supported by a back end that collects data from Yelp, sorts out food dishes, and creates a numerical rating for each item. The strength of Dishcovery is that through the method of web scraping, we are able to apply our application to find the recommended items of any customer review-based website. This includes already popular websites such as Angie's List, Trip Advisor, Google My Business and much more. Our program can easily expand beyond recommending restaurants on Yelp. Without drastically changing the code, our application can compile a list of favorite aspects from a service, trip, or company.

We have experimented with two different methods to extract user reviews from Yelp. First, we tested the Yelp Fusion API that provides free access to over 50 million businesses [11]. We initially thought this would be a good way to request the Yelp data. However, the API is restrictive and only provides a limited amount of information. For example, API requests are capped at 5000 calls per day and only three partial reviews of 160 characters total are provided for each call. We contacted Yelp to increase our daily calls and get access to more Yelp reviews, but they rejected our request for academic use. These restrictions prevented us from getting enough information to run the machine learning algorithms necessary for finding food items. We decided to try another method to get more data. Second, we used the Python requests library combined with Beautiful Soup to extract all the reviews. Although the Python web scraping gave us a plentiful amount of information from the Yelp URL pages, it was difficult to locate all the reviews because they were in separate places in the HTML code. By using the Python library Beautiful Soup to parse the HTML code, it simplified this process and grouped all the reviews together in a list. The list created for each business has more information than the Yelp Fusion API and allows for better filtering of the food items. Beautiful Soup provides the best results because the Yelp reviews can be quickly found by the HTML tag names and organized together. These experiments have helped us find the best way to obtain the Yelp reviews.

The rest of the paper is organized as follows: Section 2 details the four main challenges and limitations we encountered during the making of our program. Section 3 identifies the solutions and steps we took to combat the challenges mentioned in Section 2. Section 4 showcases the two evaluations we conducted and an analysis of the results; Section 5 cites the relevant related works from other authors for further reading as well as comparisons to our own work. At the end in Section 6, we conclude the paper, restate the main ideas, and explain the future work of our project.

2. CHALLENGES

There are a few of challenges existing in the project. They will be discussed one by one in this section.

2.1. Challenge 1: Limited information from Yelp Fusion API

A big part of Yelp's success stems from obtaining a large amount of user data and business information. Since Dishcovery recommends food dishes from restaurant reviews, it also depends on abundant data sets. We initially used the restaurant information and user reviews provided by Yelp Fusion API through Postman. Although easy to access, the Yelp API limits the amount of data developers can view and only provides developers three incomplete reviews for each business. The limited reviews do not present a complete picture of a business and provide inadequate information necessary for running the natural language processing and sentiment analysis machine learning algorithms. As a result, few food items were identified and the algorithm inaccurately analyzed the overall sentiment of the reviews.

For example, the restaurant review from Yelp Fusion API provides the text "This is one of my favorite restaurants. The food here is so different in a good way. They serve good quality food. My favorite things to get from here are…" The review is cut off and the food recommendations are not shown. This creates a problem for the machine learning algorithms because there are not enough words to analyze and no food items to extract. We addressed this by using the method of web scraping to get the full user reviews from Yelp.

```
"id": "WPuMjo9TZx5_Am1dkk6elA",
"url": "https://www.yelp.com/biz/din-tai-fung-arcadia-3?adjust_creative=J4UsXVUs5Myu_cxdiocugg&
    hrid=WPuMjo9TZx5_Am1dkk6elA&utm_campaign=yelp_api_v3&utm_medium=api_v3_business_reviews&
    utm_source=J4UsXVUs5Myu_cxdiocugg",
"text": "This is one of my favorite restaurants. The food here is so different in a good way. They serve good quality
    food. My favorite things to get from here are...",
"rating": 5,
"time_created": "2020-05-15 00:01:40",
"user": {
    "id": "Q6slEKKM1LrE9WStPD7z8w",
    "profile_url": "https://www.yelp.com/user_details?userid=Q6slEKKM1LrE9WStPD7z8w",
    "image_url": "https://s3-media4.fl.yelpcdn.com/photo/csj8IfT1Z4LCXmUAE_4_Fg/o.jpg",
    "name": "Sarina T."
```


Figure 1: A limited review from Yelp Fusion API that has its food recommendations cut off.

2.2. Challenge 2: Difficulty identifying food items

Reviews address many aspects of a business or restaurant and contain a lot of information. Therefore, it can be difficult to filter out the keywords, like food items and beverages. Additionally, slang and typos often confuse the sorting algorithm. For example, words such as ‘table’ and ‘smile’ are sometimes classified as potential dishes. Although Google Natural Language API has a powerful model that identifies parts of speech, it has trouble with sorting out nouns and often groups them into the other category. The other category requires more implementation of filtering algorithms as it stores the words the natural language processing could not identify. For our purposes, the list needs to be shortened and only contain words relevant to items on a restaurant menu. We fixed this problem by creating a blacklist of common words mistaken by the API to be a dish. The unrelated words in this list will be filtered out, leaving the food dishes.

2.3. Challenge 3: Analyzing positive and negative comments

Most restaurant and business reviews have an overall positive or negative tone. Although Yelp’s 5-star rating system does give us an idea of the reviewer’s opinion, it can occasionally be an unreliable indicator of the feelings expressed in a review. For example, every reviewer has different criteria for rating a restaurant. One customer may express that the food was great but give a three-star review while another customer may express that the food was okay and give the restaurant a 5-star review. Looking at potentially inaccurate star ratings could influence the way users initially view the restaurant. Thus, we decided to focus on the general sentiment of the review instead of using Yelp’s rating system to give us a better understanding of the reviewer’s intended tone. Using Google Natural Language API, we evaluated the sentiment analysis of each review and ranked each dish on their positive or negative attribution from a precise scale of negative one to one.

2.4. Challenge 4: Promotion of app

As a food recommendation program, we wanted Dishcovery to be accessible to the most people and decided to make it an app. We started with building the application in Flutter to make it available on both the Google Play Store and App Store [12]. Then, we improved the back-end data collection to get more information and user reviews. The app has three pages: the main page, the search page, and the about page. The main page and search page suggest nearby restaurants and show pictures of food items. The about page gives a description of the app itself and information about the creators of the app. To promote our app, we decided to create a website. However, we were unsure of how to make it available online and how to build a presentable website design. We solved this problem by finding a premade theme on Envato Market and modifying its HTML code using Sublime Text to fit the Dishcovery app theme [13]. This allowed us to share Dishcovery with more online users.

3. SOLUTION

We have created Dishcovery, an app that recommends popular restaurant dishes from Yelp reviews. When a user selects their desired restaurant or business, information such as pictures and user reviews are collected from the Yelp website. The reviews are then filtered by machine learning algorithms and a blacklist to find the popular dishes from positive Yelp reviews [9] . A JSON file returns the highly recommended food items back to the app for the user to explore.

Main Page: The main page of Dishcovery uses location to recommend the top five most nearby restaurants. Each restaurant includes the following details: restaurant name, distance to restaurant from current location, average Yelp star rating, average cost of the food items on the menu, and a picture of one of the dishes. The information inside each restaurant page includes the top five most popular and highly recommended food dishes that are filtered by machine learning algorithms addressed later in the paper.

The restaurant and business data come from the Yelp website. Yelp reviews are extracted from the actual restaurant Yelp page using the method of web scraping aided by the Python package Beautiful Soup [10]. Google Natural Language API and a blacklist sort out the food and drink items and rank the sentiment of the items on a scale from negative one to one. A higher score means that the item comes from a more positive review.

Search Page: The search page suggests hot searches that are frequently searched by Dishcovery users. A small picture and the price of the suggested food items are shown. At the top of the page, the search bar can be used to find restaurants.

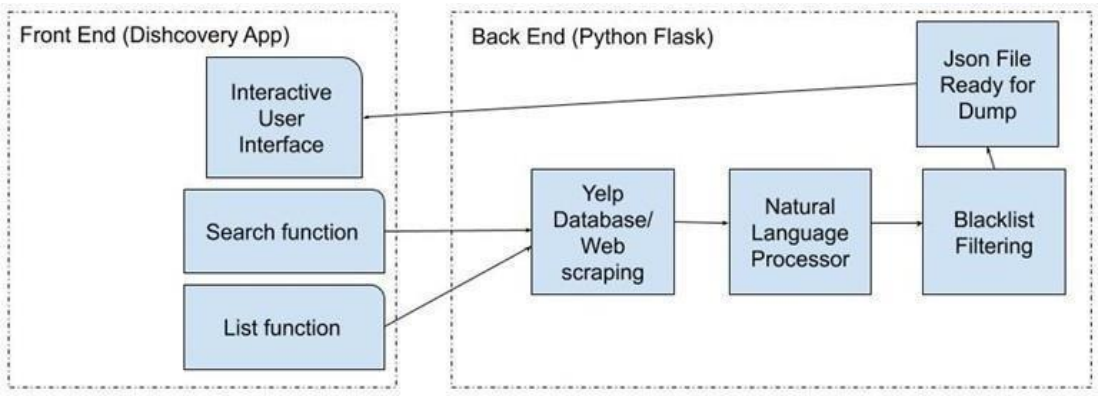


Figure 2: Overview of Dishcovery

Mobile Application: The front end of the mobile application is developed in Dart, a fast and object-oriented language that is used in Flutter apps. Created by Google, Flutter is a UI software development package that can be used to make iOS and Android apps. The interface and design of the three pages on our app are built in Flutter. The back end consists of using Beautiful Soup, Google Natural Language API, and the blacklist we made.

In Beautiful Soup, the program parses through HTML code to find information through tags and HTTP headers. As shown later in experiment 1, this web scraping process provides more user reviews than using Yelp Fusion API. After the reviews are gathered, we use natural language processing to classify food items from the Yelp reviews. We chose the syntax analysis, entity analysis, and sentiment analysis from Google Natural Language API because of its powerful deep learning models and variety of features. Its ease of access and inexpensive price are also added bonuses. The syntax analysis identifies parts of speech and sorts them into different categories.

The entity analysis labels words by type, such as person or location. This is helpful for finding the nouns and food items in the other category. The sentiment analysis creates a score of the feeling expressed in the reviews and determines whether a review has a positive or negative tone. Since the purpose of our app is to recommend popular dishes, we programmed the natural language processing to only include positive reviews that have a positive sentiment score.

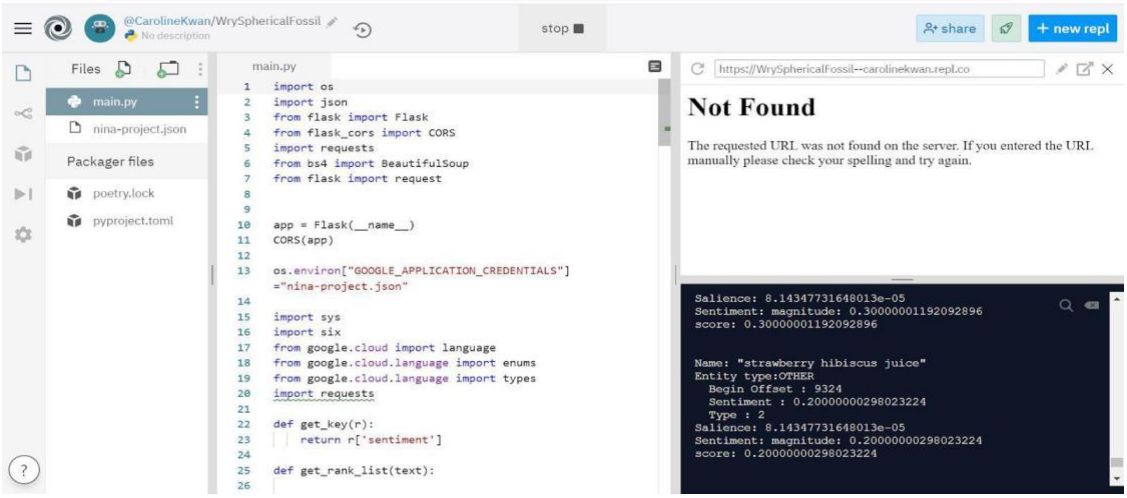


Figure 3: An analyzed food item returned from Google Natural Language API.

Although the entity analysis categorizes words into different types, many words that cannot be identified by the machine learning algorithms are sorted into the other category. Since the food items are in this category, we had to make additional parameters to sort them out. We created a dictionary of words called the blacklist that stores words unrelated to food dishes. This way, the nonfood words are sorted out through the blacklist and remaining words in the other category are limited to the food items. However, this filtering method requires a lot of time as we have to manually write out all the words in the list. After being refined by the blacklist, the condensed list of food words is then returned in a JSON file to the app and shown to the user.

We made a Dishcovery website (dishcovery.app) to promote the services of our app and allow users to easily contact us. The design of the website is created with a premade Envato Market theme. After transferring the files onto GitHub, we hosted our website on GitHub Pages, customized the layout to fit our needs, and bought a domain name. On the main page of the website, we have conveniently included links for people to download the iOS and Android versions of our app.

4. EXPERIMENT

4.1. Experiment 1: testing Data Retrieval Effectiveness

We have analyzed 50 Yelp restaurant reviews from 5 different cities to evaluate the advantages of web scraping with Beautiful Soup over Yelp Fusion API. To count the number of food items, we ran the reviews from Beautiful Soup and Yelp Fusion API through the same machine learning algorithms. As previously mentioned, Yelp Fusion API only provides an average of 160 characters for the reviews in each restaurant. This is not enough for the machine learning algorithms to find food items or accurately analyze the sentiment of the reviews. The solution is web scraping with Beautiful Soup because extracting the full reviews directly from the Yelp restaurant website page provides the most updated and accurate results. Our results show that Beautiful Soup consistently recommends more food items over Yelp Fusion API.

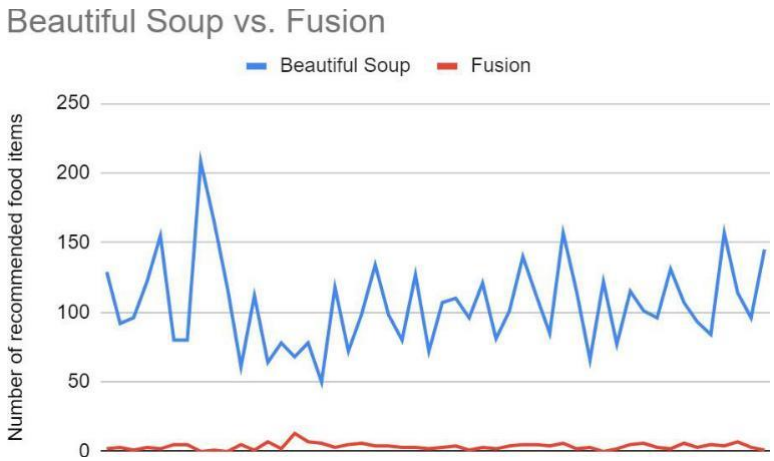


Figure 4: test results of Beautiful soup and Fusion

As shown in Figure 4, there is a significant difference between the number of food items returned by Beautiful Soup and Yelp Fusion API. Since Yelp API provides such limited data, the number of food items in the reviews is consequently very low. This can be seen by the red line (Yelp Fusion API) constantly returning numbers close to 0. The results support our conclusion that another solution is required in which more reviews are provided. We used Beautiful Soup to maximize the amount of reviews we could obtain from Yelp. By using the data shown in a restaurant's HTML page on Yelp, we were allowed to access a lot more reviews. In comparison to Yelp Fusion API, the blue line (Beautiful Soup) constantly outputs an average of 105 food items. This disparity shows that the web scraping method proposed fixes the limited amount of data Yelp Fusion API provides.

Since the main goal of our app is to recommend customer favorite food dishes to users, it is important that the recommendations are accurate. To ensure that unrelated words are not mixed in with the recommendations, we have manually created a blacklist to filter out those unrelated words. In our second experiment, we tested the effectiveness of the blacklist by collecting samples from 50 different restaurants in 5 different locations to ensure the reliability of the final food recommendations. The experiment data can also serve as a future reference for retrieval improvement to the blacklist. The outlier, being the difference of 77.6 in 'City 2', shows that we could filter out more unrelated words.

4.2. Experiment 2: Reliability of Blacklist and Retrieval Improvement

Experiment 2 focuses on measuring the effectiveness and reliability of the added blacklist. Food items in Yelp reviews from 50 restaurants in 5 different cities are collected and averaged. The numbers in the 'Without Blacklist' column signify the mean amount of dish words the algorithm finds from the restaurant reviews in each city, including all unrelated words that the algorithm picked up on as well. The numbers in the 'Blacklist' column include the mean amount of dish words in each city NOT including the unrelated words that have been filtered by the blacklist. Through the results of experiment 2, we have found that the blacklist is very effective in filtering out words that are not food dishes. In some cases, the blacklist even blocks out more than half of the original words in the retrieval sample. This can be seen in the 'Difference' column, which shows how many unrelated words the blacklist has filtered out.

	Without Blacklist	Blacklist	Difference
City 1	236.5	115.4	121.1
City 2	151.4	73.8	77.6
City 3	199.8	97.8	102
City 4	219	103.7	115.3
City 5	219.1	105.8	113.3

The results of both experiments prove that our proposed solutions are valid and beneficial to the application. In experiment 1, we tested the effectiveness of using Beautiful Soup and Yelp Fusion API to output the food words found in the Yelp reviews. We concluded that the web scraping with Beautiful Soup always produced more food words than Yelp Fusion API. The low word count from Yelp Fusion API is attributed to the restrictions from Yelp of only providing three review excerpts for each restaurant. By using Beautiful Soup instead, we are able to gather more reviews and return additional food items. Another experiment analyzed the reliability of the blacklist we created to filter out non-food words. In the results from experiment 2, the large difference between using the blacklist and not using the blacklist was expected because the blacklist is a very long list that encompasses most of the common unrelated food words. The data from this experiment confirmed that the blacklist did improve the relevancy of the food items by a large margin and the reliability of the app overall.

5. RELATED WORK

There are many methods of applying sentiment analysis in online social platforms to observe public opinion and predict potential sways in financial markets [4] [5] [6]. Although both the opinion mining project and Dishcovery use sentiment analysis for prediction, Dishcovery web scrapes from Yelp while the opinion mining project scrapes from Twitter. These two platforms, though formatted similarly in short paragraphs of user input, are functionally distinct and have different algorithms of correlation and accuracy contributing to the final prediction. Both programs are bound to have inaccuracies because they scrape solely from one platform. Public sentiment on Twitter is likely to be different from public sentiment on other social media platforms, so stock market changes cannot be determined from Twitter alone. Similarly, there are other restaurant review websites aside from Yelp that may have different review sentiment.

This paper highlights the use of web-scraping software in searching for grey literature, referring to any research that is either unpublished or non-commercially published [7]. Since searching for grey literature online is often “time-consuming” and “challenging”, web scraping can be very useful in “extracting data from multiple pages of search results.” Unlike Dishcovery, the grey literature search is not limited to any particular website and does not use sentiment analysis to recommend users particular literature. Instead, it finds patterns in the documents to extract. Dishcovery is confined to scraping on Yelp and analyzes the sentiment of reviews. Although the main purpose of these projects is dissimilar, they both use web scraping to gather the bulk of their data.

Although similar to the first related work, this paper discusses the importance and potential for polarity detection and gives more of a general overview of opinion mining. Like Dishcovery, opinion mining would likely utilize both web scraping and sentiment analysis to make predictions [8]. Though opinion mining can be extremely helpful in the financial sector, its predictions need to be very accurate and political and social preferences cannot be determined from the internet alone. The same hurdle applies to Dishcovery: although Yelp plays a big role

in public sentiment for the restaurant experience, good dishes cannot be predicted from Yelp alone. Word of mouth is still very much prevalent, which simply cannot be accounted for in opinion mining or Yelp recommendations.

6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a simple and efficient way for people to discover food dishes from different restaurants. Our mobile app encourages users to visit local restaurants and go out of their comfort zones to try food items from different cuisines. Aided by data from Yelp, this application gathers reviews, analyzes the overall tone of the reviews, and filters out the popular dishes.

Some of the challenges we faced when making our app include having access to too few Yelp reviews and having difficulty identifying food items. To address these problems and further elaborate on the methods we used to solve them, we performed two evaluations. Our first evaluation of the different approaches to obtain and analyze user reviews shows that Beautiful Soup consistently returns more food items than Yelp Fusion API. To illustrate, the average amount of food words Beautiful Soup analyzes is 105.7 while the average amount Yelp Fusion API analyzes is 3.6. The increased amount of data provided by Beautiful Soup allows for more accurate sentiment analysis of user reviews. In our other evaluation, we found that the blacklist does a good job of removing nonfood words and improves the accuracy of the output. Our results show that the filtering process filters out an average of 51% of the unrelated words from Google Natural Language API. Both of the results in our evaluations, as provided in Figure 5 and Table 1, prove that using web crawling with Beautiful Soup and the blacklist we made are useful solutions for the challenges we initially encountered.

For future work, we plan to improve our food sorting method and blacklist. We noticed that the word sorting groups in Google Natural Language API are too broad for the purposes of our app, so other natural language processing services can potentially be used to improve the efficiency of sorting. Although the current approach is applying a blacklist for narrowing down the food words from Google Natural Language API, this method is inefficient and requires us to look through the result output. Furthermore, it is impossible for the blacklist to filter out all the nonfood items from the Yelp reviews.

In the future, Dishcovery could be capable of identifying bots and fake negative reviews by competing restaurants. Also, we are considering creating a machine learning algorithm that learns to identify food items and works alongside the natural language processing service. This will improve the accuracy of the list of food words for each restaurant and possibly eliminate the blacklist. Thus, a more effective and adaptable solution could be implemented.

REFERENCES

[1] Duan, Wenjing, Bin Gu, and Andrew B. Whinston. "Do online reviews matter?—An empirical investigation of panel data." Decision support systems 45.4 (2008): 1007-1016.

[2] Chaves, Elisabeth. "The Internet as global platform? Grounding the magically levitating public sphere." New political science 32.1 (2010): 23-41.

[3] Rose, Susan, et al. "Online customer experience in e-retailing: an empirical model of antecedents and outcomes." Journal of retailing 88.2 (2012): 308-322.

[4] Feldman, Ronen. "Techniques and applications for sentiment analysis." Communications of the ACM 56.4 (2013): 82-89.

[5] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." LREc. Vol. 10. No. 2010. 2010.

[6] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012): 1-167.
[7] Haddaway, Neal R. "The use of web-scraping software in searching for grey literature." Grey J 11.3 (2015): 186-90.
[8] Suganya, E., and S. Vijayarani. "Sentiment Analysis for Scraping of Product Reviews from Multiple Web Pages Using Machine Learning Algorithms." International Conference on Intelligent Systems Design and Applications. Springer, Cham, 2018.
[9] Kumar, Naveen, et al. "Detecting review manipulation on online platforms with hierarchical supervised learning." Journal of Management Information Systems 35.1 (2018): 350-380.
[10] Nair, Vineeth G. Getting Started with Beautiful Soup. Packt Publishing Ltd, 2014.
[11] Zunnurhain, Kazi, and Arian J. Gonzalez. "Enhancement of User Experience with Mobile App." Proceedings of the 47th International Conference on Parallel Processing Companion. 2018.
[12] Kuzmin, Nikita, Konstantin Ignatiev, and Denis Grafov. "Experience of Developing a Mobile Application Using Flutter." Information Science and Applications. Springer, Singapore, 2020. 571-575.
[13] Habib, Md. "Web and Android Application Developer" in Popular IT Limited." (2018).
[14] Mellet, Kevin, et al. "A "democratization" of markets? Online consumer reviews in the restaurant industry." Valuation Studies 2.1 (2014): 5-41.
[15] Hlee, Sunyoung, et al. "An empirical examination of online restaurant reviews (Yelp. com): Moderating roles of restaurant type and self-image disclosure." Information and communication technologies in tourism 2016. Springer, Cham, 2016. 339-353.

TRUEREVIEW: AN OBJECTIVE PRODUCT RATING AND RANKING BASED ON USER REVIEWS USING AI AND DATA ANALYTICS

Lirui (Harrison) Huang[1], Yu Sun[2], Fangyan Zhang[3]

[1]University High School, Irvine, CA 92612
[2]California State Polytechnic University, Pomona, CA, 91768
[3]ASML, San Jose, CA, 95131

ABSTRACT

Since the moment human step into the high technology era, the world that people live in has changed subversively. As more and more unprecedented advancement being discovered, modern life of today's people is now incredibly convenience. However, when the internet has played a pivotal role in everyday life, among the information that we are getting, countless of them are fraudulent. This paper designs a website to filter the influence of fake comments made to the product. We applied our application to the most commonly used shopping platform, Amazon, and conducted a qualitative evaluation of the approach. With a large amount of trained data, the sentiment analysis program can filter the fraudulent and odd comments and to give a more accurate score regarding a product by reading through each comment. This will definitely help consumers determine whether a product is fine or not.

KEYWORDS

Product review, website, machine learning, JavaScript, HTML

1. INTRODUCTION

Online shopping is grueling, picking one from a giant product that fit the most can take a lot of time [1]. Imagine searching for a simple item and thousands of tabs emerged out, not a very pleasant experience [2]. Among them, there are things people cannot see that are behind the curtain. There are people making money by stealing other people's information; there are also people making money by posting false information. Fraudulent pieces of information are omnipresent on the internet, it is hard to not put yourself in a jeopardizing situation when you have no idea what your circumstance is. But people nowadays cannot really live without using an online shopping platform once in a while. So, filtering and summarizing a product is a delirious need on the internet. The ranking websites are now ubiquitous, there is IMDB [3] for movie and TV show rank, there is U.S. News for best university rank, but there is not really a shopping rank that can help customers make the right decision, so there comes The Real Shopping Ranks! There has to be some improvement in the shopping experience. A well generated informational website should be given to customers for them to make a better shopping decision. While making the information on shopping platforms more transparent, it is also a step forward for making a better cosmopolitan online community.

Some of the ranks generating techniques and systems that have been proposed to generate information, for such as movies shows or colleges, all around the internet, which allow the user to

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 89-95, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101209

see the translucent information. However, the biggest issue is that there is no rank is generated specifically for improving the shopping experience [4]. The only rank customers can use is by the shopping platform, which usually includes sponsors and fake promotions and is not trustworthy. Also, their implementations are usually biased, samples with more people buying them will increasingly get more reviews, and so on. Even if people trust the rank given by a shopping platform, the options for people are too less to see products from different perspectives; such as popularity, price, delivery time, are not available. Many times, customers may seek help from other buyers, but they are no believable either. Some of them are paid to write positive comments, or on the other way, paid to write extremely negative comments. Customers who really want to buy often swamp while reading long comments that seem to take forever. The tools that are available are so scarce, because of that, many people appeal that they would rather go shopping in store without being influenced by others. Internet is like a deep and dark mire, people with their own knowledge usually lead them to the wrong direction. A data- based rank is needed to be invented for customers as a reference when buying goods. This can not only save customers' time, help them make the right decision, but also, to discover the potential of sentiment analysis when applied to real life situation[5][6][7].

In this paper, we follow the same line of research by some of the present ranks like IBDM, US News, and Amazon's ranking. Inspired by shopping experience in Amazon, most of the data are extract from Amazon.com. There are some good features that can indicate some useful information about their goods, such as the large sample scale, pictures in different sizes, and score movements. Upon the great number of products in Amazon, we want to make it less messy than it is right now. Our goal is to provide a detailed, trustworthy shopping ranks that can allow customers to swiftly make the right shopping decision. Our rank can generate a non- biased rank by collecting product information and worth trust comments after the filtering program. The score can indicate whether a product is fine or not. This is meant to save customers time and prevent them from doing too much research and bought the product they do not actually like. Time is your money, efficiency is your life, it is especially important to do things fast in this modern society. Rank show everything, without jumping over tabs, a clean table will help people find the one they want.

In the application scenario, we demonstrate how machines can combine and generate information for shopping usage. We show the usefulness of our approach by a comprehensive case study on the evolution of sentiment analysis. Sentiment analysis is to processing natural language and study affective states from text. The machine has been applied sentiment analysis function and can give a positive/negative point for texts. By using that, machine can break down each part of the text and read through it to generate a summarizing average point. The un matching of comment and stars will no longer exist in the view of average point system. This allows customers to identify the quality without reading through all the comments.

The paper is structured as follows: Section 2 includes the challenges we faced during the experiment and designing the sample; Section 3 on the next focuses in detail on our solution in relation to our challenges; Section 4 presents the relevant details about the experimental structure, following by section 5 are related works to this research. And the end of the paper is section 6, which gives the concluding remarks, as well as pointing out the significance of our research and future works of this project.

2. CHALLENGES

2.1. CHALLENGE1

When trying to research in a new field, there will always be some issues that came up. Creating program and researching a topic requires patient. And I faced many unprecedented issues throughout the whole process. One of the issues that I struggled with was finding the crawling information. Since I am new to web design field, it took me long time to get use to the form of HTML. Also, as I was organizing the code, I often encounter the time when crawling fails. Because I was not familiar with the form, I faced many emerging problems. Fortunately, I overcome with challenges by keep trying. I can keep on doing what I was doing because I know that Sentiment Data Analytics Can offer people in real life a doable solution, bring them convenience and solve the present real-life problem. It is challenging to do something that are out of our comfort zone, but being able to contribute to the computational study and discover more potential about artificial intelligence is such a pleasure for us to do.

2.2. CHALLENGE2

When going into the real program steps, debug is a huge challenge. It happens sometimes that the program suddenly closed when it was working properly moment before. It took us hours to figure out different types of problems, even sometimes it could be little things like spacing in python. And this requires us to do a tremendous amount of independent research to construct this program. Working all the way to be able to finish the program took a lot of physical and mental strength, but the feeling of fullness when finish building the main structure is the feeling that cannot be told byword.

2.3. CHALLENGE3

As we go in depth to the research program, I noticed how lack my technical programming knowledge was. Building a intact program requires the programmer to be able to switch back and forth between different languages. So, in order to get the program on going, I spent much of time learning new programming language that I had not get in touch with. I spent a lot more time dive deeper into this topic and see the essence of it. It was difficult to balance school work and extra research study, though, I eventually made it. Besides all of the hard works, I learned many through the study. About how to do independent research and how to do a thing steadfast. Challenge will be over but the lesson from that will remain.

3. SOLUTION

The real shopping ranks is a computer-generated sentiment analysis ranking system [13]. With the navbar on the top, table, and detailed description on the bottom, product ranking detail will show up once keywords were typed and loading was done.

In current phenomenon, loading may take few minutes, and once the loading is finished, table with product details will be created along with each section Images, names, prices, and Amazon's ratings are utilized from the original product page. Other than that, the pivotal part of the rank is the sentiment score. Scores are generated by processing reviews using Sentiment Analysis. AI machine learning using the Textblob library can gives us a number range between 0 and 1 regarding to the positivity of the sentence segment [14][15]. With a large amount of trained data, we can filter the fraudulent and odd comments, and give a more accurate score using the product comments. Nonetheless, the score that shows on the website are not just about the sentiment score

given by our machine. Combining partial of Amazon score and sentiment score, the final score will be partially of both and will be a highly valued number for users. This will help consumers determine which product is truly the best.

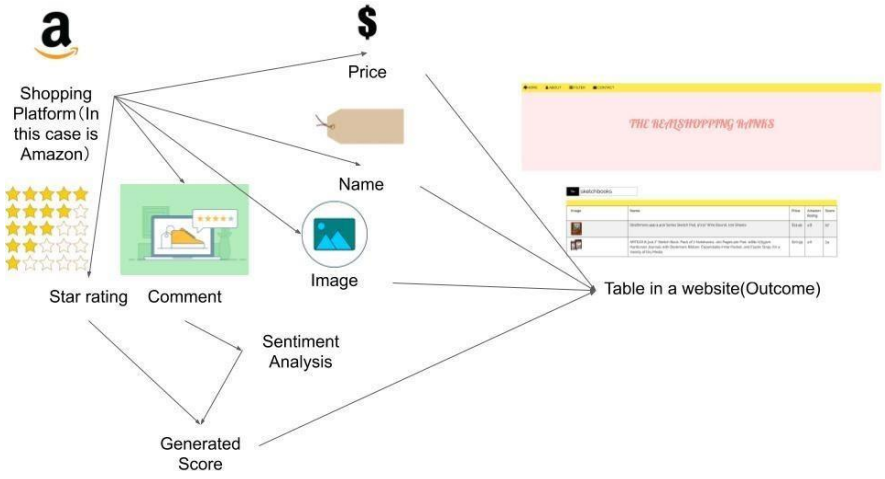


Figure 1: overview of TrueReview

The real shopping ranks were designed by using multiple coding languages, such as Python, HTML, JavaScript, and CSS. Each played an important role in this program. We also imported many different libraries to each of the programming languages which helps us more efficiently built the program. Python was mainly used in the crawling process; data were extracted from the Amazon shopping platform. We essentially used Textblob [8] and Beautiful soup [9] for most parts of our python code. Beautiful soup first helped us abstract the pieces of information so that Textblob can analyze textual information and turn it into sentiment points. The points were range between 0 and 1, 0 being the least positive and 1 being the most positive. Python was able to go into the website and captured any useful information for our project. HTML, on the other hand, played the most important role in our program.

In this platform, JavaScript and CSS were able to customize the outcomes of our program and transform it into useful work. Going into JavaScript, it can functionate all parts of the information and made the website became “live”; Liking many logical functions with data were implementing in here. CSS, on the other hand, format the website and display the functions and information to each proper place. In the final display, we would not be able to create this neat website without using many of the lovely functions from W3.CSS. On the demo picture below of our actual website, we contained three main sections and one main ranking function. Three sections are Home, About, how it works and Contact information. Each resemble the detail of the story behind the program. Beside some customized website outfit, the actual ranking table is in the pivotal part of the website page. It will be able to display once searching and loading was complete. To overview the process, the initial data information was captured by crawler using Python, beautifulsoup. Information were then processed by textblob library. And JavaScript connect the imported data with HTML.

Finally, the website was displayed by CSS with the helping from W3.CSS library. By using different sorts of methods and platforms, the real shopping ranks can only work properly and sending useful information to more users.

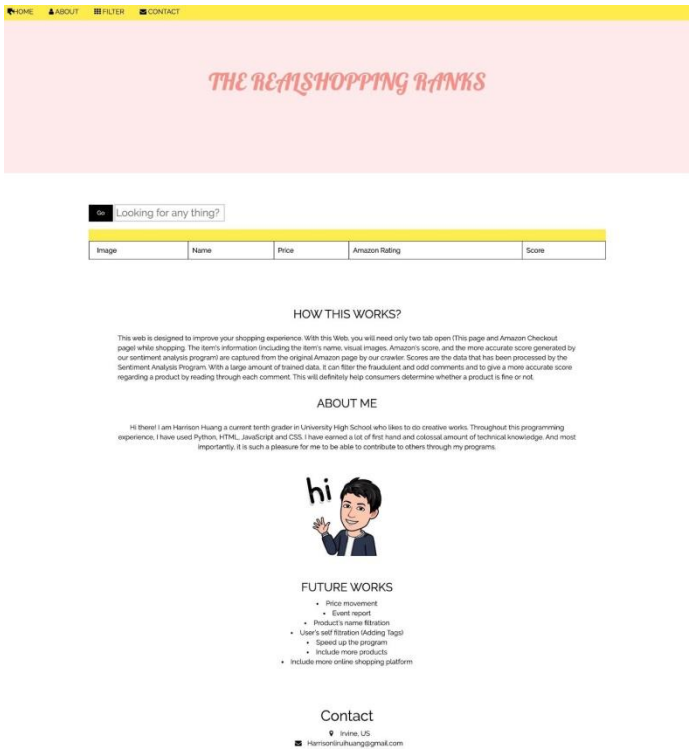


Figure 2: Interface of TrueReview

4. EXPERIMENT

The accuracy of the extracting data is extremely important. An analysis-based program can only be trustworthy when the data are extremely accurate. In our case, we want to know the percentage correctness of our crawling information. To do so, we hand record name, price, star rating of 10 products. We let the crawler to get those ten products' basic information to see if it is highly accurate. We will recode the correct number of information and report it in percentage.

After the running through the crawling process, the result was out. The data was 100% accurate compare to the hand record data. The crawling process is highly accurate. This helps us support the sentiment analysis process after the data collection, which make the analysis more accurate andmoretrustworthy.Theaccuracyofthedataisextremelyimportant,iftheinformationextract from the page is not exactly the same, then any more future analysis is not believable. Fortunately, our product information is accurate and our rating system is very authentic.

To test the accuracy of our program, we collected product comments from total of 100 products. Using our crawler, we were able to utilize the product information from the original shopping platform. But in this experiment, we focused on the product review data. We trained the machine and made it read through product comments and to compare the result with the original rating stars.

After the collection of data and the processing of the scores, the results came out. Most of the sentiment scores are lower than the actual score to our surprise. Because of the limitation of Amazon's rating system, there are only 5 stars to choose for rating a product. By using our sentiment system, we are able to range the score more widely and make the score more accurate and fit to what reviewers' original thoughts.

By doing the two experiment above, we were able to prove that our essential data and extensive data are accurate. The accuracy of the extracting data is huge. An analysis base ranking program cannot be trusted if the information given are fraudulent. The purpose of our program is to give accurate information and point out right directions for customers. We would not want our data to be inaccurate and trick the customers for second time. Although there are still space to improve, overall, we are very satisfied about how accurate our crawling and analysis system are.

5. RELATED WORK

Pranking with Ranking is a research paper written by Koby Crammer and Yoram Singer [10]. In the paper, they discuss about the evaluation of ranking system. They tried to find a more accurate ranking system while we are trying to create a sentiment based ranking system with the actual existed data. Our project focus more on generating product review rating which can give customers a better look at the products' quality. On the other hand, Koby and Yoram's work based on creating a new way of generating the ranks which covers more various field.

Adaptive ranking system for information retrieval is a research paper adapted by Shih-Chio Chang, Anita Chow and Min-Wen Du [11]. They talked about how importing customized weighted system and tags into ranks can make it more accurate. Our project, on the other hand set a 3:7 rate weighted correlation between sentiment score and amazon's score. We can easily provide a general information while their project can give more possibility to customers and make customers decide what kind of rank they want to see.

The team that research on Ranking products by mining comparison sentiment dive deep in the sentiment analysis in ranking products [12], and this is actually very similar to what we are doing. In their program, provide different ways of ranking a product instead of actually doing a program. We, on the other hand, extract all reviews and average them. In their research, they conclude multiple ways that sentiment analysis can be involved into ranking system. And for us, although our program is not perfect, but we were able to actually apply sentiment analysis to ranking system and send useful message to other people.

6. CONCLUSION AND FUTUREWORK

To conclude, there many gruelling issues in modern shopping platforms, customers may spend hours finding a product that fits them the most. The main reasons are that the unrelated products, fraudulent comments, and too many buying options slow down their shopping speed. Time is even more valuable than gold, in wanting to help others saving time and realizing how bad online shopping experience in our daily life is, we proposed a sentiment-based ranking program. We wanted to conclude product information from multiple websites into one single table as well as a generated sentiment score. By using HTML, Python, JavaScript, and CSS, we were able to construct the main structure of the rating system. After crawling data by Python and Textblob, analysis text using beautifulsoup we turned the piece of information into an actual website by using JavaScript, HTML, CSS, and W3.CSS library. Along with wrapping up the program, two experiments were being done. One proof that our crawling information was accurate and other proof our sentiment score being real and trustworthy. Since we are doing a data-based ranking analysis program, we will need to be giving out high valued information. A shopping program will not be successful if it cannot give accurate information and point out the right direction for customers. By performing experience, we were able to prove experimentally that our program is believable.

In the current state, our program made many compromises which led to limitations of our program. The running speed of our program is at a low level. Because of that, we can only have two products comparison in one time, which limit many of the potential possibility of our program. Similarly, because of the loading speed, we cannot let our program to complete reading every single comment under one product, which made the sentiment score a little bit off of where it should be.

Speeding up our program is the main future work we were going to try to solve in the future. The process of crawling data, processing scores, and loading the page takes a lot of time, but we must optimize our algorithm to speed up our program. With the speeding up of the program, we could make the website more valuable and discovering its more potential possibilities.

REFERENCES

[1] Limayem, Moez, Mohamed Khalifa, and Anissa Frini. "What makes consumers buy from Internet? A longitudinal study of online shopping." IEEE Transactions on systems, man,and Cybernetics-Part A: Systems and Humans 30.4 (2000):421-432.
[2] Morales, Lisa. "Method and System for Online Shopping and Searching For Groups OfItems." U.S. Patent Application No. 12/907,617.
[3] Dodds, Klaus. "Popular geopolitics and audience dispositions: James Bond and the internet movie database (IMDb)." Transactions of the Institute of British Geographers 31.2 (2006): 116- 130.
[4] Khandelwal, Shashikant. "Method for relevancy ranking of products in online shopping." U.S. Patent No. 8,510,298. 13 Aug.2013.
[5] Bakshi, Rushlene Kaur, et al. "Opinion mining and sentiment analysis." 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE,2016.
[6] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012):1-167.
[7] Maas, Andrew, et al. "Learning word vectors for sentiment analysis." Proceedings of the 49th annual meeting of the association for computational linguistics: Human languagetechnologies. 2011.
[8] Loria, Steven. "textblob Documentation." Release 0.15 2(2018).
[9] Nair, Vineeth G. Getting Started with Beautiful Soup. Packt Publishing Ltd,2014.
[10] Crammer, Koby, and Yoram Singer. "Pranking with ranking." Advances in neural information processing systems.2002.
[11] Chang, Shih-Chio, Anita Chow, and Min-Wen Du. "Adaptive ranking system for information retrieval." U.S. Patent No. 5,321,833. 14 Jun.1994.
[12] Sun, Jian-Tao, et al. "Ranking products by mining comparison sentiment." U.S. PatentNo. 8,731,995. 20 May2014.
[13] Edlund, Stefan B., et al. "Metadata search results ranking system." U.S. Patent No. 6,718,324. 6 Apr.2004.
[14] Michie, Donald, David J. Spiegelhalter, and C. C. Taylor. "Machine learning." Neural and Statistical Classification 13.1994 (1994):1-298.
[15] Alpaydin, Ethem. Introduction to machine learning. MIT press,2020.

A NOVEL INDEX-BASED MULTIDIMENSIONAL DATA ORGANIZATION MODEL THAT ENHANCES THE PREDICTABILITY OF THE MACHINE LEARNING ALGORITHMS

Mahbubur Rahman

Department of Computer Science, North American University,
Stafford, Texas, USA

ABSTRACT

Learning from the multidimensional data has been an interesting concept in the field of machine learning. However, such learning can be difficult, complex, expensive because of expensive data processing, manipulations as the number of dimension increases. As a result, we have introduced an ordered index-based data organization model as the ordered data set provides easy and efficient access than the unordered one and finally, such organization can improve the learning. The ordering maps the multidimensional dataset in the reduced space and ensures that the information associated with the learning can be retrieved back and forth efficiently. We have found that such multidimensional data storage can enhance the predictability for both the unsupervised and supervised machine learning algorithms.

KEYWORDS

Multidimensional, Euclidean norm, cosine similarity, database, model, hash table, index, K-nearest neighbour, K-means clustering.

1. INTRODUCTION

Searching, classifying, predicting the multidimensional data have been the most interesting applications of today's machine learning algorithms [1],[2]. As the number of dimensions, size of data increase, so does the overall complexity of pre-processing, searching, classifying or predicting such data sets using machine learning algorithms [3]. To overcome such complexity, we have introduced a data organization model that can enhance overall learning outcome while keeping the necessary information associated with the learning in a reduced dimensional space.
The multidimensional data requires high dimensional data storage, access models. There are several high dimensional data organization models as SR-tree [4], R-tree [5], kd-tree [6]. These are tree-based data organization models that follow the complexity of tree while inserting, deleting multidimensional data. One of the recent patents from Google [7] has explained the methods of searching the multidimensional dataset by mapping the data to the nodes in the multidimensional hyperspace. However, we are interested to design a data organization model for the multidimensional data in a reduced space that has efficient data access, manipulation schema along with the stored information that can be used in the learning outcome of the machine learning algorithms.

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 97-108, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101210

To design such multidimensional data organization model, we have introduced a multidimensional index-based data organization model that can be used in the machine learning algorithm. This data organization model is ordered on the Euclidean norm and stores the original index position along with the Euclidean norm of the datasets. This ordering allows searching in logarithmic time and index mapping to the original database in constant time. As a result, the overall data organization model provides a reduced space along with the original index-Euclidean norm pair that can be reused in the machine learning computation enhancing the overall learning outcome.

The following sections and subsections explain the overall data organization model, its applications on the supervised, unsupervised machine learning models, implementations, results and analysis.

2. DATA ORGANIZATION MODEL

The data organization model is a two-dimensional map (e.g. reduced space) of the original multidimensional dataset (e.g. Fig. 1) that keeps the information (e.g. original index position and Euclidean norm) by the increasing order of the Euclidean norm of the datasets. As a result, this information can be accessed and reused efficiently. As for example, the Euclidean norm can be reused in the Euclidean distance [8], cosine similarity [9] calculation etc. Additionally, such ordered data organization model can provide the searching spaces for a point of interest that can enhance finding its nearest neighbours, clusters etc. Thus, the ordered index-based database provides important information that can enhance the predictability of the machine learning algorithms. We have provided further details in the following subsections.

Index	Dim 1	Dim 2	Dim 3	...	Dim n
0	a_1	a_2	a_3	...	a_n
1	b_1	b_2	b_3	...	b_n
2	c_1	c_2	c_3	...	c_n
...	...	...	...	...	...
m	z_1	z_2	z_3	...	z_n

Original N Dimensional Datasets

Index	Original Index	Euclidean Norm
0	2	$\sqrt{c_1^2 + c_2^2 + \cdots + c_n^2}$
1	0	$\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$
2	m	$\sqrt{z_1^2 + z_2^2 + \cdots + z_n^2}$
...	...	...
m	1	$\sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}$

Ordered index-based Database

Figure 1. Original multidimensional datasets and ordered index-based data organization model.

2.1. Ordered Index-Based Database

The ordered database is organized from the original data source. It is ordered by the Euclidean norm and stores both the original index position and the Euclidean norm. As a result, the Euclidean norm can be mapped to the original index position of the data directly. This index and Euclidean norm can be stored as a pair to a hash table that forms the ordered database (e.g. Fig. 1).

The Euclidean norm of an n dimensional vector (e.g. x) has the following formula:

$$\text{Euclidean\_norm}(x) = \sqrt{x1^2 + x2^2 + \cdots + xn^2} \quad (1)$$

The Euclidean distance between the two multidimensional vector (e.g. x, y) has the following formula:

$$\text{Euclidean\_distance}(x,y) =$$

$$\sqrt{(x1 - y1)^2 + (x2 - y2)^2 + \cdots + (xn - yn)^2} \quad (2)$$

$$= \sqrt{x_1^2 + \ldots + x_n^2 + y_1^2 + \ldots + y_n^2 - 2\sum_{i=1}^{n} x_i y_i^T} \quad \text{[e.g. expanding } (x-y)^2 \text{ formula] (3)}$$

$$= \sqrt{Euclidean\_norm(x)^2 + Euclidean\_norm(y)^2 - 2\sum_{i=1}^{n} x_i y_i^T}$$

[e.g. from (1)] (4)

Hence, the unknown part of formula (4) is the dot product of the two vectors. The rest is available from the Euclidean norm of the database. This can be reused in calculating the cosine similarity of the two multidimensional vectors as suggested by the following cosine formula:

$$\frac{\sum_{i=1}^{n} x_i y_i}{Euclidean\_norm(x) \quad Euclidean\_norm(y)} \quad (5)$$

The denominator part of (5) is also available from the database. Additionally, such organization of the multidimensional data sets provides the nearest neighbour searching space for a dataset. It means that the nearest neighbours for a data can exist within a minimum distance between two immediate neighbours of the dataset as it is explained in Figure 2.

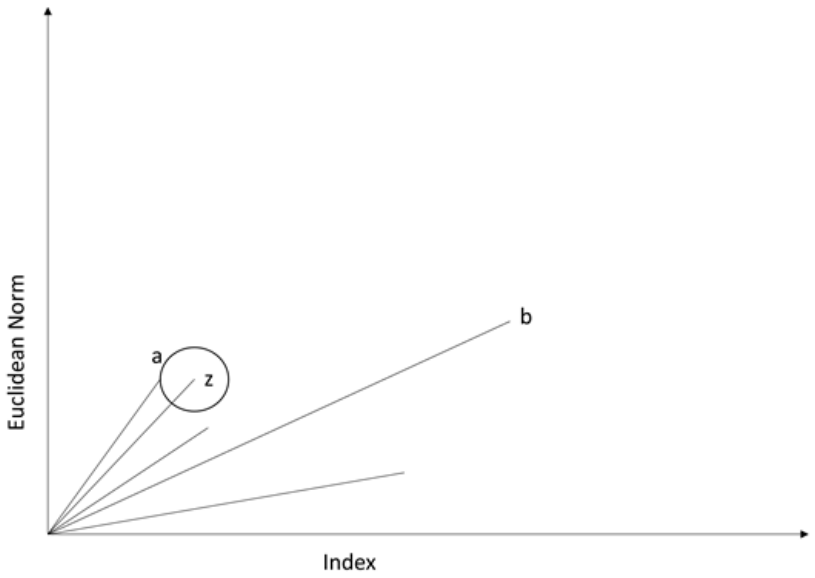


Figure 2. Index vs Euclidean Norm of the ordered database. The nearest neighbours of z exist in the circle having the radius of the minimum distance of its immediate neighbour (e.g. a).

We find from the Fig. 2 is that the two immediate neighbours are one index up and down (e.g. left and right of z in Fig. 2) respectively from the index position of the data of interest in the ordered database. Next, the minimum Euclidean distance between these two immediate neighbours provides the searching space (e.g. radius of the circle in Fig. 2) for the nearest

neighbours of the dataset of interest. The index positions of the dataset in the ordered database can be retrieved in logarithmic time by the binary search and then their original index position in the database in constant time (e.g. Fig. 1). As a result, the ordered database can provide the searching space information that can be used in the supervised and unsupervised machine learning algorithms. The overall searching space for a point of interest along with the reusage of the above computations has been listed as Algorithm 1: Nearest Neighbour Searching Algorithm (NNSA) below.

ALGORITHM 1: Nearest Neighbor Searching Algorithm (NNSA)

Function NNSA (x, ordered_db, original_db):
Input: A point (x), ordered database (ordered_db), original database (original_db).
Output: Two immediate neighboring index (e.g. ui, di) of x.
index_pos_x, euclidean_norm_x = ordered_db[x]
index_up = index_pos_x+1
index_down = index_pos_x-1
xu, euclidean_norm_up = ordered_db[index_up]
xd, euclidean_norm_down = ordered_db[index_down]
x1 = original_db[xu]
x2 = original_db[xd]

$$d1 = \sqrt{\text{euclidean\_norm\_up}^2 + \text{euclidean\_norm\_up}^2 + \cdots - 2\sum_{i=1}^{n} x_{i.} x_{1i}^{T}}$$

$$d2 = \sqrt{\text{euclidean\_norm\_down}^2 + \text{euclidean\_norm\_down}^2 + \cdots - 2\sum_{i=1}^{n} x_{i.} x_{2i}^{T}}$$

d = min (d1, d2)
d_up = euclidean_norm_x - d
d_down = euclidean_norm_x+d
ui = ordered_db[d_down]
di = ordered_db[d_up]
return ui, di

2.2. An Organized Data Model Based Supervised Learning

The supervised learning model requires the labelled data sets for learning and predicting the outcome. This labelling information is used during the training steps. The model executes different types of computations (e.g searching, comparing etc.) at the training phase. We have explained how the ordered index-based database can enhance such computations as well as learning the outcome in this subsection.

We have used the K nearest neighbour (KNN) supervised learning model [10] as this is one of the most fundamental supervised learning algorithms. In KNN algorithm, the neighbours for a new point of interest in the dataset are the K closest instances. To locate such K nearest neighbours, the algorithm must first calculate the Euclidean distance between each record of the dataset and the point of interest. Finally, it must sort all the datasets by their distances to the point of interest to get the K required instances. There are several attempts to optimize the KNN using a multi-step query processing strategy [8], GPU computing [11]. However, we are interested to optimize the learning outcome of the KNN using the NNSA.

The NNSA can provide the KNNs for a point of interest while searching around the point. The searching space can be increased by the same minimum amount of the radius of the searching

space if the required number of nearest neighbours is not found within the initial searching space. Additionally, the ordered database can be reused again to find the KNNs for a new point of interest. Thus, this model helps avoiding repetitive expensive sort, computation operations. The overall KNN algorithm using the NNSA has been listed as Algorithm 2: KNN Algorithm using NNSA below.

ALGORITHM 2: KNN Algorithm using NNSA

```
Function K_NNSA (x, ordered_db, original_db, k, neighbor_list):
Input: A point (x), ordered database (ordered_db), original database (original_db), number of neighbors (k) of x, list of nearest neighbors (neighbor_list) of x.
Output: List of nearest neighbors (neighbor_list) of x.
ui, di = NNSA (x, ordered_db, original_db)
neighbor_counter = 0
neighbor_list = []
u0 = ui
d0 = di
while neighbor_counter  <= k  do
        if u0 – k <= 0 then
                u0 = u0 – k
        end if
         if d0 + k < length(original_db) then
                d0 = d0 + k
        end if
        up = u0
        down = d0
        while up <= down do
                index_pos_up = ordered_db[up]
                label_pos_up = original_db[index_pos_up]
                if  label_pos_up == x[label_index] then
                        neighbor_counter = neighbor_counter + 1

        neighbor_list.
                add(original_db[index_pos_up])
        end if
        index_pos_down = ordered_db[down]
        label_pos_down = original_db[index_pos_down]
         if label_pos_down == x[label_index] then
                neighbor_counter = neighbor_counter + 1
                neighbor_list.
                add(original_db[index_pos_up])
        end if
                up = up + 1
                down = down -1
                if length(neighbor_list) == k then
                        return neighbor_list
        end while
end while
return neighbor_list
```

2.3. An Organized Data Model based Unsupervised Learning

The unsupervised learning algorithm does not require labelled datasets [12], rather can learn from the unlabelled datasets. The existing K-means clustering algorithms calculate the Euclidean distance among the centroids of the clusters and the points until the convergence [13], [14]. As a result, the overall computation involved here are repetitive and thus, expensive. However, we have organized the data such that this repetitive calculation can be reduced from multiple times to single one. Additionally, the NNSA can also determine the minimum no. of clusters and finally, this can be used to determine the optimum no. of clusters.

The ordered database can provide the minimum number of clusters by selecting the mid index point (e.g. centroid) of the ordered database and then determining its searching space (e.g. Fig. 3) by the NNSA. Next, this step can be repeated for the rest of the datasets above and below the searching space of the mid index point until it covers the ordered database. Finally, this minimum no. of clusters can be reduced to the optimum no. of clusters by applying agglomerative [15], divisive [16], hierarchical [17] clustering algorithms.

The optimum number of clusters can be determined where the number of clusters and within cluster sum of squares (WCSS) [18] begins to level off (e.g. elbow method). The WCSS is defined as the sum of the squared distance between each member of the cluster (e.g., x) and its centroid (e.g., c) and has the following formula:

$$\sum_{i=1}^{n}(xi - ci)^2 \qquad (6)$$

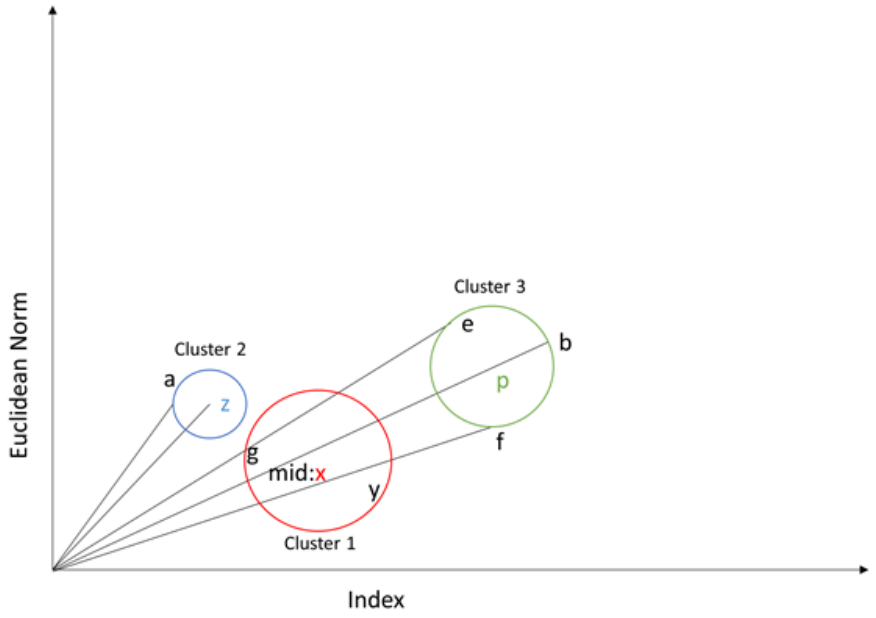


Figure 3. Cluster formation by the ordered index-based database. Cluster 1 starts from the mid indexed vector x (e.g. mid: x in the figure). Similar vectors belong to different clusters depending on the Euclidean distance between the centroid and the cluster

Our proposed ordered database can provide the Euclidean norms to the formula (6) and thus, can enhance the overall computation. The overall K-means algorithm using the NNSA has been listed as Algorithm 3: K-means Algorithm using NNSA below.

ALGORITHM 3: K-Means Algorithm using NNSA

```
Function K-Means_NNSA(x, ordered_db, original_db, centroid_list):
Input: Midpoint (x), ordered database (ordered_db), original database (original_db), list of centroids (centroid_list).
Output: List of centroids (centroid_list).
up, down = NNSA (x, ordered_db, original_db)
centroid_list = []
centroid_list.add(original_db[x])
up = up / 2
n_points = down – up
max_size = length(original_db)
while up > 0 do
        l, r = NNSA (up, ordered_db, original_db)
        l = l – (n_points/2)
        if l < 0 then
                l = 0
        end if
        r = r + (n_points/2)
        if r >= down-1 then
                r = down-1
                add the point at the up index along with l, r to the centroid list
                centroid_list.add(original_db[up])
                break
        end if
        up = up / 2
end while
down = down + (n_points/2)
while down < max_size do
        l, r = NNSA (down, ordered_db, original_db)
        l = l – (n_points/2)
        if l < 0 then
                l = 0
        end if
        r = r + (n_points/2)
        if r >= max_size then
                r = max_size - 1
                centroid_list.add(original_db[down])
                break
        end if
        down = down + n_points / 2
end while
return centroid_list
```

3. IMPLEMENTATIONS

All the algorithms above have been implemented in python version 3 [19]. The ordered database of the NNSA is a python dictionary of the pair of original index position and Euclidean norm of the multidimensional dataset. As result, this original index position can be mapped to the position of the dataset in the original database directly. We have used the existing python libraries to sort the dictionary by the Euclidean norm. The above three algorithms are three different python

subroutines where the KNN and K-Means subroutines both interact with the NNSA subroutine. Finally, all the subroutines along with the databases have been incorporated into a python class file.

4. RESULTS

We have introduced an index-based data organizational model for the multidimensional data in a reduced space to improve the predictability of the learning algorithms. The reduced space is a two-dimensional space (e.g. a pair of original index and Euclidean norm) for the multidimensional dataset. The organization of this reduced space requires the dataset to be sorted by the Euclidean norm. Hence, there are some computations involved in the organization of the data model. The computations include calculating the Euclidean norm, sorting by that norm and finally, inserting as a pair of original index and Euclidean norm of the dataset into a dictionary. We have defined the time required for such computations as the data model organization time and analysed the effect of increasing the no. of dimensions, sizes of the datasets (e.g. Fig. 4, 5 respectively).

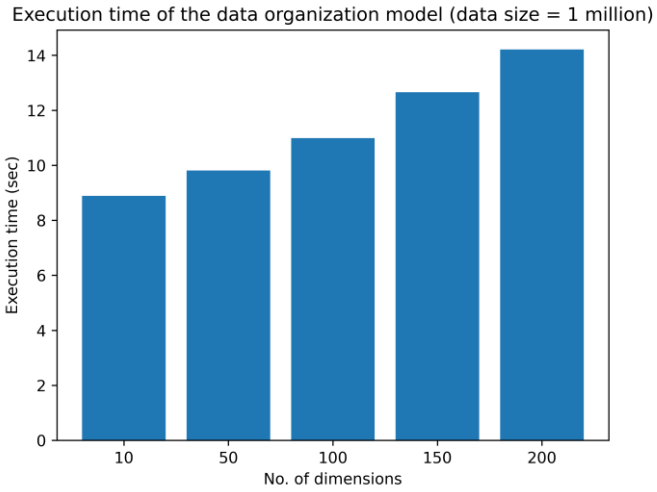


Figure 4. No. of dimensions vs Execution time of the data organization model for a fixed size data (e.g. 1 million).

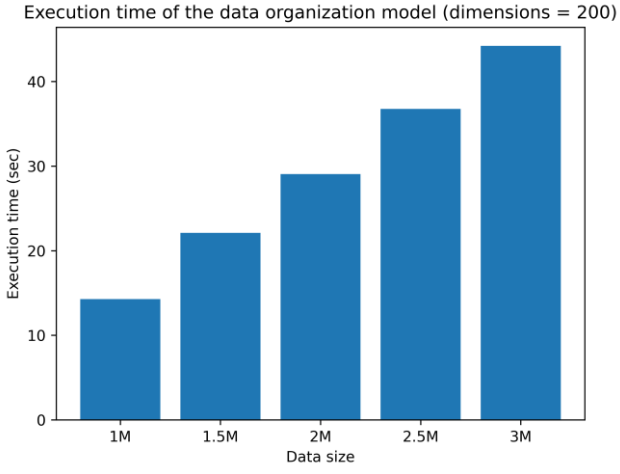


Figure 5. Data size vs execution time of the model for a fixed dimension size (e.g. 200).

We have executed the NNSA on an eight-core intel i7 2.60 GHz, 16 GB main memory PC and received the above performance. It is evident that the NNSA algorithm requires more computing resources for large size multidimensional datasets and can be further improved while incorporating the computing power of the graphical processor unit (e.g. GPU) [20].

Next, we have compared our model with the existing models from the scikit-learn package [21]. We have used the iris dataset [22] to compare between the models that use the NNSA and that don't. We have found that the KNN model that uses the NNSA show better accuracy irrespective to the no. of neighbours than that doesn't (e.g. Fig. 6). As we have already explained in the previous sections that the NNSA provides the searching space of the nearest neighbours for the point of interest enhancing the overall accuracy of the model. The higher accuracy of the KNN with the NNSA requires some little extra time as well (e.g. Fig. 7).

Next, we have tested the above dataset in the K-means algorithms that use the NNSA and don't respectively. We found that the minimum no. of clusters suggested by the K-means with NNSA are eight whereas it is four as suggested by the traditional K-means algorithm (e.g. Fig. 8). The NNSA can determine the minimum no. of clusters by providing the minimum searching space for a point of interest and thus, can help to predict the optimum no. of clusters by forming single clusters following agglomerative, hierarchical approaches [23].

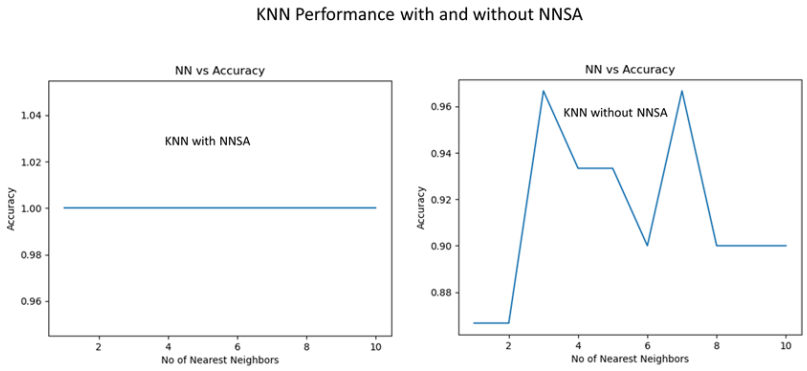


Figure 6. Model performance with and without the NNSA. The KNN with the NNSA shows constant accuracy over the no. of nearest neighbours.

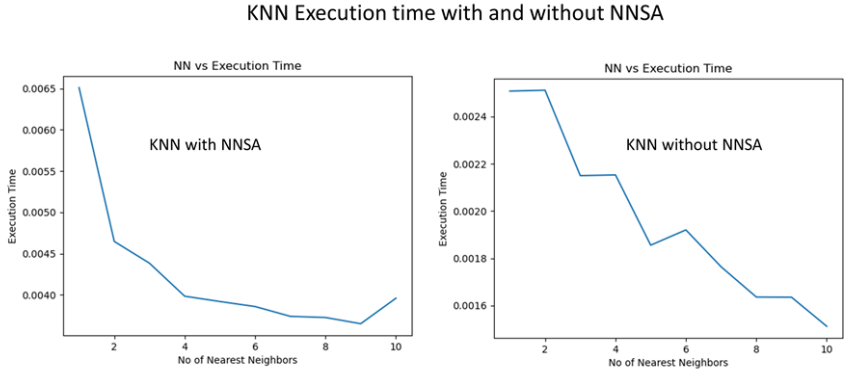


Figure 7. KNN with and without NNSA versus Execution Time (s). The higher accuracy of the KNN with NNSA is responsible for some extra time

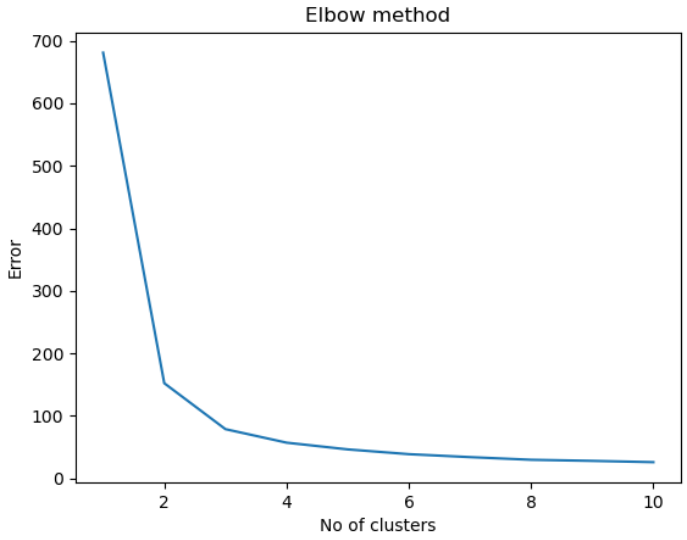


Figure 8. Error vs no. of clusters to determine optimum no. of clusters required for the iris dataset.

5. CONCLUSIONS

We have introduced an index-based data organization model for the multidimensional dataset and explained how this organization can enhance the predictability of both the supervised and unsupervised machine learning algorithms. The analysis of the multidimensional data increases as the no. of dimension increases [24]. We have explained that this complexity of analysis can be reduced by organizing the dataset in a reduced space. Our approach is simple, easy to understand and can be integrated with the existing machine learning algorithms. We have used the Euclidean norm as a metric of the data sorting and next, this has been reused in the calculations involved in the distance-based learning algorithms. Similar way, other metrics that have repeating usages in the learning algorithms can be applicable in this data organization model to further improve the predictability, data accessibility and overall computing efficiency of such algorithms.

ACKNOWLEDGEMENTS

The authors would like to thank the department of computer science at the North American University for supporting the research.

REFERENCES

[1] Pat Langley and Herbert A. Simon, “Applications of Machine Learning and Rule Induction”, Communications of the ACM November 1995/Vol. 38, No. 11, 1995.
[2] MI Jordan, TM Mitchell, “Machine learning: Trends, perspectives and prospects”, Science 2015.
[3] G. J. Grevera and A. Meystel, “Searching in a multidimensional space”, Proceedings. 5th IEEE International Symposium on Intelligent Control Philadelphia, PA, USA, 1990, pp. 700-705 vol.2, 1990.
[4] Norio Catyama and Shin’ichi Satoh, “ The SR-Tree: An Index Structure For High-Dimensional Nearest Neighbor Queries”, *ACM*, 1997.
[5] Antonin Guttman, “R-Trees: A Dynamic Index Structure For Spatial Searching”, *ACM*, 1984.
[6] Jo-Mei Chang, King-Sun Fu, “Extended K-d Tree Database Organization: A Dynamic Multiattribute Clustering Method”, *IEEE Transactions on Software Engineering*, 1981.

[7] EG Sirer, NL Caruso, B Wong, R Escriva, “System and methods for mapping and searching objects in multidimensional space”, US Patent 9,317,536, 2016 - Google Patents
[8] Thomas Seidl, Hans-Peter Kriegel, “Optimal Multi-Step k-Nearest Neighbor Search” , Proceedings of the 1998 ACM SIGMOD.
[9] Peipei Xia, Li Zhang, Fanzhang Li, “Learning similarity with cosine similarity ensemble”, Information Sciences,Volume 307,2015,Pages 39-52, ISSN 0020-0255.
[10] D. Sculley, “Web-Scale K-Means Clustering”, 2010, Proceedings of the 19th international conference on World wide web, 2010, Pages 1177–1178.
[11] V. Garcia, E. Debreuve and M. Barlaud, “Fast k nearest neighbor search using GPU”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, AK, 2008, pp. 1-6.
[12] K Wagstaff, C Cardie, S Rogers, S Schrödl, “Constrained K-means Clustering with Background Knowledge”, Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577–584.
[13] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74.
[14] D. Sculley, “Web-Scale K-Means Clustering”, 2010, Proceedings of the 19th international conference on World wide web, 2010, Pages 1177–1178.
[15] Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, Andy Song, “Efficient agglomerative hierarchical clustering”, Expert Systems with Applications, 2015, Volume 42, Issue 5.
[16] T Xiong, S Wang, A Mayers, E Monga, “DHCC: Divisive hierarchical clustering of categorical data”, Data mining and knowledge discovery, 2011.
[17] Jianxin Wang, Min Li, Jianer Chen, and Yi Pan, “A Fast Hierarchical Clustering Algorithm for Functional Modules Discovery in Protein Interaction Networks”, IEEE/ACM Transactions on computational biology and bioinformatics, 2010.
[18] M. N. Vrahatis, B. Boutsinas, P. Alevizos, G. Pavlides,“The New k-Windows Algorithm for Improving the k-Means Clustering Algorithm”, Mathematical and Computer Modelling, journal of complexity 18, 375–391 (2002).
[19] Mark Lutz, Learning Python, Third Edition, 2008, Published by O’Reilly Media, Inc.
[20] N Govindaraju, J Gray, R Kumar, Dinesh Manocha, “GPUTeraSort: High Performance Graphics Co-processor Sorting for Large Database Management”, Microsoft Technical Report MSR TR-2005-183, 2006.
[21] Scikit-Learn: Machine learning in Python.
[22] R. A. Fisher, Sc.D., F.R.S., “The use of multiple measurements in taxonomic problems”, 1936, Annals of eugenics, Wiley Online Library.
[23] Mark Ming-Tso Chiang, Boris Mirkin, “Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads”, Journal of Classification 27 (2009).
[24] P Indyk, R Motwani,” Approximate nearest neighbors: towards removing the curse of dimensionality”, Proceedings of the thirtieth annual ACM, 1991.

AUTHORS

Mahbubur Rahman completed his masters from the Texas Tech University in 2010 and PhD from the University of Houston in 2015 in computer science. His research interest is in machine learning, data mining, algorithm optimization. He developed a multidimensional searching algorithm during his master's at the Texas Tech University. He also developed and enhanced computational modeling using machine learning algorithm and HPC environment at the University of Houston during his PhD. He developed a clustering-based prediction algorithm during his postdoctoral research at the University of Texas Medical Branch at Galveston (UTMB).

He also developed and managed two scientific websites:
https://cdssim.chem.ttu.edu/
https://dynamic-proteome.utmb.edu

Linkedin profile:
https://www.linkedin.com/in/mahbubur-rahman-26544a27

Github link: https://www.github.com/shawonmr

Google scholar: https://scholar.google.com/citations?user=w7x71U4AAAAJ&hl=en

AUTOMATED ESSAY SCORING SYSTEM USING MULTI-MODEL MACHINE LEARNING

Wilson Zhu[1] and Yu Sun[2]

[1]Diamond Bar High School, Diamond Bar, California, USA
[2]California State Polytechnic University, Pomona, California, USA

ABSTRACT

Standardized testing such as the SAT often requires students to write essays and hires a large number of graders to evaluate these essays which can be time and cost consuming. Using natural language processing tools such as Global Vectors for word representation (GloVe), and various types of neural networks designed for picture classification, we developed an automatic grading system that is more time- and cost-efficient compared to human graders. We applied our application to a set of manually graded essays provided by a previous competition on Kaggle in 2012 on automated essay grading and conducted a qualitative evaluation of the approach. The result shows that the program is able to correctly score most of the essay and give an evaluation close to that of a human grader on the rest. The system proves itself to be effective in evaluating various essay prompts and capable of real-life application such as assisting another grader or even used as a standalone grader.

KEYWORDS

Automated Essay Scoring System, Natural Language Processing, Multi-Model Machine Learning.

1. INTRODUCTION

Automated essay scoring originated with work of Ellis Batten Page. Page suggested the possibility of such a system in 1966 [1] and that such a system can match the performance of human judges.

Page created the Project Essay Grade (PEG) in 1968 [2-5] but technology at the time would not have allowed his system to be cost-effective [6] and he eventually sold his system, PEG, to Measurement Inc. Other systems have been developed such as the Intellimetrics by Vantage Learning which was first used in 1998[7] and the E-rater offered by Educational Testing Service that was first used in 1999 [8].

More recently, the 2012 Kaggle competition, Automated Student Assessment Prize, sponsored by Hewlett Foundation [9] saw numerous teams attempting to develop a program that is capable of scoring an essay to the same ability that a human grader could. The winning team achieved a kappa of 0.81407. There has been very few studies and breakthrough in this field after the Kaggle competition, and these recent studies are mostly based on this Kaggle competition.

Previous automated essay scoring systems such as the e-rater used feature extraction to obtain necessary information from an essay [10-12]. These systems are effective but lack the ability to accurately evaluate the content of the essay as it measures shared vocabulary between the prompt

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 109-117, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101211

and the essay and cannot reach the level of content-comprehension that a word vector model can achieve. However, e-rater does have error analysis and style analysis which word vector models cannot achieve. Despite its advantages, the feature extraction model ultimately falls short in the content evaluation aspect which is arguably the most crucial aspect of the essay.

In this paper, we attempt to combine the approaches of feature extraction model and word vector model. Using feature extractions to obtain word count, grammar mistakes, and part of speech count and implementing word vectors such as GloVe, this method can measure both the numerical features of the essay as well as the contextual feature and its relatedness to the topic. Compared to previous methods where only either feature extraction or word vector is present, our method combines both and takes advantage of both methods while minimizing the shortcomings of only using one. Therefore, we believe our automated essay grading has potential to be less vulnerable to 'study to test' essays.

To evaluate the result of the program, we test our data on the validation set consisting of 20 percent of the overall data and obtain a kappa score through the following function:

$$\kappa = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Where the quadratic weighted kappa is calculated, and the mean is taken across all data. A kappa score of 1 demonstrates total agreement between the raters - machine and human - while 0 demonstrates random agreement and a kappa less than zero indicates agreement is less than that by chance.

The rest of the paper is organized as follows: Section 2 gives detail regarding the challenges faced during data transformation, data tokenization and network evaluation; Section 3 focuses on our methodology and solutions to the challenges offered in Section 2; Section 4 present the result of evaluation of the program on the validation set followed by presenting related work in Section 5; Finally, Section 6 gives conclusion remarks as well as possible future works of this project.

2. CHALLENGES

There are a multitude of challenges that exist in the project. They will be discussed in this section.

2.1. Challenge 1: Tokenizing the data

The data can be tokenized in a variety of ways. The essay can be tokenized through the use of a GloVe embedding layers and word tokenizer with ease. However, to add to the GloVe representation with numerical representation such as word count and grammar mistake count is a challenge. Since using the GloVe embedding layer required the input to be solely the word index of the individual word of the essay and convert the word indexes into a 2-dimensional array that represents the context of the essay. Thus, adding other information proves to be impossible and requires an overhaul of the tokenization method, either find a way to represent this additional numerical information within the 2-dimensional array or find another method to replace the 2-dimensional array.

2.2. Challenge 2: Choosing the neural network

There are numerous ways to create a neural network and choosing the optimal network proves to be a major challenge. There are multiple aspects to consider when choosing the network - the type of neural network, the activation function, and the number of neurons. The type of neural network depends heavily on the application. For example, Recurrent Neural Network is best suited for sequential data while Convolution Neural Network works best on image data. In addition, choosing the right activation can be crucial. Depending on the type of network, Rectified Linear Unit, Sigmoid, or SoftMax are utilized to maximize the accuracy. Furthermore, the number of neurons in each layer may influence the outcome as well. To optimize the neural network, choosing the correct elements can be difficult.

2.3. Challenge 3: Training and Evaluation

When training the model, there are multiple things to consider - epoch, batch size, optimizer, and metrics. It is imperative to find the optimal epoch and batch size combination to minimize the training time since the training can be extremely time-consuming. The training time and accuracy is also affected by the optimizer used. Therefore, the most efficient optimizer for this specific application needs to be selected. The most challenging aspect of training is choosing the correct metric to maximize the evaluation score. Since the model is being rated through quadratic weighted kappa, it is necessary to find or create the closest metrics available to the evaluation function and maximize the score.

3. SOLUTION

The Automated Essay consisted of two major components, the essay tokenizer and the neural network model. The tokenizer converts the essay, a string, into two vectors, one containing the numerical representation and another one containing the word vector representation of the essay. The vector is then passed into the neural network which evaluates the model using the trained model and returns a score.

In detail, the preprocessing process outputs a numerical representation for feature extraction as well as a sequence for the word vector model. The numerical representation consists of the following:

- Grammar Counts: Counts of various parts of speech using spaCy to reflect the writer's ability to utilize a wide aspect of the English language. After evaluating the essay's grammar and spelling mistakes, those mistakes are corrected for further processing.

- Numerical Counts: Total word count, character count, unique word count, average word length, sentence count, paragraph count, and comma count of the corrected essay; Stop words from the spaCy library are removed and total word count, character count, unique word count, and average word length are calculated and the difference between the pre-removal and post-removal are calculated as well.

The input sequence for word vector neural network consists of a sequence of fixed size from Keras text tokenizer. The overview of our solution is represented in the figure below:

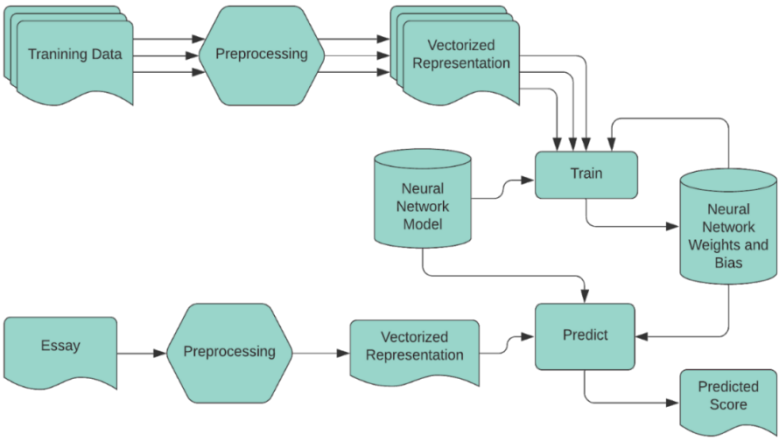


Figure 1: Overview of the Solution

Using Keras Functional API, we created a custom multi-model neural network consisting of a two-layer neural network to processes the numerical representation of the essay and a word vector neural work that processes the sequence from Keras tokenizer. Then the outputs of both neural networks are concatenated and outputted into another two layers neural network to produce the final score. Graphical representation of the multi-model neural network with Long Short-Term Memory (LSTM) as word vector neural network and a GloVe embedding dimension of 50:

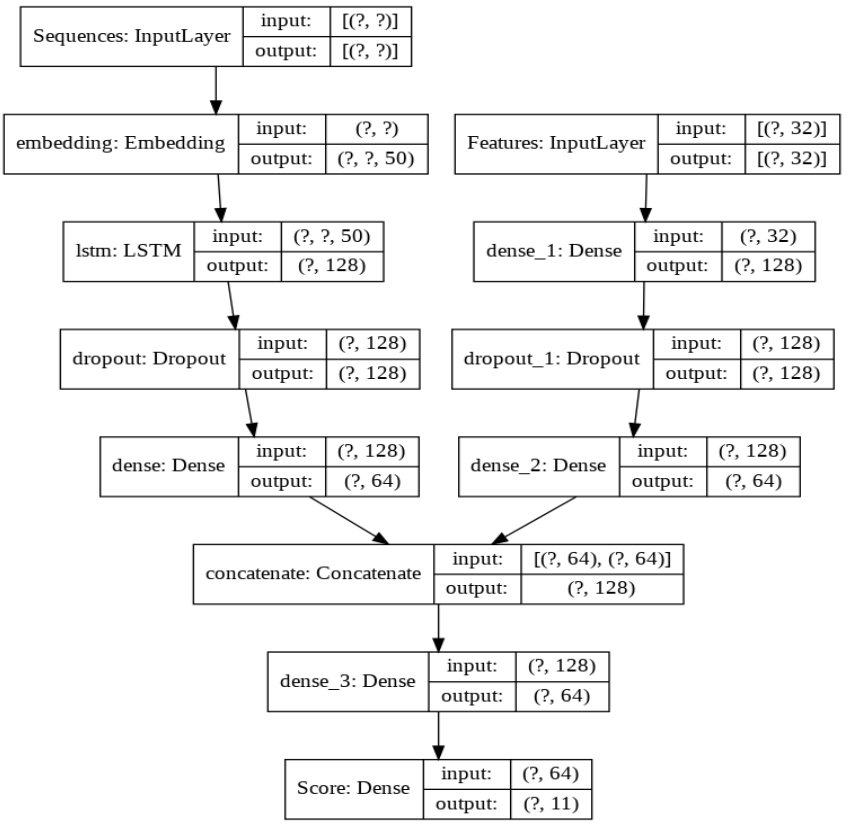


Figure 2: Graphical Representation of the Neural Network with LSTM

We used 'adaptive moment estimation' ('adam') optimizer and in order to minimize the training time we trained each neural network with a batch size of 50 essays, epoch of 40, and a varying learning rate as follows:

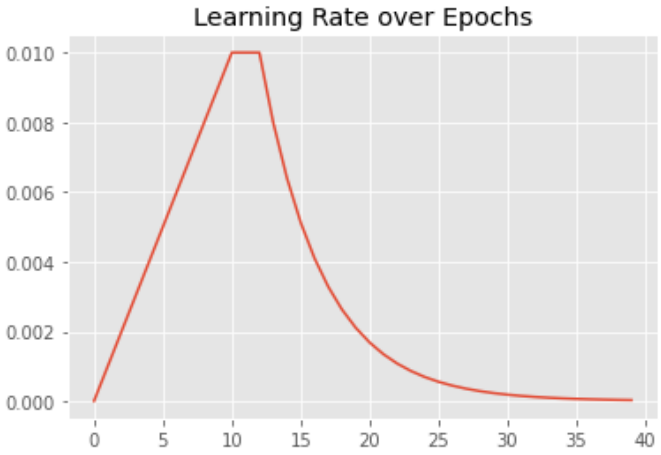


Figure 3: Custom Learning Rate over Epoch using 'adam' optimizer

4. EXPERIMENTS

To maximize performance, we decided to test two parameters – type of neural network used for GloVe embedding, and length of the GloVe embedding dimension. The types of neural network we tested are Long Short-Term Memory, Gated Recurrent Unit (GRU), and Bi-Directional LSTM (BiLSTM). The various length of GloVe embedding dimensions are 50, 100, 200 and 300.

Various neural network has their own advantages and disadvantages. For example, LSTM contains an input gate, an output gate, and a forget gate, whereas GRU contains a reset gate and an update gate. This difference leads to GRU having a shorter training time but LSTM having better performance on longer sequences. BiLSTM is LSTM with a bidirectional layer that reads the given text from beginning to end and end to beginning, which allows the neural network to capture more information since LSTM tend to 'forget' information in the beginning of the text. However, BiLSTM is more computationally expensive. Thus, we needed to test for the best neural network for our system.

Different sizes of GloVe word embedding can impact the neural network performance as a longer embedding dimension would allow the input essay to be expressed more thoroughly whereas a shorter embedding dimension allows the essay to be expressed more concisely. This difference could impact the performance since a larger embedding dimension would allow for too much unnecessary details, but a shorter embedding dimension may lead to crucial information being lost. Therefore, we decided to use the embedding dimension as one of the parameters for our experiment.

4.1. Neural Networks

To test for the most optimal neural network, we took the average kappa score across the various sizes of GloVe embedding dimension of each network. LSTM obtained an average of 0.69479, GRU obtained an average of 0.63776, and BiLSTM obtained an average of 0.67321. Furthermore, the best performing model for LSTM, GRU, and BiLSTM obtained a kappa of 0.70026, 0.68525, and 0.70024, respectively, with LSTM being the best performing neural

network overall. As the figures below demonstrate, LSTM has the highest training accuracy and lowest training categorical cross entropy in one of the eight essay set, with other training essay set having a similar trend.

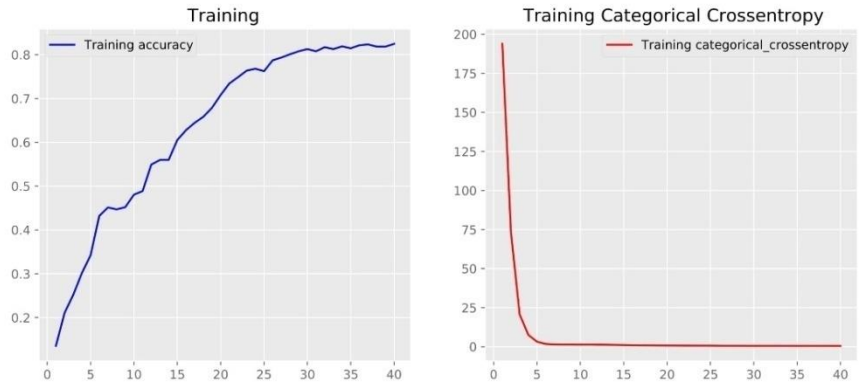


Figure 4: Training based on LSTM Model

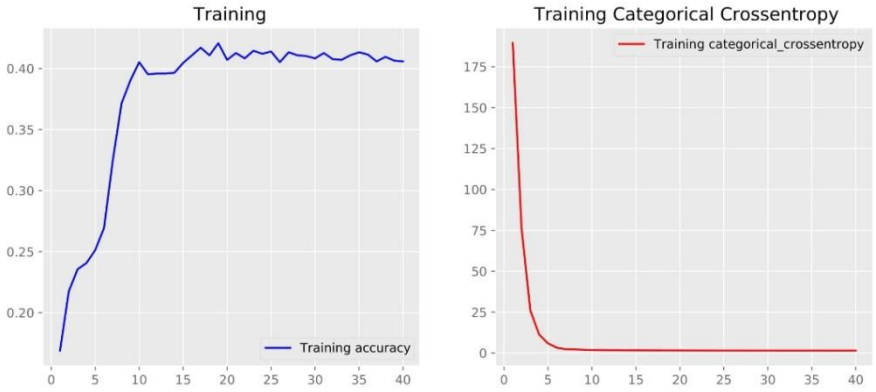


Figure 5: Training based on GRU Model

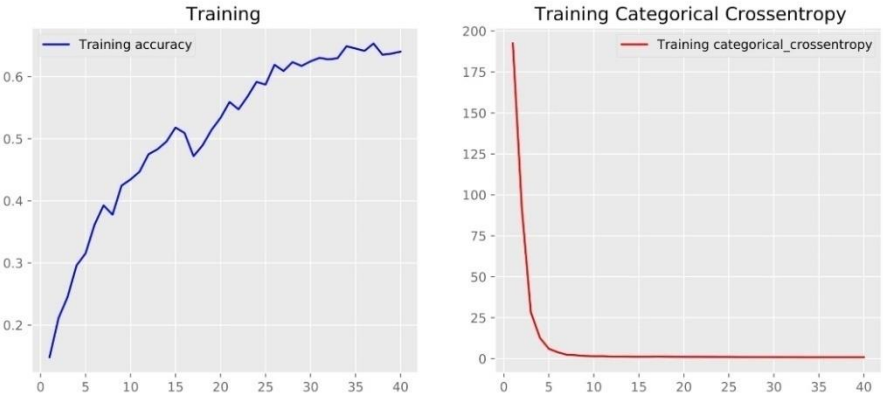


Figure 6: Training based on Bi-Directional LSTM Model

The result shows LSTM being the most optimal neural network. GRU performed worse than the other two because as stated above, it has a faster training time but is not better suited for longer text model than LSTM. BiLSTM on the other hand may have retained more information, but the kappa score evaluates agreements with a human reader. This poor result may be explained by human grader reading the essay from top to bottom. Although BiLSTM is able to retain most information, even the top of the essay since BiLSTM reads the essay in both directions, the human grader tends to remember information later in text as people only read from top to bottom, which is similar to LSTM. Thus, LSTM performed best on this application as it is most similar to how a human grader read and evaluate an essay.

4.2. GloVe Embedding Dimension

To find the most optimal GloVe embedding dimension, we compared the quadratic weighted kappa of each embedding dimension with different neural networks. The table below shows the scores:

Table 1: Quadratic Weighted Kappa of Tested Neural Networks

Dimension\Network	LSTM	GRU	BiLSTM	Average
50	0.70003	0.57328	0.70024	0.65785
100	0.68477	0.68525	0.67184	0.68062
200	0.70026	0.61821	0.68884	0.66910
300	0.69411	0.67478	0.63192	0.66694

From the table, we can see that the embedding dimension does not greatly affect the performance of LSTM model as much. However, the embedding dimension did affect GRU greatly and BiLSTM to a lesser extent. The result shows GRU preferring an embedding dimension of 100 and 300, which indicates that specific categories within the 100 and 300 embedding dimension allows for better understanding for GRU. The trend is clearer for BiLSTM with a strong preference for shorter embedding dimensions. This may be a result of BiLSTM also retaining much of the information and does not need for longer embedding dimensions to express the details.

In conclusion, we used LSTM neural network with 200 embedding dimensions as our final model since it achieved the highest quadratic weighted kappa at 0.70026.

5. RELATED WORK

Other methods have been written regarding the automated essay grading competition on Kaggle. For example, a team from Rice University obtained a kappa score of 0.63 through extraction of features such as word occurrence, word count, Kullback–Leibler divergence and using linear regression [13]. Similarly, another team from Stanford also attempted the problem using machine learning, with feature extraction and linear regression, scored a kappa of 0.72. The second used a different approach as they included features such as bag of words and part of speech count [14].

In contrast, a team from Stanford University used word vectors and a 2 layers neural network to achieve a score of 0.9447875. They utilized the GloVe word vector with various dimensions and types to tokenize the essay combined with various different types of neural networks and managed a high score [15]. However, this team does not evaluate the various features of the essay such as word count, unique word count, grammar mistakes, etc.

The first two teams were less effective in their method as they only accounted for the numerical features of the essay and cannot accurately account for the context due to the lack of word vectors. The third team can thoroughly evaluate the content and achieved a high kappa score as a result.

6. CONCLUSION AND FUTURE WORK

In conclusion, we developed an easy grading system using machine learning and natural language processing that achieve a quadratic weighted kappa of 0.70026. We used both feature extraction and word vectors to represent the essay and convert the essay into its vectorized representation and tokenized sequences. Then a LSTM neural network is utilized to evaluate the tokenized sequences and a 2-layer neural network to evaluate the vector representation. The results are concatenated, and another 2-layer neural network is used to predict the final score. The resulting kappa demonstrates that the current model is capable of real-world application but still has some shortcomings.

For example, the current model performs vastly better on certain prompts than others and performs significantly worse on longer essays compared to shorter ones since the word vector has a vaguer representation of longer essays and cannot accurately extract the context of the longer essays.

We hope to combat the longer essay problem by giving more weights to certain words using Term Frequency - Inverse Document Frequency. This approach allows less significant words, which appears more frequently in longer essays, to have less weight and thus the content of longer essays to become more expressed.

REFERENCES

[1] Page, E. B. (1966). "The imminence of... grading essays by computer". The Phi Delta Kappan. 47 (5): 238–243.
[2] Page, E.B. (1968). "The Use of the Computer in Analyzing Student Essays", International Review of Education, 14(3), 253-263.
[3] Hsieh, Kevin Li-Chun, Chung-Ming Lo, and Chih-Jou Hsiao. "Computer-aided grading of gliomas based on local and global MRI features." Computer methods and programs in biomedicine 139 (2017): 31-38.
[4] Slotnick, Henry B. "Toward a theory of computer essay grading." Journal of Educational Measurement 9, no. 4 (1972): 253-263.
[5] Czaplewski, Andrew J. "Computer-assisted grading rubrics: Automating the process of providing comments and student feedback." Marketing Education Review 19, no. 1 (2009): 29-36.
[6] Nguyen, Tien Dzung, Quyet Hoang Manh, Phuong Bui Minh, Long Nguyen Thanh, and Thang Manh Hoang. "Efficient and reliable camera based multiple-choice test grading system." In The 2011 International Conference on Advanced Technologies for Communications (ATC 2011), pp. 268-271. IEEE, 2011.
[7] "IntelliMetric®: How it Works", Vantage Learning. Retrieved 28 February 2012.
[8] Burstein, Jill (2003). "The E-rater(R) Scoring Engine: Automated Essay Scoring with Natural Language Processing", p. 113. In Shermis, Mark D., and Jill Burstein, eds., Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates, Mahwah, New Jersey, ISBN 0805839739
[9] Hewlett prize" Archived 30 March 2012 at the Wayback Machine. Retrieved 5 March 2012.
[10] Attali, Y. & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. Journal of Technology, Learning, and Assessment, 4(3).
[11] Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. "Enriching automated essay scoring using discourse marking." (2001).

[12] Aluthman, Ebtisam S. "The effect of using automated essay evaluation on ESL undergraduate students' writing skill." International Journal of English Linguistics 6, no. 5 (2016): 54-67.
[13] Lukic, A., & Acuna, V. (n.d.). Automated Essay Scoring, Rice University.
[14] Manvi Mahana, Mishel Johns, and Ashwin Apte. (2012). Automated essay grading using machine learning. Mach. Learn. Session, Stanford University.
[15] Nguyen H. & Dery L. (2018). Neural Network for Automated Essay Grading. Stanford University.

A GENETIC PROGRAMMING BASED HYPER-HEURISTIC FOR PRODUCTION SCHEDULING IN APPAREL INDUSTRY

Cecilia E. Nugraheni[1], Luciana Abednego[1], and Maria Widyarini[2]

[1]Dept. of Computer Science, Parahyangan Catholic University, Bandung, Indonesia
[2]Dept. of Business Adm., Parahyangan Catholic University, Bandung, Indonesia

ABSTRACT

The apparel industry is a type of textile industry. One of scheduling problems found in the apparel industry production can be classified as Flow Shop Scheduling Problems (FSSP). GPHH for FSSP is a genetic programming based hyper-heuristic techniques to solve FSSP[1]. The algorithm basically aims to generate new heuristics from two basic (low-level) heuristics, namely Palmer Algorithm and Gupta Algorithm. This paper describes the implementation of the GPHH algorithm and the results of experiments conducted to determine the performance of the proposed algorithm. The experimental results show that the proposed algorithm is promising, has better performance than Palmer Algorithm and Gupta Algorithm.

KEYWORDS

Hyper-heuristic, Genetic Programming, Palmer Algorithm, Gupta Algorithm, Flow Shop Scheduling Problem, Apparel Industry

1. INTRODUCTION

Generally, the textile industry can be divided into two major segments, namely the textile industry and the apparel industry [2,3]. The production of fabric from raw material is done through a series of processes such as spinning, weaving, knitting, etc. These processes are the focus of the textile industry. Meanwhile, the apparel industry transforms fabrics into ready-to-use goods, in particular ready-to-wear clothes. The activities that belong to the apparel industry are pattern making, cutting, sewing, and finishing.

Based on the flow of work activities, the production system of the apparel industry is usually included in the flow shop production or job shop production group. Job Shop is a type of production process flow that is used for producing goods with small production quantities but have many models or variants. "Custom-made" products that have to follow the unique design and special specifications of the customer at a specified time and cost usually use this type of production process flow. Flow Shop Production is a type of production process that is used to produce products that are assembled or produced in large quantities and successively (continuous). All products are manufactured with the same standards and processes.

This work focuses on Flow Shop Scheduling Problems (FSSP) which is a class of scheduling problems found in the manufacturing industry whose workflow follows the Flow Shop Production. Given a number of jobs that must be processed in a series of stages, the goal of FSSP

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 119-128, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101212

is to find a sequence of jobs that meets certain optimal criteria. There are many approaches proposed for these problems that are based on heuristic techniques. There are three types of heuristics. The first type is called low-level heuristics such as standard dispatching rules (First Comes First Served, Shortest Processing Time, etc.) and several algorithms such as NEH, Palmer, Gupta, Dannenbring, Pour, etc.[4,5,6], The second type is meta-heuristics such as Genetic Algorithm, Simulated Annealing, Ant Colony Algorithm, etc. [7,8]. The last type is hyper-heuristics such as Genetic Programming [9,10,11]. Different from the other heuristics, hyper-heuristic does not work directly on the problem domains, rather on the heuristics. This is why hyper-heuristic is usually called heuristics to choose heuristics. This characteristic enables hyper-heuristic to do the searching in a more flexible way. Also, hyper-heuristics offers the ease of application to a larger scope of problems.

In our previous work [1], we proposed a technique for solving FSSP problems which is a hyper-heuristic based genetic programming. The main idea of the technique is to generate new low-level heuristics by combining two low-level heuristics namely Palmer Algorithm and Gupta Algorithm with the use of the Genetic Programming technique. The combination of two algorithms has been not reported before. Continuing the work, we have developed a program that implements the proposed technique. This paper reports the implementation as well as the results of experiments conducted for measuring its performance.

The rest of this paper is organized as follows. Section 2 briefly describes the GPHH algorithm that we proposed to solve FSSP problems. We briefly describe what FSSP is, the types of heuristics used to solve the FSSP and the algorithm we proposed. A more complete explanation of this can be found in [1]. Section 3 explains the implementation of GPHH. Section 4 presents the experimental results. Last, conclusion and future work are given in Section 5.

2. PROPOSED TECHNIQUE

Given m machines and n jobs to be processed on each machine, the objective of the FSSP is to find job sequences that meet certain optimality objectives. In this study, the objective used is makespan, which is the time needed to process the entire job starting from the first job in the first machine to the final time of processing the last job on the last machine. It is assumed that at each stage there is only one machine provided for processing the jobs.

Palmer's algorithm and Gupta's algorithm have very similar working principles. In order to generate job orders, both algorithms use a value that is known as the slope index. The two algorithms consist of two stages, namely calculating the slope index of each job and sort the jobs according to the slope indices.

Genetic programming uses the same idea as the genetic algorithm. It is inspired by Darwin's theory of evolution. The difference between genetic programming and genetic algorithm lies in the role of the chromosomes. In genetic programming, a chromosome represents a computer program (function) that can be used to find solutions to problems. The application of genetic operators (crossover, mutation, and reproduction) generates new computer programs. Figure 1 gives two examples of computer programs represented as syntax trees (p-2 and p/1+g), as well as the results of the crossover (p/1-2 and p+g). For mutation operation, a randomly generated sub-tree will replace the part of original tree based on a mutation point. Figure 2 shows an example of mutation operation over computer program. The replacement of constant 2 by p+g resulting in a new computer program which is (p/1 – (p + g)).

The algorithm of GPHH is given in Figure 3. Very similar to genetic algorithms, the algorithm starts with an initial population consisting of some individuals representing a set of computer

programs. Then, iteratively, a new population is generated by applying some genetic operations. Given a set of inputs namely a set of operands (constants and variables), a set of operators, population size, pool size, maximum depth of the syntax tree, the number of the run, the number of generation, FSSP problem, crossover rate, and mutation rate, this algorithm returns the schedule or the order of the jobs and the makespan corresponding to it.

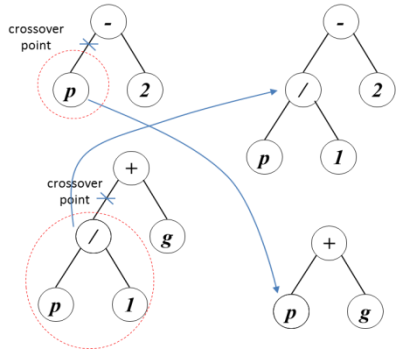


Figure 1. An example of crossover operation.

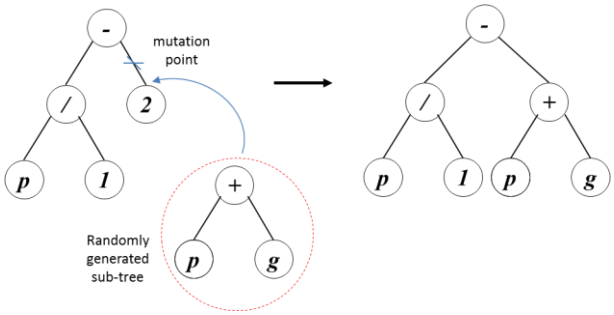


Figure 2. An example of mutation operation.

function GPHH (*ops*, *oprs*, *popSize*, *poolSize*, *depth*, *maxRun*, *maxGen*, *problem*, *COrate*, *Mrate*) → (schedule, makespan)

1. Generate a pool of computer programs based on *ops* and *oprs* which are a set of operands and a set of operators, respectively. Two parameters are used in this stage, namely *depth* representing the depth of programs' syntax trees and *poolSize* representing the size of the pool or the number of generated programs.

2. **for** i = 1 **to** *maxRun* **do**

a. Create the initial population with the size of the population is *popSize* by randomly picking computer programs in the program pool.

b. **for** j = 1 **to** *maxGen* **do**

i. Generate a new population by applying genetic operations over computer programs in the current population and using the COrate and Mrate as the rate of cross over and mutation operation, respectively.

ii. Apply fitness measure to individuals in the new population.

iii. Record the best individual so far as well as the schedule and the makespan corresponding to it.

iv. j = j +1

c. i = i + 1

3. **return** the schedule and the makespan of best individual.

Figure 3. GPHH Algorithm.

3. IMPLEMENTATION

We have developed a computer program implementing the algorithm of GPHH as described above. In the implementation we assumed that the set of the operands and the operators, ops and

oprs, are fixed, which are {p, g, 1, 2} and {+, -, *, /}, respectively. The symbol p is used to represent the slope index of Palmer Algorithm and the symbol g is used to represent the slope index of Gupta Algorithm.

The user interface of the programs is given in Figure 4. The InputFile button is used for choosing a file representing the problem to be solved. There are two tabs, namely Problem Tab and Solution Tab. The Problems tab displays the problem in a tabular form which describes the time required for each machine to process each job. The Solution Tab is used for displaying the schedule and makespan resulted by Palmer Algorithm, Gupta Algorithm, and GPHH as shown in Figure 4.

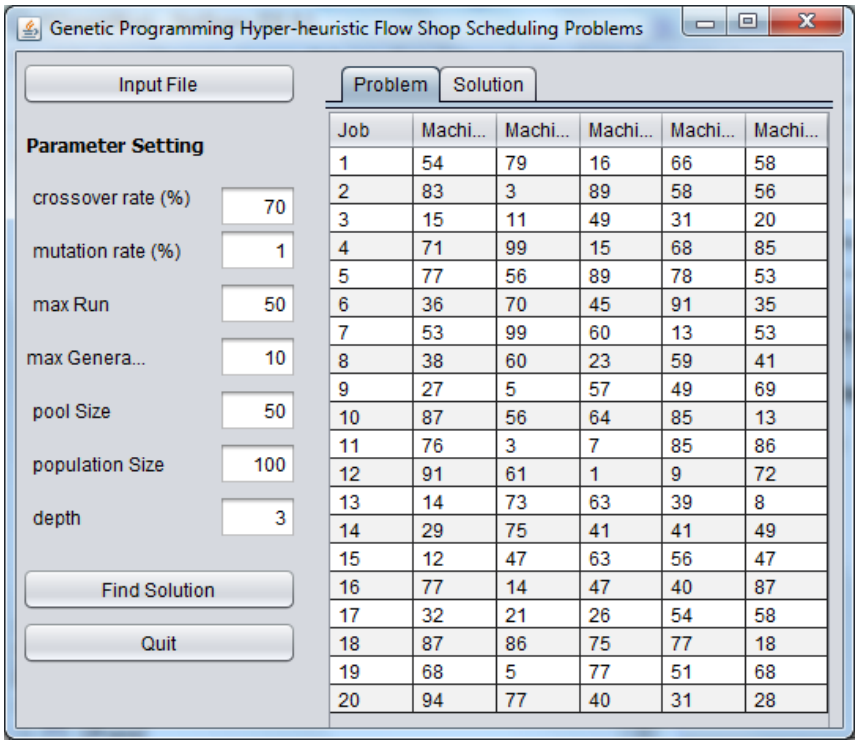


Figure 4. The problem tab for displaying the problem.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In order to measure the performance of the heuristics generated by GPHH, a number of experiments were carried out. In carrying out the experiments, a benchmark proposed by Taillard et al. [12]. This benchmark provides a number of scheduling problems grouped by the number of jobs and machines as shown in Table 1. Each group consists of 10 problem instances. So in total there are 120 problem instances.

Experiments were carried out with different crossover rate COrate, namely 75%, 80%, and 85%. The rate of the mutation operation Mrate is set 5%. For each COrate we have used three depth values of syntax tree which are 2, 3, and 4. Meanwhile, the values for the other parameters are fixed, namely: maxRun = 50, maxGen = 10, poolSize = 250, and popSize = 200. The objective of the experiments is to compare the makespan generated by the GPHH algorithm with the makespan from Taillard's benchmarks, the makespans produced by the Palmer Algorithm and the makespans produced by Gupta Algorithm. Besides, the experiments also aim to know the effect of the crossover rate and the depth of syntax tree on the resulting makespan.

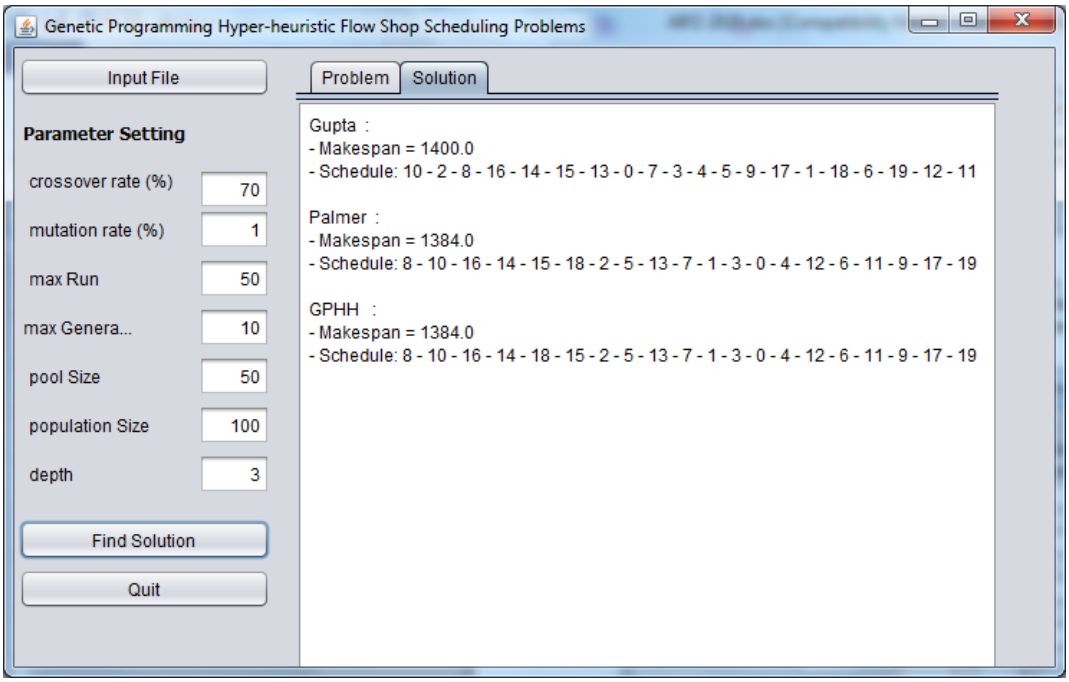


Figure 5. The solution tab for displaying the schedule and makespan of each algorithm.

Table 1. Problem group of Taillard Benchmark.

Group	The number of the jobs	The number of the machine
1	20	5
2	20	10
3	20	20
4	50	5
5	50	10
6	50	20
7	100	5
8	100	10
9	100	20
10	200	10
11	200	20
12	500	20

Figures 6, Figure 7, and Figure 8 compare the average makespans resulted from GPHH Algorithm with the corresponding makespans produced by other algorithms for each COrate value. For each syntax depth tree, depth, Figure 9 and Figure 10 show the comparison of average makespans of each instance problem group and the total average makespan.

From the experimental results, it can be concluded that although the performance of GPHH is still below the benchmark, in general, GPHH produces better makespan compared to Palmer Algorithm and Gupta Algorithm. The worst performance of GPHH occurs in the case of 500 jobs 20 machines. It can be seen that for the three experiments (Figure 6, 7, and 8), GPHH takes the last position. It can be observed, for the case of 500 job 20 machines, the benchmarks are also worse than Palmer and Gupta. Whether cases with many very large jobs will reduce the performance of GPHH still needs to be further analyzed and tested.

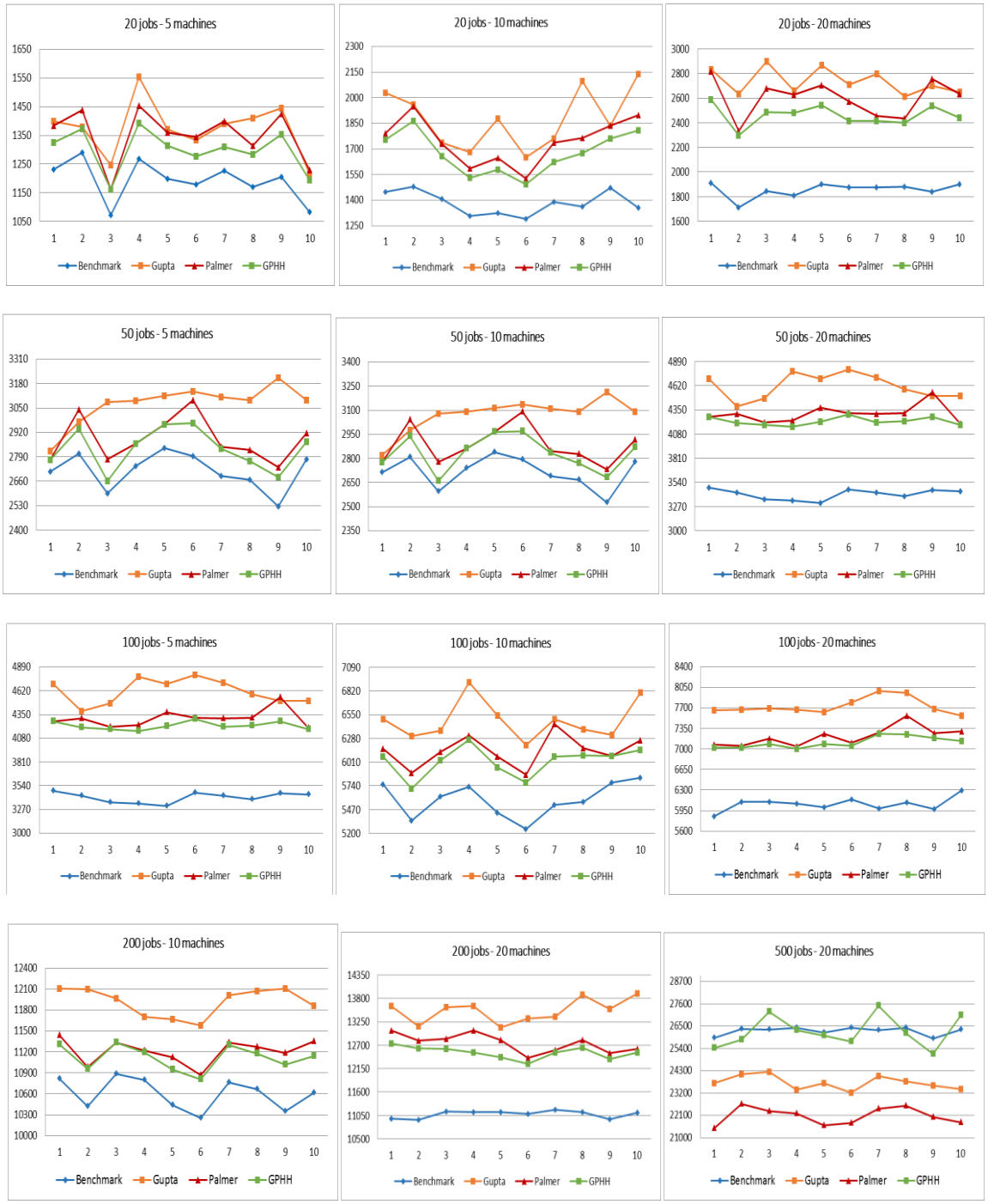


Figure 6. Experimental Results for Cross Over Rate 0.75.

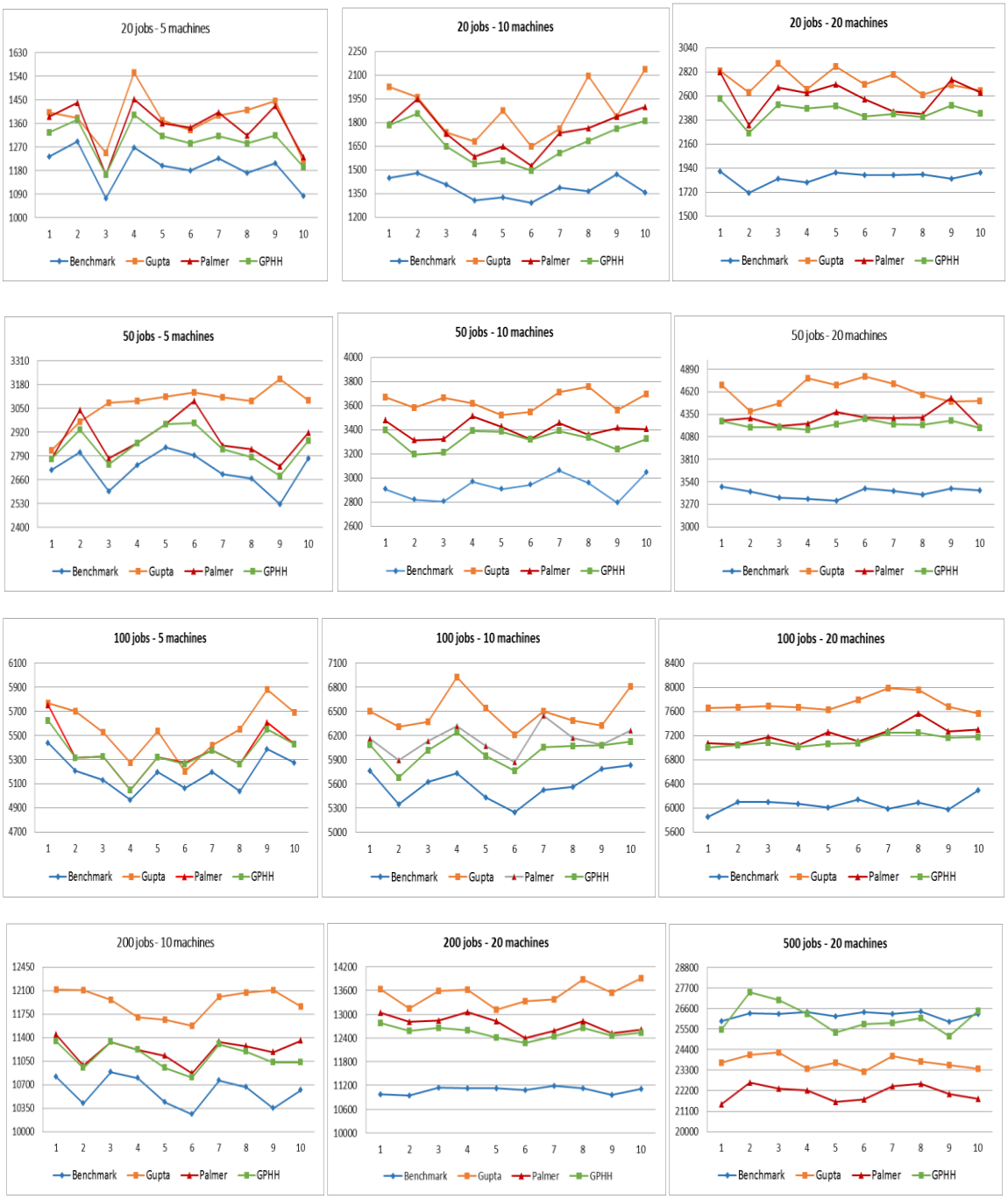


Figure 7. Experimental Results for Cross Over Rate 0.80.

Figure 8. Experimental Results for Cross Over Rate 0.85.

In terms of the depth of the syntax tree, from Figure 9 it can be seen that the three tree depth values show results that are not too different from one another for each problem instance group. However, from the total average makespan, the value 3 yields the best result compared to values 2 and 4 as shown in Figure 10.

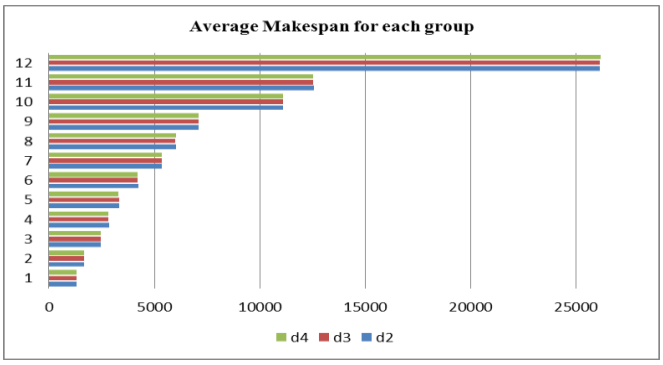


Figure 9. Average Makespans for each problem instance group.

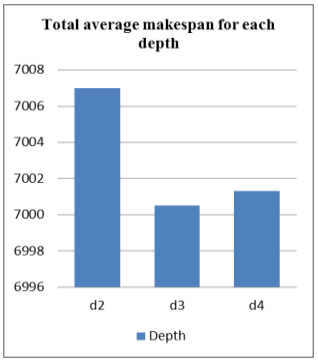


Figure 10. Total average for each syntax tree depth.

5. CONCLUSIONS

Scheduling problems in the textile industry in general belong to the Flow Shop Scheduling Problem (FSSP). Given a set of machines and a set of jobs, the objective of FSSP is to find a job order that meets some optimization criteria. In this work, a computer program that can be used to solve FSSP has been developed. This program implements the GPHH Algorithm which is genetic programming based hyper-heuristic for solving FSSP proposed in [1]. The experimental results show that the GPHH algorithm is promising. Even though it has not outperformed the benchmarks that are used as reference, this algorithm has better performance than the two basic heuristics, namely Palmer Algorithm and Gupta Algorithm.

Currently, we are working on the application of genetic programming hyper-heuristics for multi-objective FSSP. Not only makespan, but we also consider the lateness (tardiness and earliness) of completing the total jobs.

ACKNOWLEDGEMENTS

This work was supported by Indonesian Ministry of Research, Technology and Higher Education (RistekDikti) under research scheme Penelitian Terapan Unggulan Perguruan Tinggi year 2019-2021 Contract Nr. III/LPPM/2020-04/105-PE-S.

REFERENCES

[1] Cecilia E. Nugraheni and Luciana Abednego. On the Development of Hyper Heuristics Based Framework for Scheduling Problems in Textile Industry. International Journal of Modeling and Optimization, Vol. 6, No. 5, October 2016.

[2] Robert, N. Tomastik, Peter, B. Luh, and Guandong, Liu. Scheduling Flexible Manufacturing System for Apparel Production. IEEE Transaction on Robotics and Automation. 12(5): 789-799.

[3] Scholz-Retter Bernd et al. 2015. Applying Autonomous Control in Apparel Manufacturing. Proc. Of 9th WSEAS Int. Conference on Robotics, Control and Manufacturing Technology. 73-78.

[4] C. E. Nugraheni and L. Abednego, "A survey on heuristics for scheduling problem in textile industry," in Proc. ICEAI 2015.

[5] C. E. Nugraheni and L. Abednego, "A comparison of heuristics for scheduling problems in textile industry," Jurnal Teknologi, vol. 78, no. 6-6. 2016.

[6] Said Aqil and Karam Allali. Three metaheuristics for solving the flow shop problem with permutation and sequence dependent setup time. Proc. Of Conference: 2018 4th International Conference on Optimization and Applications (ICOA). 2019.

[7] Peter Bamidele Shola and Asaju La'aro Bolaji. A metaheuristic for solving flowshop problem. International Journal of Advanced Computer Research, Vol 8(37).

[8] Le Zhang and Jinnan Wu. A PSO-Based Hybrid Metaheuristic for Permutation Flowshop Scheduling Problems. The Scientific World Journal. Vol. 2014.

[9] Ochoa G., Rodriguez J.A.V, Petrovic S., and Burke E. K. 2009. Dispatching Rules for Production Scheduling: a Hyper-heuristic Landscape Analysis. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2009), Montreal, Norway.

[10] C. E. Nugraheni and L. Abednego, "Collaboration of multi-agent and hyper-heuristics systems for production scheduling problem,"International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 7, no. 8, pp. 1136-1141, 2013.

[11] C. E. Nugraheni and L. Abednego, "A combined meta-heuristic with hyper-heuristic approach to single machine production scheduling," International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 8, no. 8, pp. 1322-1326, 2014.

[12] E. Taillard. Some efficient heuristic methods for the flow shop sequencing problem. European Journal of Operational Research 47 (1990) pp. 65-74.

AUTHORS

Cecilia E. Nugraheni received her bachelor degree (1993) and master degree (1995) from Dept. of Informatic Engineering, Bandung Institute of Technology (ITB), Bandung, Indonesia. She has received PhD Degree (2004) from Dept. of Informatics, Ludwig Maximilians Universität, Munich, Germany. Her research interest includes formal methods, intelligent systems, machine learning, meta-heuristic and hyper-heuristic techniques.

Luciana Abednego received her bachelor degree from Dept. of Informatics, Parahyangan Catholic University, Bandung, Indonesia. She has done her Master in Informatics from Bandung Institut of Technology, Bandung, Indonesia. Currenty, she is working as a lecturer at the Dept. of Informatics, Parahyangan Catholic University. Her research interest includes machine learning and intelligent systems.

Maria Widyarini received her PhD from School of Business Management, Bandung Institute of Technology (ITB). She was involved in SME particularly as trainer and researcher in microfinancing issues. Her specializing is in Microfinance for SME. She was Director of Center of Excellence – Small Medium Enterprise and Development (COE SMED) and Head of BA and MBA Dept in Parahyangan Catholic University (2018 - up to now).

SMARTTANK: AN INTERNET-OF-THINGS (IOT) APPLICATION TO AUTOMATE THE WATER TANK REFILLING USING COMPUTER VISION AND AI

Henry Hamilton[1], Yu Sun[2], Fangyan Zhang[3]

[1]Capistrano Valley High School, CA 91765, USA
[2]California State Polytechnic University, Pomona, CA, 91768, USA
[3]ASML, San Jose, CA, 95131, USA

ABSTRACT

This system provides a method of automatically keeping water bowls full and refilling every time it is detected that they are not. This is highly useful for anyone who owns a pet, as it decreases the amount of work the owner will need to do. The system uses an AI model, trained with over a thousand images of water bowls. This allows it to accurately determine when a bowl needs filling. When an empty bowl is spotted, a subsystem consisting of a valve and other electronic parts releases stored water into the bowl. Through experimentation it has been shown the accuracy of the system is about 97% under optimal lighting conditions. Without a light source, the system does not function. Currently, the components are not of the highest quality and the system only works with the bowl used in testing. There are future plans to train the model with new pictures featuring an assortment of bowls. Additionally, an LED could be added to the system to solve the issue of it not working without external light.

KEYWORDS

Artificial Intelligence, image detection, RPI system processor

1. INTRODUCTION

Pet owners face a variety of difficulties on a daily basis, which is expected, as they are responsible for maintaining the life of another organism. For the most part, these difficulties are specific to what type of animal you own, although there are a few that are universal. One of the more problematic of these is providing your pet drinking water while you are not physically present. Whether it be a vacation or business trip, for a week or for a month, your pet still needs fresh water to drink. Some animals, namely rodents, have specialized bottles which do not require constant attention, but most use water bowls. Keeping these bowls full may be a matter of life or death. Even if the situation is not so grave, needing to regularly complete this task is tedious and unenjoyable.

There are a few systems currently available that can accomplish a similar purpose, though all have limitations [1-5]. As mentioned earlier, an existing method is a rodent water bottle [6]. This involves a bottle held upside down with a short metal tube leading towards the rodent. The tube's width decreases as it moves downward. Normally, the end of the tube is sealed with a small rubber ball, held in place by gravity. When the rodent is thirsty, it can use its snout to push up the rubber ball into the wider section of the tube, leading to the flow of water around it as the seal is

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 129-136, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101213

broken. While simplistic yet efficient, the problem lies in the fact that it cannot be used by other animals. It dispenses water at an extremely slow rate, making it useless for larger animals. Additionally, many animals cannot push up the ball, or even comprehend how the system functions. The other solution in existence is a larger water bowl with a button that when pressed, pumps out water from a tank nearby [7]. This solution is limited by the strength required to push the button. Only larger animals, like dogs or cats can utilize this tool. Another assumption made is that the animal in question can even equate an action like pushing a button with being able to drink water. For most reptiles, this is nearly impossible.

My tool involves a raised platform supported by legs and a base. A water bowl can be placed under the platform. The base serves two purposes: firstly, it provides stability to the platform and prevents it from falling over. The second purpose is that it provides a clearly defined location to place the bowl, so that it is within the view of the camera. The platform uses a camera to firstly identify the bowl, then decide whether the bowl is empty or full of water. If the bowl is empty, a valve residing on the platform will open, and water from a nearby tank will flow through it. The water is then channeled into the bowl. The process occurs every 30 minutes, thus autonomously filling the bowl and satisfying whatever animal depends on it without human intervention. An additional feature is that it shuts off at night and turns back on in the morning. This provides a multitude of advantages. Firstly, the tool doesn't work under low light conditions, so being active at the specified times brings no advantages. Additionally, without sunlight, the water will not evaporate at a noticeable rate, so refilling the bowl isn't necessary at night. Finally, by shutting it off energy can be saved.

Two experiments were done to prove the effectiveness of the solution. In the first experiment, the AI was subjected to 60 pictures, broken down into the 3 categories "Filled" "Partially Filled" and "Empty". It sorted the images and accuracy was recorded. These three categories are important, as the AI will most likely encounter all 3 during usage. The second experiment tested accuracy in different light levels with a constant water level. The bowl remained completely filled while the brightness of the room's light was adjusted. Once again, 60 pictures were tested this time under 3 light levels. Results showed that darkness and empty bowls caused the most confusion, but accuracy was still acceptably high in all cases.

The rest of the paper is ordered in this fashion: Section 2 describes a number of obstacles met when attempting to design a solution to the problem; Section 3 provides a detailed overview on how the solution works and what components are in play, along with how the challenges noted in Section 2 were overcome; Section 4 provides information regarding experiments done to prove the solution's effectiveness, while Section 5 presents some related works. Section 6 concludes the paper and predicts the project's future. It also provides a few examples regarding how to solve issues listed earlier in the paper.

2. CHALLENGES

Throughout the system development, we ran into several challenges that needed overcoming. Here is a brief overview of some of the most difficult challenges that we faced when developing this application.

2.1. Challenge 1: Differentiate between Filled and Empty Bowls

When the light levels are too low, the AI has trouble differentiating between filled and empty bowls. This was due to a variety of causes. Initially, when training the AI, all pictures used were taken during midday in a brightly illuminated room. The model was therefore unable to recognize

objects in a different environment. An attempt was made to remedy this by taking pictures during the morning and using those for training. There was very limited success to this though, since the camera equipment was not of the highest quality and therefore look blurrier photos in the absence of light. The final aspect to this issue was that filled and empty bowls looked very similar from above. Even when a bowl is empty, the small amount of water lining the bottom reflects light just like a full bowl would. Since the camera is directly above the bowl looking down into it, the two were hard to differentiate. Darkness only made this harder.

2.2. Challenge 2: Proper Camera Position

Positioning the camera correctly was a challenge. In the initial design, there was no raised platform and the camera was haphazardly taped to a table. This was obviously an issue, since the camera was not secure and instead wobbled around. Additionally, the height of the table meant that the bowl would be far away from the camera, leaving an undesirable amount of space around the bowl. These issues were fixed when the platform was implemented. Now the camera was fixed to a specific spot and could no longer move. The platform was not too high, so the bowl was more clearly visible. Finally, the base helped keep the bowl position consistent, so it would always remain the pictures center. The only remaining challenge would be to remove the extraneous space around the bowl. This was done by zooming in slightly.

2.3. Challenge 3: Water Control System

While the AI posed the most difficulty, the water control system was no easy feat either. The setup was simple initially, involving a tube, a tank and a valve. Unfortunately, one important aspect was overlooked: water doesn't just flow on its own. There was no force to propel the water out of the tank and through the tubing, so even when the valve was open water would not flow. A solution was devised where a tire pump would be used to pressurize the water, and it worked somewhat, but with limited success. The other more obvious solution would be to use gravity. This led to more consistent results, but the water pressure was still unfavorably low.

3. SOLUTION

My project is a system whose purpose is to refill a bowl of water once it is detected that the bowl is no longer full. It is made up of multiple components that work in conjunction towards this goal. The first is the software. The system uses an AI model trained on a vast amount of bowl pictures with varying water levels. When an image, captured with an RPI camera, is sent to the AI, it returns whether the bowl is considered full or empty, along with a confidence level. Both of these values are key in determining the next action. If the AI's confidence exceeds 0.65, the result returned will be used. However, if the AI's confidence is below the previously specified number, the system considers the bowl's state to be the one not returned by the AI, e.i., if the AI returns "Full" the rest of the system will run as if the bowl was empty. The rest of the software side of the system contains some code which activates a pin on the RPI if the bowl is believed to be empty. A series of hardware components open a valve once a signal from the pin is received, which eventually results in water pouring into the bowl from a nearby storage tank. A general list of components involved in the completion of goals are:

- A trained AI model to determine water level in bowls
- Additional software to take pictures and take actions in accordance with the AI's decision
- Multiple pieces of hardware that control the valve.
- Water tank in the form of a milk jug.
- RPI camera to capture bowl photos.
- A wood and plastic stand which acts as a base for all the components.

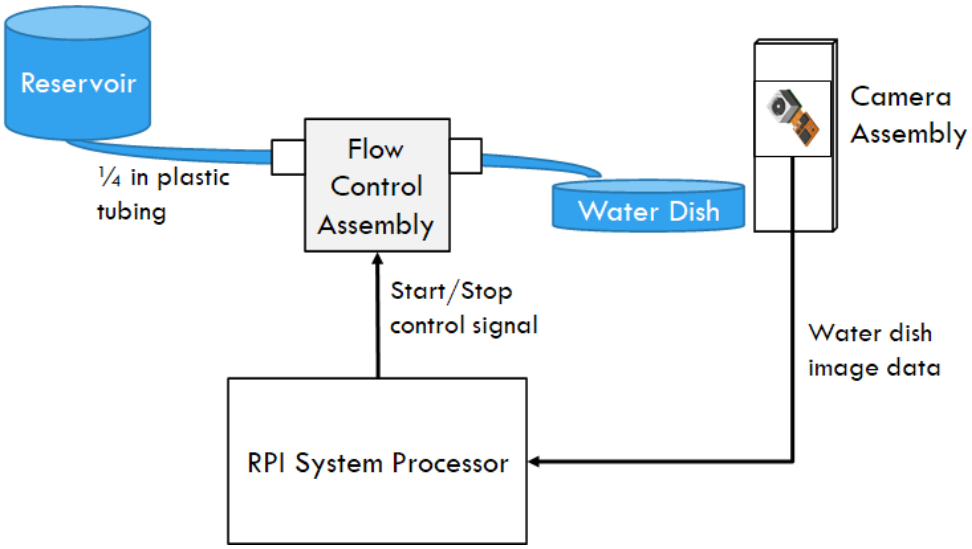


Figure 1: Overview of the components

The first component to be noted is the trained AI model. This was created using Google Cloud's Vision API and was trained with over 1500 images. These images were an assortment of filled, partially filled and empty bowls with varying light levels. To collect these photos, a short program was written that would continuously take pictures of a bowl throughout an entire day. As the day progressed, the movement of the sun would slightly change the light level of each picture in order to make the pictures unique. This was done over multiple days, each day using a different water level. The completed model was then downloaded onto an RPI. An RPI camera attachment was purchased online and also added to the RPI. The code works rather simply. First, a picture is captured. Next, the AI determines whether the bowl is filled or empty and the confidence level. Finally, a function is used to activate a pin on the RPI if the bowl is empty. The next component to be addressed is the valve control system. The electrical components for it are pictured below. The pins on the RPI are linked with a MOSFET switch.

Before further information on the switch, it must be noted that the solenoid valve requires 15V to function correctly. More specifically, it remains closed as long as it receives a constant 15V. The RPI cannot supply 15V, as its maximum is 5V, so the power supply is used to that end. When the RPI sends 5V to the switch, it breaks the connection between power supply and valve, resulting in the valve opening to allow water flow. The water is stored in a milk jug at higher elevation than the rest of the system. This is important, because otherwise the water does not have enough pressure to flow. The water is transported through plastic tubing, eventually emptying into the water bowl. The above-mentioned steps occur multiple times per day at 30 second intervals. The system is deactivated at night since the camera does not operate without light. It should be noted that for each cycle, only a small amount of water is released. To fill an entire bowl the system must be left on for close to an hour.

A four-legged platform is the body of the project, and on it sits all the electrical components. It is made of a flat board filled with holes, 4 PVC pipe legs and 4 white wooden 1 x 2 in planks that surround the bottom of the legs, forming a base. The top wooden platform provides a surface for the RPI and other components, but its other purpose is to elevate the camera to the optimal height. In order for the AI to best make decisions regarding the bowl, it must take a picture with only the bowl within view. The camera is attached to the bottom on the wooden platform, and points straight down at the bowl. At a height of 18 in, the camera perfectly captures the bowl with

few spaces around it. The wood planks around the legs provide stability and also wall off a square area directly under the platform for the bowl.

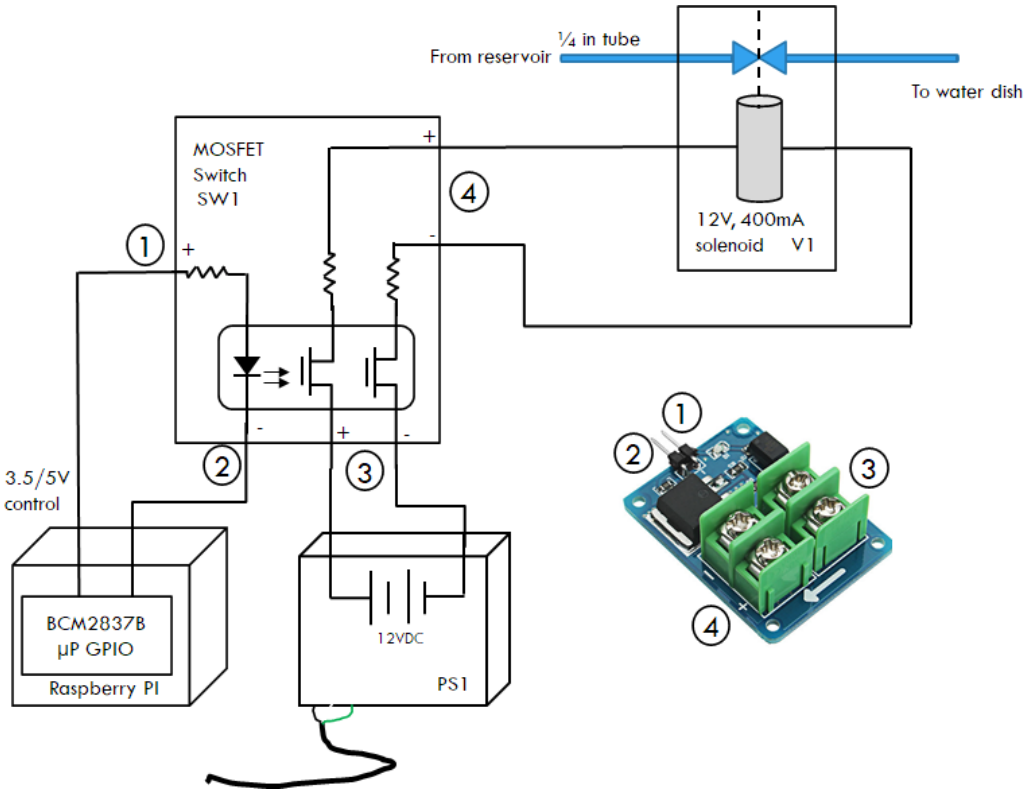


Figure 2: Details of the components

4. EXPERIMENT

The first experiment's goal is to test the AI model's accuracy under 3 different conditions (Table 1). The three conditions are "Completely Full", "Half Full" and "Empty". For each condition, 30 trials were done in order to obtain statistically significant results. All other factors were held constant. This testing was done in a brightly lit room during the day, so the bowl was exposed to both natural and artificial light. The bowl was positioned directly under the camera to avoid interference from other factors. On the results table, T shows that the AI's deduction was correct, or true, while F shows it was incorrect, or false.

Table 1: AI model's accuracy under 3 different conditions

Condition	Results	Accuracy
Completely Full	T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T	100%
Half Full	T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, F, T, T, T, F, T, T, T, T, T	93%
Empty	F, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, F, T	93%

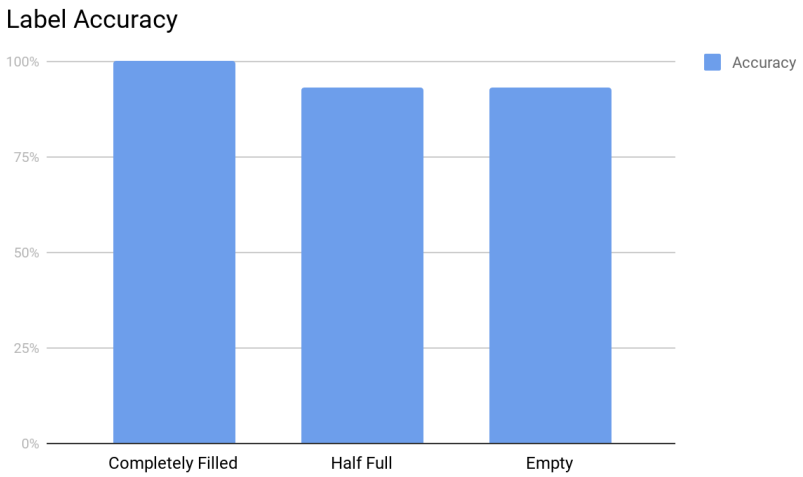


Figure 3: Label Accuracy in bright room

Accuracy appears to be higher for full bowls than empty bowls. This is further made clear by the confidence intervals for each result. On average, the AI had a confidence of about 0.7 for full bowls, only 0.6 for empty bowls and half full bowls.

This experiment's goal is to test the AI's accuracy under multiple light levels (Table 2). There are three levels tested in the experiment, one with light in the room completely off, one with the light at half brightness and one with the light at full brightness. The lights in question are LED lamps and the curtains are closed, making sunlight negligible. Water level was held at a constant level of completely filled. 30 trials were done for each light level. T shows that the deduction was correct, while F shows it was incorrect.

Table 2: AI's accuracy under multiple light levels

Condition	Results	Accuracy
Light at Full Strength	T, T, T, T, T, T, T, T, T, T, T, T, T, F, T, T, T, T, T, T, F, T, T, F, T, T, F, T	84%
Light at Half Strength	T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T	100%
Dark	F, F, T, F, T, F, F, T, F, F, T, F, T, F, F, F, T, T, F, T, T, F, T, F, T, T, T, T, F, T, T, F, F	47%

The results in this experiment were slightly unexpected. While a dark room showed a great decrease in accuracy, a brightly lit room led to a drop in accuracy as well. The reasoning behind this could be that most of the training pictures were done in a room with natural lighting from the outside, as opposed to a bright artificial light, hence the AI better classifies a dimmer room more similar to a naturally lit one.

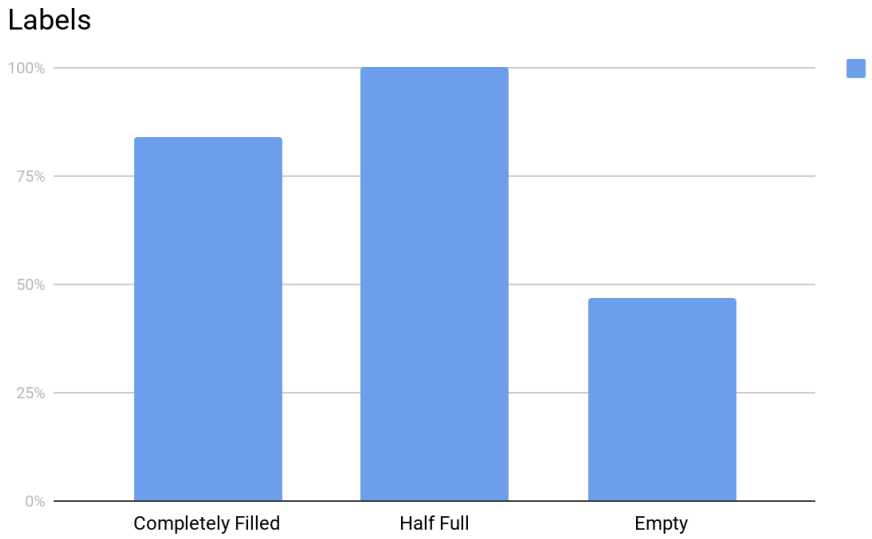


Figure 4: Label Accuracy in dark room

The experiments confirmed a few notions that were previously held. Firstly, having the system deactivated during night hours is a good idea, since it clearly does not work without light. Conversely, however, too much light has a negative impact on accuracy as well. The optimal place for the system would likely be outdoors in a shaded area or indoors but only exposed to ambient light. The expectation is that this system would be able to be used with most animals, indoors or outdoors. To this end it seems operational, and just requires careful positioning when initially set up. On the subject of water level, it appears to perform decently well at all water levels, even though confidence might be lower when the bowl is not full. The lack of confidence, however, does not appear to greatly impact performance.

5. RELATED WORK

Automatic water dispenser for pets is an automatically refilling pet bowl that flushes the water and refills it daily in accordance with a timer. My tool is similar as it is used for an identical purpose, refilling water bowls. Norris' tool [9] is superior in the aspect that it not only refills but also drains existing water to always maintain fresh, clean water in the bowl. The downside with Norris' tool is that it will refill water at each time interval even if there is water still in the bowl. This leads to water being wasted.

A Cloud Image Classification for Thrash Collecting LEGO Mindstorms EV3 Robot that uses Google Cloud's Vision API [10-11] to identify plastic waste. Additionally, it is able to identify trash bins and differentiate between trash bin types to correctly dispose of collected trash. Both my project and this one-use Vision to train our AI models, though the difference lies in that this project utilizes it for two functions: identifying waste and identifying trash cans. The prototype shown in the report appears to mainly use color to easily identify trash cans, rather than the symbols that are used in public areas. Compared to my own project, this one's AI is most likely simpler and possesses fewer images, as color identification is far easier than identifying water level.

International Journal of Landscape Architecture Research uses Google Vision API to detect landmarks and capture the visual character of landscapes. For example, the model would be able

to determine whether a picture of a landscape contains forests, or the ocean, or a city etc. This project's similarity with mine lies that it also uses Google Vision API to derive data from a picture. The difference lies in that this AI uses 320 labels compared to my 2 labels. Additionally, because of the number of labels, along with the complexity of the pictures entered, the number of training images far exceeds mine. Overall, this project shows the potential of Google Vision.

6. CONCLUSION AND FUTURE WORK

My project utilizes AutoML Vision to train an AI model that can identify the water level of a bowl. This is combined with a system that opens and closes a valve, which is able to then fill said bowl when the water level is detected to be low. Together these components provide a method for refilling water bowls for pets autonomously.

Currently the method, while functional, is not the most efficient way to accomplish the task. Its accuracy decreases heavily when there is not sufficient lighting. The completed project is bulky and difficult to transport. It takes up a fair amount of space, making it unsuitable for smaller animal enclosures. However, its supports also prevent larger animals from accessing the bowl. Additionally, the supports may encourage animals to climb on it, which would damage it or the animal. Finally, it depends heavily on the bowl being the same in all aspects except for the water. If the bowl is moved off center, or if something is placed within the bowl, the AI will cease to function correctly. Using a different bowl leads to similar issues.

A few solutions have been thought up to combat the current issues. An LED could be installed onto the raised platform, so that it illuminates the bowl, solving the lighting issue. A plastic shell could be used to cover the sensitive components, so climbing animals will not damage it. The legs and base of the platform could be altered to allow animals to access the bowl more easily. A possible solution could be to remove the base and configure the legs to be at an angle.

REFERENCES

[1] Poffenroth, Kevin. "Animal drinking water supply apparatus." U.S. Patent 5,452,683 issued September 26, 1995.
[2] Ewell, Anthony S. "Automatic pet food dispenser." U.S. Patent 5,433,171 issued July 18, 1995.
[3] Krishnamurthy, S. "Automatic pet waterer." U.S. Patent 6,928,954 issued August 16, 2005.
[4] King, Wayne. "Automatic waterbowl for pets." U.S. Patent 6,253,709 issued July 3, 2001.
[5] Graves, James. "Endless water bowl." U.S. Patent Application 10/367,019 filed July 28, 2005.
[6] Honeycutt, Jennifer A., Jenny QT Nguyen, Amanda C. Kentner, and Heather C. Brenhouse. "Effects of water bottle materials and filtration on Bisphenol A content in laboratory animal drinking water." Journal of the American Association for Laboratory Animal Science 56, no. 3 (2017): 269-272.
[7] Sexton, James E. "Pet food dish elevating assembly." U.S. Patent 5,584,263, issued December 17, 1996.
[8] Olde, Jarl Rune. "Automatic water dispenser." U.S. Patent 3,868,926, issued March 4, 1975.
[9] Norris, J. (2003). U.S. Patent Application No. 10/426,865.
[10] BENLİAY, A., & ALTUNTAŞ, A. (2019). Visual Landscape Assessment with the Use of Cloud Vision API: Antalya Case. International Journal of Landscape Architecture Research (IJLAR) E-ISSN: 2602-4322, 3(1), 07-14.
[11] Othman, Z., Abdullah, N. A., Chin, K. Y., Shahrin, F. F. W., Ahmad, S. S., & Kasmin, F. (2018). Comparison on Cloud Image Classification for Thrash Collecting LEGO Mindstorms EV3 Robot. International Journal of Human and Technology Interaction (IJHaTI), 2(1), 29-34.

A SMART INTERNET-OF-THINGS APPLICATION FOR SHOE RECOMMENDATIONS USING PRESSURE SENSOR AND RASPBERRY PI

Yutian Fan[1], Yu Sun[2] and Fangyan Zhang[3]

[1]Milton Academy, Milton, MA, 02186, USA
[2]Department of Computer Science California State Polytechnic University, Pomona, CA, 91768, USA
[3]ASML, San Jose, CA, 95131, USA

ABSTRACT

Running is one of the most important and simple sports spanning various ages, which can train throughout body and muscle. For running, proper shoes not only improve runners' performance but also protect them from injury to some extent. However, runners have difficulty in finding a pair of shoes which fit runners' gait patterns and feet shape very well. The process of selection of shoes is not effective and necessarily accurate. In this paper, we propose a new tool which facilitates the process by employing electronic sensors to the insoles of shoes and collecting feet information for runner accurately. It is helpful for runners to find the best fit shoes.

KEYWORDS

Machine learning, Firebase, Mobile application, Model fitting

1. INTRODUCTION

1.1. The Importance of Shoes

Many say that running is a sport that little to no requirement for equipment: all that one needs is a pair of shoes [1][2]. However, this one pair of shoes can have an immense long-term impact to the runner. Shoes that don't fit the runner can cause injuries varied from average shin splints to achilles tendonitis and insertional achilles tendinopathy. Other than preventing health risks, a fitting pair of training shoes can also increase the speed of a runner. Different models of shoes targets to support different kinds of feet, and the right kind of support can provide the runner better balances and lighter strides [3][4]. Competitive athletes, even in middle schools, are suggested to buy a new pair of shoes every year to avoid the disadvantages from overused, unbalanced shoe sole. Professionals always have multiple pairs of customized shoes to maximize the result of both their training and competing.

1.2. Current Ways of Shoe Picking

Currently, a lot of non-competitive runners simply go to sports stores such as Dicks and purchase whatever shoes that has appealing appearance. Often, these runners get shoes that don't actually fit with their gait patterns and feet shape. They either suffer from injuries previously mentioned or simply can't reach their full potential. Competitive runners usually go to stores like marathon sports and consults the staff there to get the right shoe model. However, the process is time

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 137-142, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101214

consuming and not necessarily accurate. The staff looks at the runner's feet with different shoes on, asks about the runner's events or distances, and then suggest the runner several models for their subjective opinion on the comfort level. The shoe-picking process is good but can be better- the staff can't really see the foot or feel the pressure applied to the shoes directly.

1.3. Goals of This Project

This project aims to improve the shoe selecting process by employing electronic sensors to the insoles of shoes to collect all the data needed to judge the fitting model for a runner. The shoes picked according to these data can be the best fits for the runners and maximizes both their comfort and their performance during practices and competitions.

The project has 4 main components: The Raspberry Pi, the Firebase [5][6], the Repl server and the mobile app.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the experiment and designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the related work. Finally, Section 5 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

2.1. The Hardware

The shoe insoles are going to take the full weight of the user's body, therefore where to put what kind of sensors is the first question. Currently, the project contains a Raspberry Pi and the breadboard full of circuits. Both the Pi and the breadboard are too big to be in a shoe. The size of the wire can be solved by making customized chips, but the alternative for the Raspberry Pi is not yet find.

2.2. The Software

There are a few requirements for the current software to function properly: 1, The Raspberry Pi [7][8] has to be connected to the WIFI for it to be able to send data to the Firebase. 2, The Firebase has to be checked and free of excessive data. 3, The Repl server has to be on and running (if the browser tab is closed, the server can't run). These requirements are easy to satisfy when the product is used by only one person. When the number of users increases, parts such as the Firebase and the Repl server needs better alternatives to support large quantity of access.

3. SOLUTIONS

3.1. Overview of the Solution

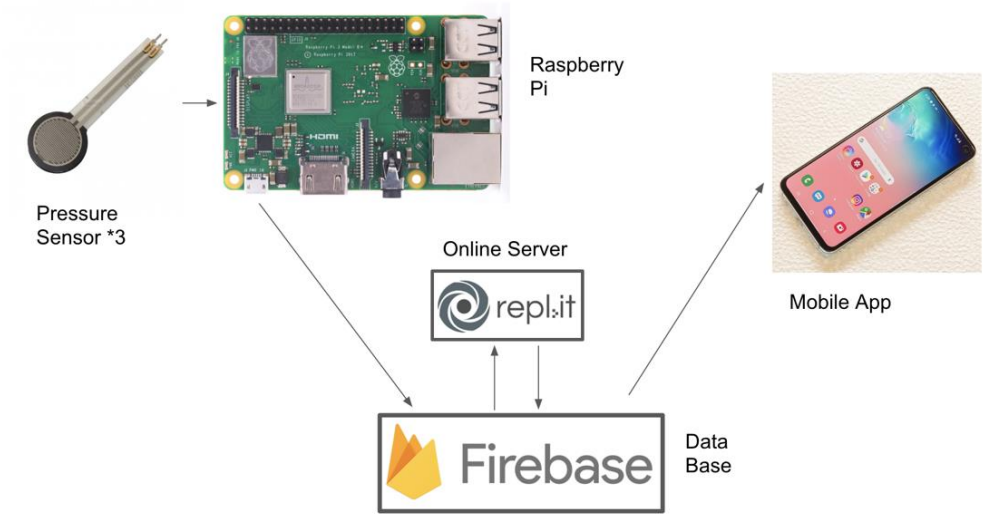


Figure 1: Overview of the solution

3.2. The Raspberry PI

The Raspberry Pi is the main hardware of this project. Currently, 3 pressure sensors are connected to the Pi via breadboard. The more pressure that are applied to the sensors, the more frequently the sensors send signals to the Pi. The Pi is programmed to calculate the timespan between each signal and therefore judge the pressure that is applied to the sensors. Then, the Pi sends the calculated pressure to the Firebase, an online Database.

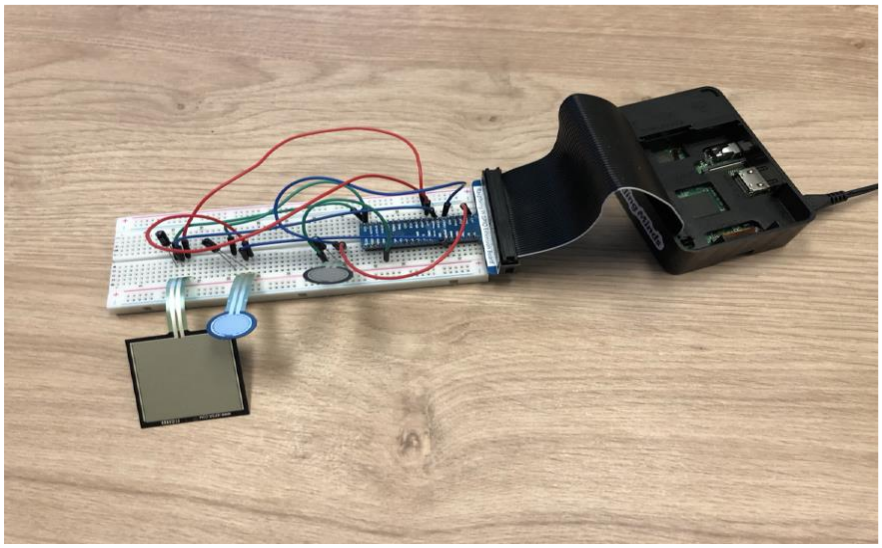


Figure 2: Hardware of Raspberry Pi

```
def rc_time (pin_to_circuit):
    count = 0

    GPIO.setup(pin_to_circuit, GPIO.OUT)
    GPIO.output(pin_to_circuit, GPIO.LOW)
    time.sleep(0.05)

    GPIO.setup(pin_to_circuit, GPIO.IN)

    start_time = time.time()
    while (GPIO.input(pin_to_circuit) == GPIO.LOW):
        count += 1
        if time.time() - start_time > 0.1:
            return 0
    return count
```

Figure 3: Code to get reading from one sensor

3.3. The REPL Server

The Repl is the online server, and most of the calculation happens here. The raw data in the Firebase are extracted and converted to average pressure and stride lengths. These data can be used to determine where the pressure point of the foot is, and which model is to the most suitable for the foot shape. After the calculations, Repl sends this information back to Firebase to store.

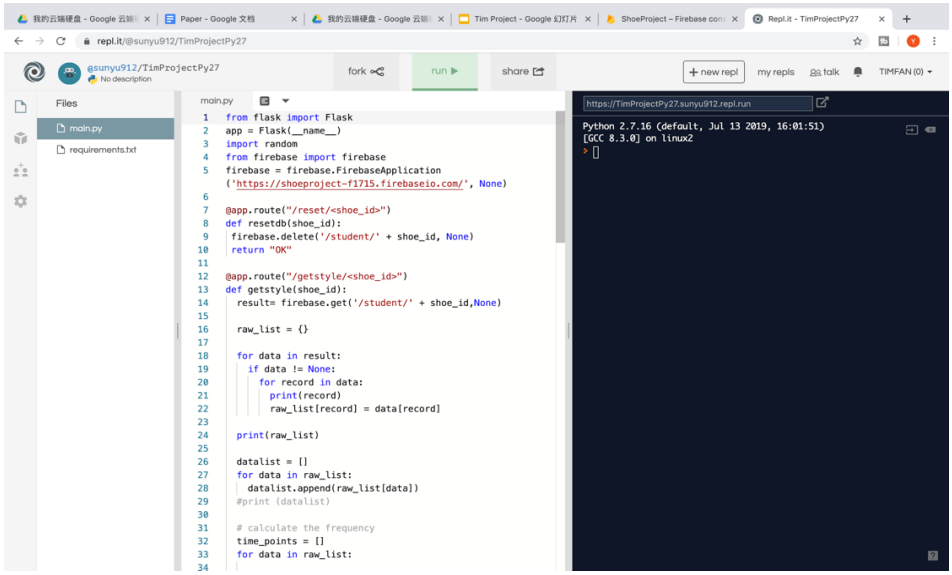


Figure 4: Overview of Repl Server running

3.4. The Mobile Application

The Mobile App [9][10] is the direct connection between this system and the users. The users, after uploading their data from the smart insole, can access their data by imputing the serial number of the insole. Then the mobile app extracts all the information calculated by Repl from the Firebase and presents the information to the user. The information includes and not limited to: the average pressure on each sensor, the stride frequency and the suggested model.

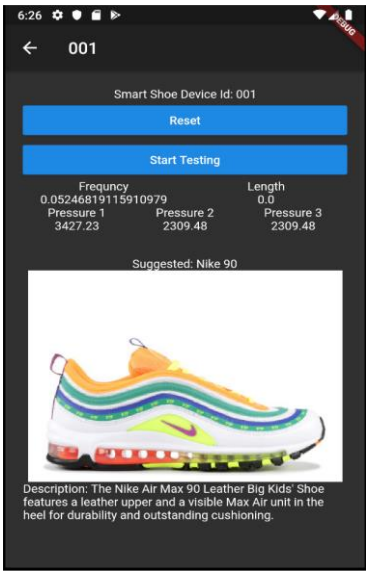


Figure 5: Screenshot of Mobile App

4. RELATED WORK

Smart sole [11] designed its product for a wide range of uses and people. The shoes include health analytics, smart connectivity, and sneaker design. The shoes have movement sensors to analyze pronation, supination, propulsion levels, impact force, fatigue, posture, steps, calories, and more creating precise data that prevents injuries and improves the user's health. The shoes can be connected via Bluetooth and a multi-function mobile app. The shoes have auto-lacing, temperature regulation with heating, and more. The Smartshoe is designed into an "ultra-light, premium leather and Neotech EVA shoe".

The shoes have pressure sensors in the soles that sense when to put the foot inside and triggers an algorithm that allows an automatic lacing. With integrated LEDs, the shoes can alert a user of low battery or a tight fit. Moreover, these shoes do not need charging every day, and the charge can last up to two weeks.

HOVR Phantom and HOVR Sonic [12] shoes released in February 2018 by Under Armour have inbuilt sensors to record a number of metrics important for runners. These include pace, distance, steps, stride, and cadence. These chip-laden shoes can be easily synced to the Map My Run app and are compatible with iOS [9] and Android [8] phones. The users can experience zero gravity and a great energy run with the shoes' excellent cushioning properties and comfort.

These designs are advanced smart shoes. They are created by big company with professional scientists and designers. However, most of these products are not for athletic uses. And each of these shoes still has just one model. They just can't possibly fit everyone's feet. The Smart Sole project aims not to create the best pair of shoes but the best way for the users to find their own best fit shoes. As for shoe picking, nobody has yet to make a successful and popular product.

5. CONCLUSION

Currently, the Smart Sole project is nothing but an experiment and a prototype. The size of the hardware needs to be shrunk down and the software needs to be upgraded for public use. However, the potential of this project is immense. People don't need to create the perfect shoe.

There are so many models or running shoes in the world that hardly anyone don't have a fit. The tricky part is to find that right fit for everyone. Smart Sole aim to solve this problem and link the runners to the best shoes they never had. If Smart Sole is proved to be successful, it can benefit all the runners in the world. While, this project has a lot of space for improvement. I plan to take some steps forward in the following fields:

- Find a replacement for the Repl server.
- Shrink down the size of the hardware, definitely replacing the breadboard with chip and potentially replacing the Raspberry Pi.
- Make shoes in a pair, measure and separate the data from both feet.
- Make the pairing of the shoe and Wi-Fi connecting easier for none-coders.

ACKNOWLEDGEMENTS

I would like to give special thanks to Doctor Yu Sun, who introduced me and taught me a lot of ways to utilize Flutter, Repl, and Firebase. The hours of my research time in his university lab and his kind inspiration made this project possible. I am also grateful to FangyanZhangand Dylan Lazar for helping me with the circuit building and Pi software testing. Also, thanks to Mr. Chris Hales for providing me with great support while I was doing my further research of this project in my school (Milton Academy). I was able to build this project from scratch only because of their help.

REFERENCES

[1] Young, W. B., R. James, and I. Montgomery. "Is muscle power related to running speed with changes of direction?" Journal of Sports Medicine and Physical Fitness 42, no. 3 (2002): 282-288.
[2] McKenzie, D. C., D. B. Clement, and J. E. Taunton. "Running shoes, orthotics, and injuries." Sports medicine 2, no. 5 (1985): 334-347.
[3] Wezel, Frank V., and Terry Mackness. "Running shoes." U.S. Patent 4,624,061, issued November 25, 1986.
[4] Richards, Craig E., Parker J. Magin, and Robin Callister. "Is your prescription of distance running shoes evidence-based?." British journal of sports medicine 43, no. 3 (2009): 159-162.
[5] Alsalemi, Abdullah, Yahya Al Homsi, Mohammed Al Disi, Ibrahim Ahmed, FaycalBensaali, Abbes Amira, and Guillaume Alinier. "Real-time communication network using firebase cloud IoT platform for ECMO simulation." In 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 178-182. IEEE, 2017.
[6] Ferdoush, Sheikh, and Xinrong Li. "Wireless sensor network system design using Raspberry Pi and Arduino for environmental monitoring applications." Procedia Computer Science 34 (2014): 103-110.
[7] Upton, Eben, and Gareth Halfacree. Raspberry Pi user guide. John Wiley & Sons, 2014.
[8] Butler, Margaret. "Android: Changing the mobile landscape." IEEE Pervasive Computing 10, no. 1 (2010): 4-7.
[9] Seabrook, Heather J., Julie N. Stromer, Cole Shevkenek, AleemBharwani, Jill de Grood, and William A. Ghali. "Medical applications: a database and characterization of apps in Apple iOS and Android platforms." BMC research notes 7, no. 1 (2014): 573.
[10] Janssen, Mark, JeroenScheerder, Erik Thibaut, AarnoutBrombacher, and Steven Vos. "Who uses running apps and sports watches? Determinants and consumer profiles of event runners' usage of running-related smartphone applications and sports watches." PloS one 12, no. 7 (2017): e0181167.
[11] Mial, Yurri. "Shoe sole." U.S. Patent Application 29/664,224 filed May 14, 2019.
[12] What It's Like To Run In Under Armour's HOVR Sonic Running Shoes https://www.besthealthmag.ca/best-you/running/hovr-sonic/

FOREX DATA ANALYSIS USING WEKA

Luciana Abednego and Cecilia Esti Nugraheni

Department of Informatics, Parahyangan Catholic University, Indonesia

ABSTRACT

This paper conducts some experiments with forex trading data. The data being used is from kaggle.com, a website that provides datasets for machine learning and data scientists. The goal of the experiments is to know how to design many parameters in a forex trading robot. Some questions that want to be investigated are: How far the robot must set the stop loss or target profit level from the open position? When is the best time to apply for a forex robot that works only in a trending market? Which one is better: a forex trading robot that waits for a trending market or a robot that works during a sideways market? To answer these questions, some data visualizations are plotted in many types of graphs. The data representations are built using Weka, an open-source machine learning software. The data visualization helps the trader to design the strategy to trade the forex market.

KEYWORDS

forex trading data, forex data experiments, forex data analysis, forex data visualization, weka

1. INTRODUCTION

When planning a forex trading system, a trader needs to carefully design the system and extensively test it. Besides the help of some technical indicators and fundamental analysis [1][2], a trading system needs to set many risk management parameters, such as stop loss and take profit [3]. These parameters play an important rule to determine the trader's target profit and limit the loss risk of each open trade.

To investigate the ideal level for risk management parameters and the trading system, this research tries to find the answer to those questions. Some experiments are conducted in the H1 timeframe, which updates the price hourly. Two main currency pairs with different time ranges are used in this paper: EUR/USD (1 year) and USD/JPY (20 years). The experiments use the dataset from Kaggle, a website that provides many kinds of datasets for machine learning and data scientists [4]. Some data visualization techniques are used to represent the result of these data using Weka. Weka is open-source software that provides many machine learning techniques and data visualization tools [5].

2. RELATED WORKS

Some previous researches have been conducted. In [1], we developed some forex robot with technical analysis. Then in [2], we tried to compare the technical robot performance with a fundamental robot that extract fundamental news that affect forex prices from a website. The fundamental robot makes decisions based on the updated news. In [3], we compared some techniques of money managements in forex trading. In this paper, we further investigate the characteristics of some major pairs in forex trading. The result can then be the basis for the next robot algorithms.

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 143-155, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101215

3. PROBLEM DESCRIPTION AND ANALYSIS

This section describes the problem domain and data that want to be investigated in this research.

3.1. Forex Trading

Forex (foreign exchange) is a global marketplace where the banks, corporations, investors, and individual traders exchange foreign currencies for a variety of reasons. The fluctuations of these currencies are the target for traders for making some profit. But at the same time, a trader risks their account when the market moves against his open position. The currencies are traded in pairs. The four major currency pairs are EUR/USD, USD/JPY, GBP/USD, and USD/CHF. Figure 1 shows the approximate volume breakdown per currency pair [6].

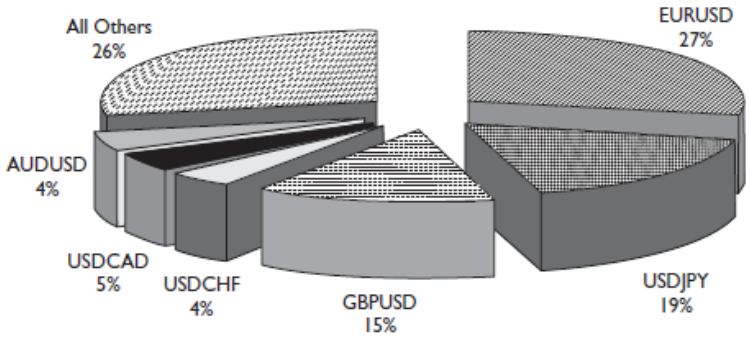


Figure 1. Estimated Trading Volume by Currency Pair

The forex market works 24 hours a day, 5 days a week. Table 1 shows the opening and closing times [6].

Table 1. Global Trading Hour Schedule

Time Zone	New York	GMT
Tokyo Open	7:00 p.m.	00:00
Tokyo Close	4:00 a.m.	09:00
London Open	3:00 a.m.	08:00
London Close	12:00 p.m.	17:00
New York Open	8:00 a.m.	13:00
New York Close	5:00 p.m.	22:00

3.2. Forex Risk Management

When a trader opens a position in the forex market, two actions can be taken: buy or sell. If the trader thinks that the price will go upward, he is supposed to open a buy position. On the contrary, if the trader considers that the price will go downward, he is supposed to open a sell position. After a trader chooses one of that action, but unluckily the market moves against its open position, the trader will lose. In this case, he must protect his account by limiting the loss he's suffered. There are many types of risk management strategies [3]. Some parameters that can be set to limit the loss of any open trade are stop loss and target profit. In this paper, some experiments are conducted to investigate the ideal level to set these parameters.

3.3. Data Preparation and Mining

This research uses past forex data that is gained from Kaggle, a website that provides many kinds of datasets for machine learning and data scientists [4]. We choose the H1 timeframe of the two top biggest volume currencies traded in the global market: EUR/USD & USD/JPY [6]. This raw data is then cleaned, transformed, and represented in some visualizations charts by using Weka. Weka is an open-source data mining and visualization framework. Weka was developed at the University of Waikato, New Zealand. Figure 2 shows the user interface of Weka. This paper uses Weka as a tool for data visualization and mining.

Figure 2. Weka Interface

4. EXPERIMENT SETUP AND RESULT

Experiments are conducted to the top two biggest volume traded currency pairs: EUR/USD and USD/JPY. The H1 timeframe for 1 year is used for all the experiments. As mention before, the datasets that are used in these experiments are from Kaggle.com [3], a website that provides many kinds of datasets for machine learning and data science purposes. These datasets use pip (price in percentage), which is the smallest value by which a currency may fluctuate in the forex market [5]. The goals of these experiments are explained in the following sections.

4.1. Experiment with Information Gain

The goal of this experiment is to sort the most important attributes to the price change above 10 pips. Table 2 shows the experimental result.

Table 2. Information Gain Experiment.

No.	Attribute	Information Gain
1	Date	0.1438
2	Volume	0.125
3	Low	0.0661
4	Close	0.0661
5	High	0.0657
6	Open	0.0656
7	Hour	0.0599

Two top attributes are date and volume. This shows that in some certain times, the forex market is trending (the price change above 10 pips) and the number of volumes influences this trend.

4.2. Experiment With 10 Pips of Currency Fluctuation

Based on the first experiment, the dataset is categorized based on the pip change that shows whether the market is on the condition of trending or sideways. So, in this experiment, a new attribute, **Class 10 Pip Change**, was added based on the open price of the next candle minus the close price of the previous candle. This attribute has three possibilities of value:

- **-10pips**: the price decrease above 10 pips
- **ranging**: the price change below 10 pips
- **+10pips**: the price increase above 10 pips

Table 3 and Figure 3 show the result of this experiment.

Table 3. Experiment Data based on10 Pips Change Class

No.	Attribute Value for Class 10 Pip Change	Number of Records
1	-10pips	955
2	ranging	4722
3	+10pips	841
Total		6518

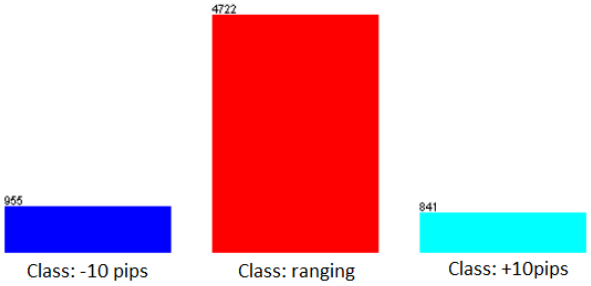


Figure 3. Experiment Data based on 10 Pips Change Class

From the data distribution, it can be concluded that most of the time EUR/USD is fluctuated below 10 pips, which shows the condition of sideways or ranging. From 6,518 different records, 4,722 records of it (72%) is the change below 10 pips. This data can be used to determine the algorithm of how to trade the forex currency pair. The algorithm must be dealt with ranging market. From this experiment, the trader can decide how many percent of winning chance if he set the forex parameters such as stop loss or target profit level at a certain position. The chance of uptrend or downtrend can be calculated as (955+841) / 6518 =0.275=27.5%.

4.3. Experiment Data Based on 25 Pips of Currency Fluctuation

Similar with the previous experiment, the dataset is categorized based on the 25-pip change that shows whether the market is on the condition of trending or sideways. So, in this experiment, a new attribute, **Class 25 Pip Change**, was added based on the open price of the next candle minus the close price of the previous candle. This attribute has three possibilities of value:

- **-25pips**: the price decrease above 25 pips
- **ranging**: the price change below 25 pips
- **+25pips**: the price increase above 25 pips

Table 4 and Figure 4 show the result of this experiment. From the data distribution, it can be concluded that most of the time EUR/USD fluctuates below 25 pips, which shows the condition of sideways or ranging. From 6,518 different records, 6,081 records of it (93%) is the change below 25 pips. This data can be used to determine the algorithm of how to trade the forex currency pair. The algorithm must be dealt with the ranging market. From this experiment, the trader can decide how many percent of winning chance if he set the forex parameters such as stop loss or target profit level at a certain position. The chance of uptrend or downtrend can be calculated as (235+202)/ 6518=0.067=6.7%.

Table 4. Experiment Data based on 25 Pips Change Class

No.	Attribute Value for Class 25 Pip Change	Number of Records
1	-25pips	235
2	ranging	6081
3	+25pips	202
Total		6518

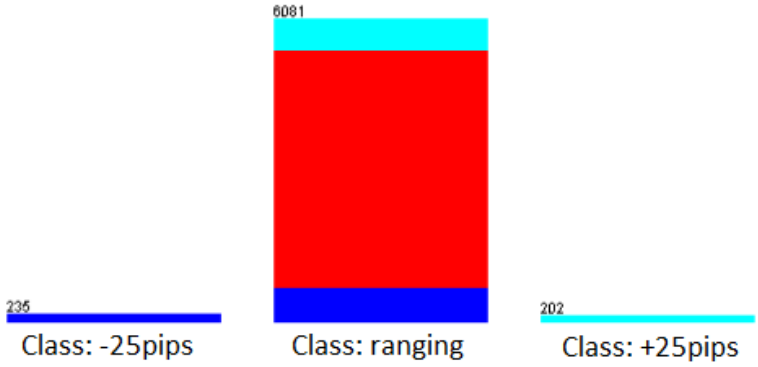


Figure 4. Experiment Data based on 25 Pips Change Class

4.4. Experiment Comparison of Uptrend to Downtrend EUR/USD

The goal of this experiment is to know the comparison of the up prices to the down prices in EUR/USD pairs. A new attribute **Class Price Up** was added in this experiment, with two possibilities of value: TRUE or FALSE. TRUE means the next close price is higher than the previous close price. FALSE means the contrary. Table 5 and Figure 5 show the result of this experiment.

Table 5. Data Experiment based on Price Up Class

No.	Attribute Value for Price Up Class	Number of Record
1	FALSE	3362
2	TRUE	3156
Total		6518

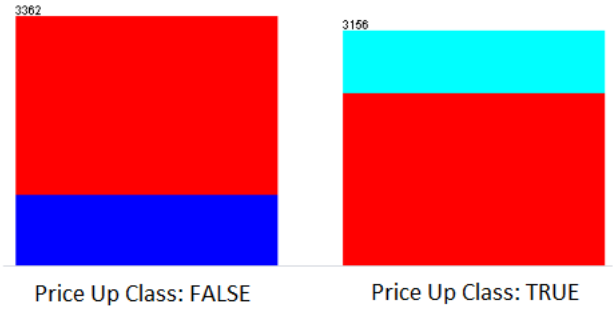


Figure 5. Experiment Data based on Price Up Class

This experiment shows that the number of up prices almost comparable with the number of down prices. From this experiment, the trader has a 50:50 percent chance to buy or sell decisions.

4.5. Experiment of EUR/USD Trending Market Time

The goal of this experiment is to know the tendency of the time when the EUR/USD is trending during a day. The price of each transaction to the time of a day is plotted in the chart below (see Figure 6).

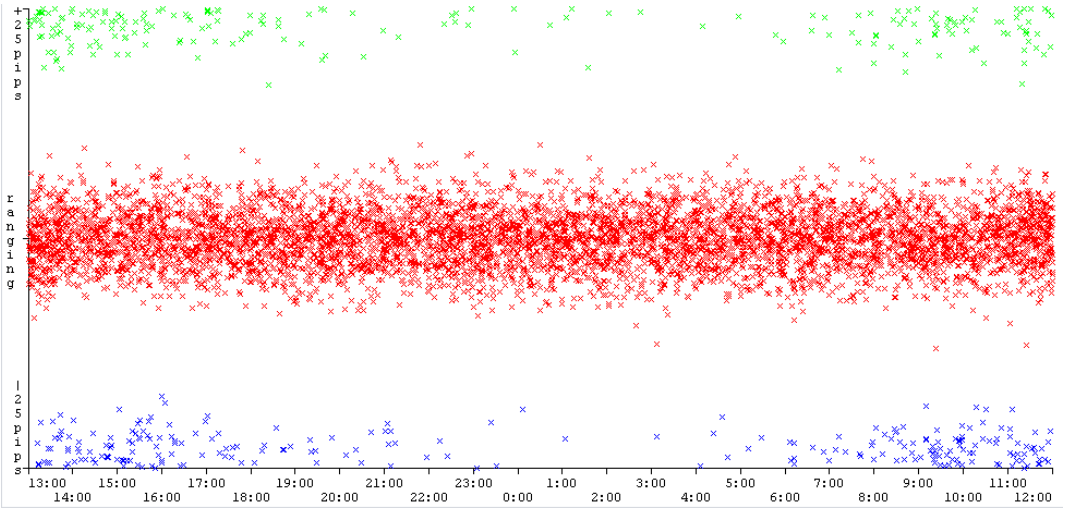


Figure 6. Market price to the time chart

The X-axis shows the time of the days and the Y-axis shows the attribute value of Class 25 Pips Change: -25pips, ranging, or +25pips. The red dots show the ranging market that happens most of the time of any day. The green dots represent the up-trending market that moves above 25 pips. From the chart, most of the trending market happened during office hours (7a.m. to 5 p.m.). Outside that time, the trend rarely happened.

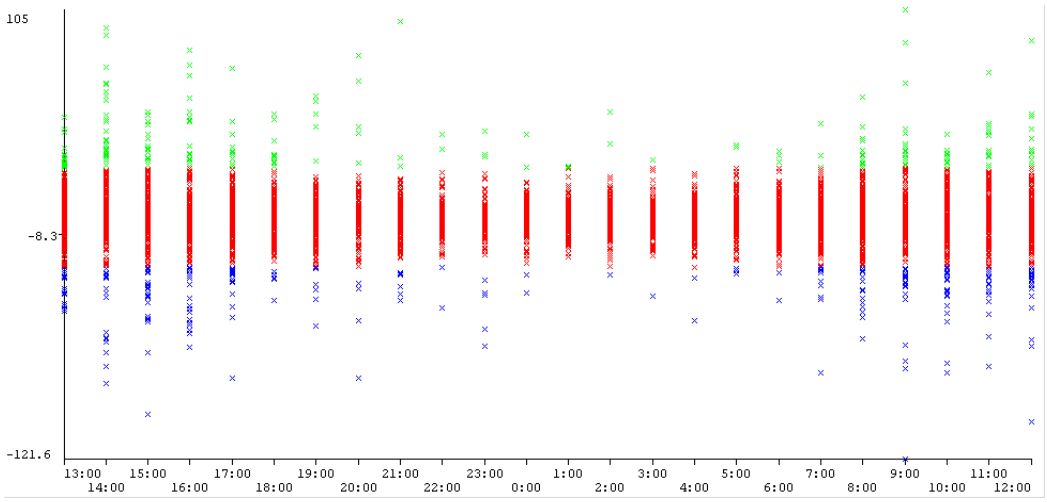


Figure 7. EUR/USD Fluctuation Range (in pips)

From this experiment, if the trader's used the trending algorithm, it would be better to apply it during office hours. On the other hand, if the trader uses an algorithm that can be dealt with ranging markets, it can be applied most of the time of the day. The trader can set the forex parameters, such as stop loss and take profit below 25 pips to gain more profit or reduce the risks.

Figure 7 shows the pip change range to time in the EUR/USD forex market. From this figure, the most trending market happened at about 14:00 - 15.00. If some of this data is selected (see Figure 8), when the market starts to open, the possibility of the downtrend is more often than the uptrend.

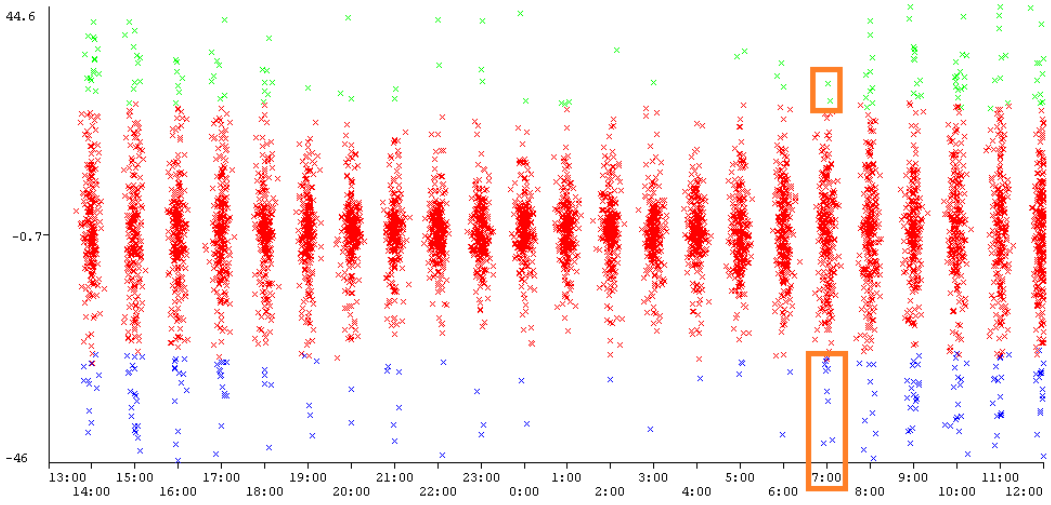


Figure 8. The Possibility of the Downtrend When the Market Starts to Open

4.6. Experiment Time of EUR/USD Trending Market

The goal of this experiment is to know the characteristics of another major currency pair in forex: USD/JPY. In this experiment, we used a large dataset (H1 timeframe, for 20 years from 1999 to 2019), that consists of 128,800 records of transactions. This data is categorized into groups of attribute Class 10 Pip Change:

- **-10pips**: the price decrease above 10 pips
- **ranging**: the price change below 10 pips
- **+10pips**: the price increase above 10 pips

Table 6 and Figure 9 show the result of this experiment. From the data distribution, 69.6% of all the transactions fluctuated below 10 pips, which shows the condition of sideways or ranging. While the other 15.3% and 15.1% each is the up and down trend (more than 10 pip change). This shows that the opportunity to buy and sale is comparable for each new open position.

Table 6. Experiment with Class 10 Pips Change.

No.	Attribute Value for Class 10 Pip Change	Number of Records
1	-10pips	19, 494
2	ranging	89, 618
3	+10pips	19, 688
Total		128, 800

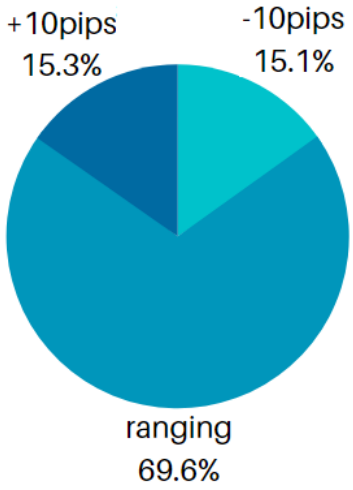


Figure 9. Experiment with Class 10 Pips Change in USD/JPY

4.7. Experiment With 25 Pips of Currency Fluctuation

Like the previous experiment, the dataset is categorized based on the 25-pip change that shows whether the market is on the condition of trending or sideways. So, in this experiment, a new attribute, **Class 25 Pip Change**, was added based on the open price of the next candle minus the close price of the previous candle. This attribute has three possibilities of value:

- **-25pips**: the price decrease above 25 pips
- **ranging**: the price change below 25 pips
- **+25pips**: the price increase above 25 pips

Table 7 and Figure 10 show the result of this experiment. From the data distribution, 93.8% of all the USD/ JPY transaction records fluctuated below 25 pips, which shows the condition of sideways or ranging. This data can be used to determine the forex risk management parameter such as stop loss and take profit. If they are set above 25 pips, the winning possibility is below

6.1%. It also can be concluded that if the trader uses an algorithm or strategy that can work when the market is ranging/ sideways, the profit gain will be bigger than an algorithm that only works in a trending market. This is because 93.8% of all the last 20 years transactions are ranging below 25 pips.

Table 7. Experiment with Class 25 Pips Change.

No.	Attribute Value for Class 25 Pip Change	Number of Record
1	-25pips	4,125
2	ranging	120,865
3	+25pips	3,810
Total		128,800

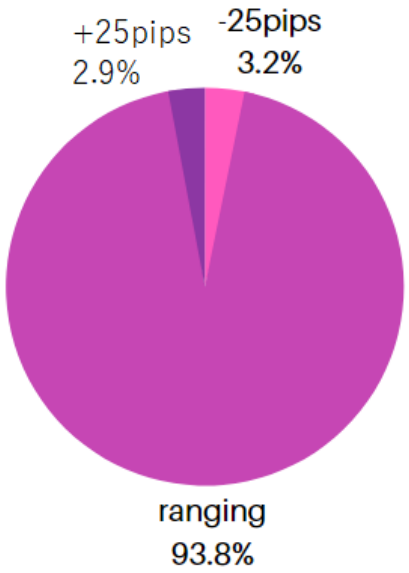


Figure 10. Experiment with Class 25 Pips Change in USD/JPY

4.8. Experiment with USD/JPY Trending Time

The goal of this experiment is to know the best time to trend USD/JPY if a trader uses an algorithm that counts on-trend. Figure 11 shows the transactions plotted against time. The red dots show the upward trend above 25 pips. The green dots show downward trends of more than 25 pips. The blue dots represent the ranging market. The trend not only happened during the office hours (7 a.m. - 5 p.m.) but also during midnights (11 p.m. - 3 a.m.).

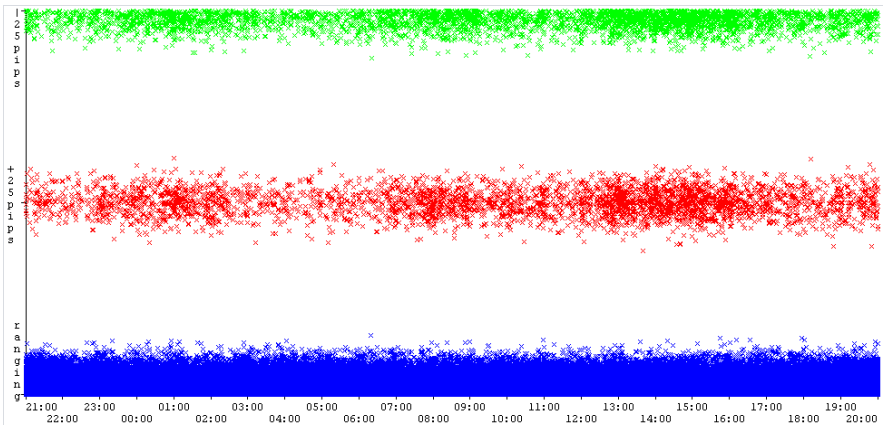


Figure 11. Experiment with Class 25 Pips Change in USD/JPY

If we decrease the threshold to 10 pips, the chart will look like Figure 12. This data can be used to determine the level of stop loss and take profit. Most of the time, the market fluctuates between -10 pips to +10 pips.

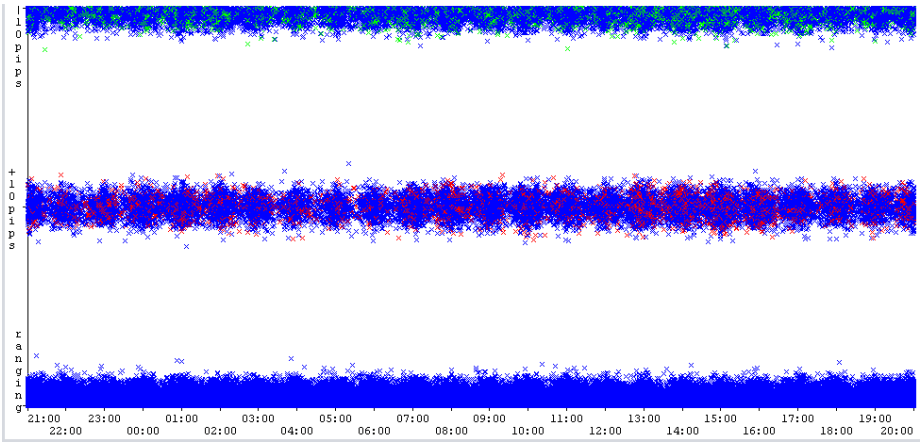


Figure 12. Price movement (in pip) plotted to time

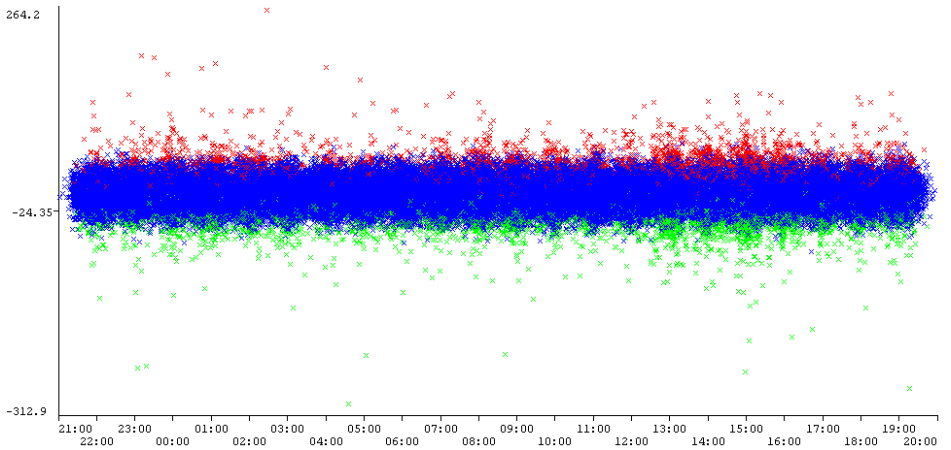


Figure 13. Price Movement (in pip) Plotted to Time (USD/JPY)

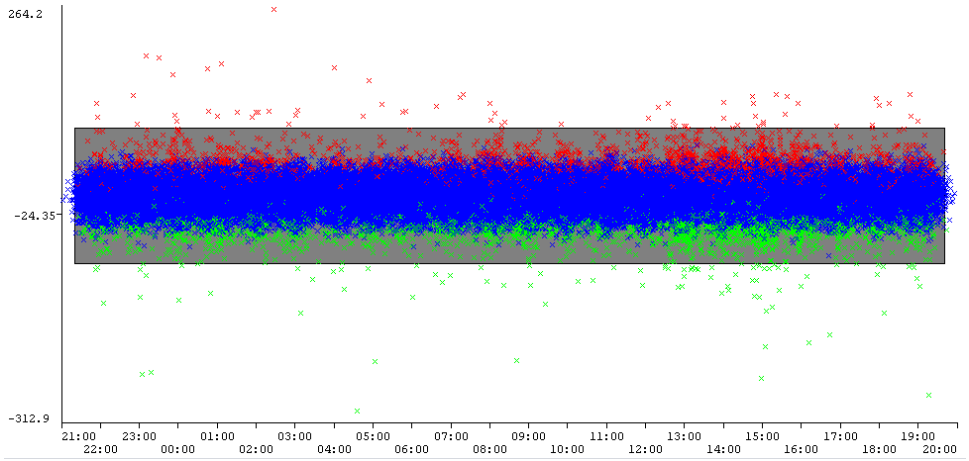


Figure 14. Choose only TheGray Area of this Data

Figure 13 shows the pip change plotted to the time of a day. From this plot, it can be seen that most of the trending market happened at about 1 p.m. to 4 p.m. If we ignore the outliers of Figure 13 (see Figure 14), we get the chart that is shown in Figure 15. The red dots show the upward trend above 25 pips, while the green dots show the downward trend.

Figure 16 shows the uptrend fluctuation range and Figure 17 shows the downtrend fluctuation range (both in pip).

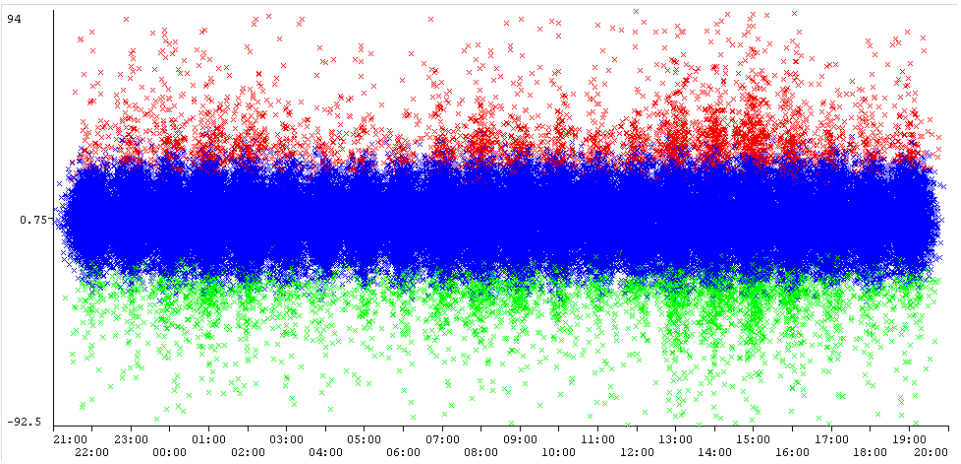


Figure 15. The Data without The Outliers

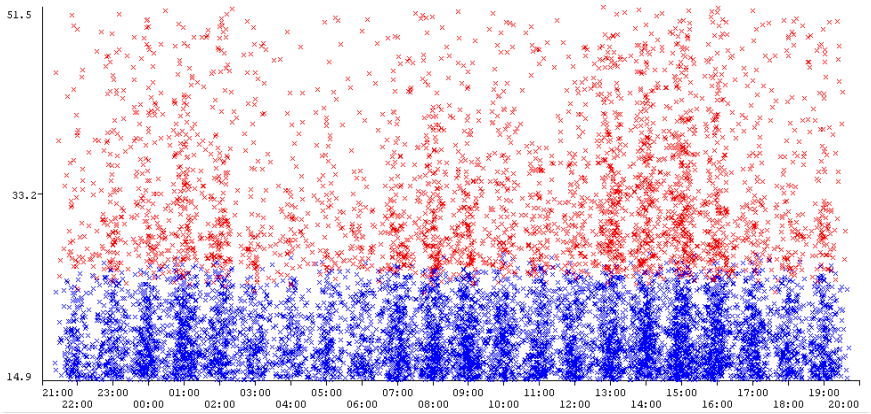


Figure 16. Range of the Uptrends of USD/JPY Pair (in Pip)

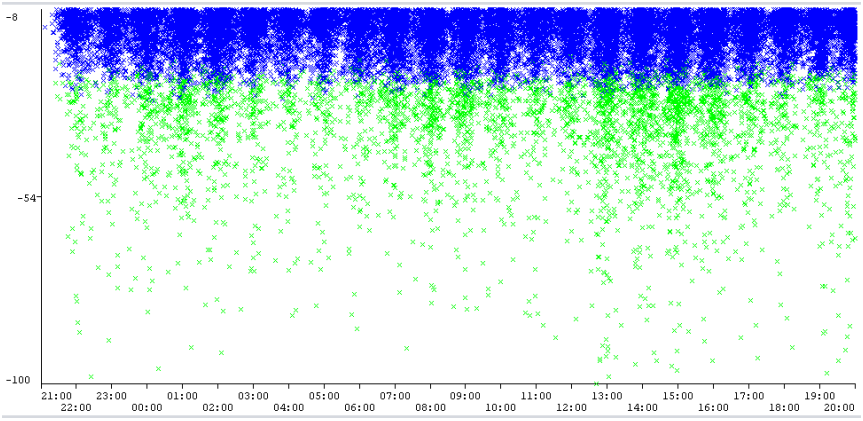


Figure 17. Range of the Downtrend of USD/JPY Pair (in Pip)

5. CONCLUSIONS

From the experiments, it can be concluded that most of the time, the forex market is ranging below 10 pips. This can be used to determine how a trading algorithm works. A forex trading robot that can deal with ranging markets is preferable than the one which only waits for the trending market. Most of the market trends happened during office hours (7 a.m. to 5 p.m.) for EUR/USD, but almost all the time for USD/JPY. The possibilities of winning between buy and sell actions are comparable for both major currencies pairs.

ACKNOWLEDGEMENTS

The authors would like to thank LPPM Parahyangan Catholic University for the research grant and the Department of Informatics Parahyangan Catholic University which supports the research.

REFERENCES

[1] L. Abednego, C. E. Nugraheni (2015). *Development of Forex Robot in MetaTrader 4*. International Congress on Engineering and Information Proceeding.

[2] L. Abednego, C. E. Nugraheni, I. Rinaldy (2018). *Forex Trading Robot with Technical and Fundamental Analysis*. Journal of Computers JCP 2018 Vol.13(9): 1089-1097 ISSN: 1796-203X. DOI: 10.17706/jcp.13.9.1089-1097

[3] L Abednego, CE Nugraheni (2018). *Development of Forex Trading Robot with Money Management.* Proceeding of Higher Education, Sydney, Australia.
[4] D. F. Jimenez (2020). *Forex currencies M1, M5, M15, M30, H1, H4, D1*. https://www.kaggle.com/lehomme/forex-currencies-m1m5m15m30h1h4d1/notebooks
[5] I. Witten, E. Frank, M. Hall, C. J. Pal (2016). *Data Mining Practical Machine Learning Tools and Techniques.* Morgan Kaufmann. Fourth Edition.
[6] J. Norris, T. Bell, A. Gaskill (2010). Mastering the Currency Market: Forex Strategies for High- and Low-Volatility Markets. McGraw-Hill.

AUTHORS

Full-time lecturer at Department of Informatics, Parahyangan Catholic University, Bandung, Indonesia.

RECODEZ: AN INTELLIGENT AND INTUITIVE ONLINE CODING EDITOR USING MACHINE LEARNING AND AI

Justin Kim[1], Yu Sun[2] and Fangyan Zhang[3]

[1]Los Osos High School, Rancho Cucamonga, CA 91739, USA
[2]California State Polytechnic University, Pomona, CA 91768, USA
[3]ASML, San Jose, CA 95131, USA

ABSTRACT

Recent years have seen a large increase in the number of programmers, especially as more online resources to learn became available. Many beginner coders struggle with bugs in their code, mostly as a result of a lack of knowledge and experience. The common approach is to have plenty of online resources that can address these issues. However, this is inconvenient to the coder, who may not have the time or patience to look for a solution. In this project, we address this problem by integrating the coding and error resolving environment. A website has been developed that examines code and provides simpler error messages that give a more comprehensive understanding of the bug. Once an error has been added to the database, the program can display the error more understandably. Experiments show that given several sample programs, our tool can extract the errors and report a more easily understandable solution.

KEYWORDS

Programming Environment, Python, Server, Database

1. INTRODUCTION

Programming is the act of writing instructions for a computer [1][2]. This often includes several processes, such as writing code and fixing errors (debugging). With the increasing popularity of programming, especially as more online resources become available that teaches beginners how to code, the necessity of making coding a simpler process becomes more apparent. Whereas many resources give help on the writing code aspect of programming, few exist that work on making errors easier to fix. However, fixing bugs is difficult even for the experienced programmer, and therefore this process must be simplified further for beginner programmers.

When a beginner coder writes their first programs, there are several large challenges that they face [4]. One of these is that code must be error-free. When a program is run, if there are errors or bugs, the program will not function as expected. Often, this results in the program failing to complete its task. However, a beginner programmer will often not know how to fix these errors. They lack the knowledge and experience necessary to resolve the bugs.

As another of the challenges that a beginner coder faces are that they lack the motivation to create programs, a simple bug could cause enough frustration for them to quit programming [12][13]. This is not ideal, as programming is an essential skill as more and more jobs become mechanized.

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 157-164, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101216

Without more programmers learning how to code, many will be left behind by those who have studied some computer science. Thus, it is important to improve the simplicity of errors and make them easier to understand.

Several resources attempt to create simpler errors. Some websites allow a user to ask questions and receive answers from other experienced coders [3][11]. These allow people who need help to reach other people who may have experienced similar issues and were able to solve them. However, this is not a convenient method. If the programmer cannot find another person with the same issue, then they must ask the question themselves. This can result in their question taking time to answer as few have the same problem and fewer have a solution. In the case that a solution exists, they often are not tailored to fit the programmers' needs, which can lead to confusion when trying to implement this fix into their own code.

Another system that creates simpler to understand errors is the programming environment [5][6]. They often have built-in software that checks for errors as you write code, and create error messages that are given to the user. The error messages are usually short and often highlight the location necessary to fix. However, they don't usually describe how to fix the error, and just notify the user that a problem exists. They must go to another resource to find out how to truly fix the error. Additionally, these error messages are often too short to be very descriptive, such as not including the definition of the error. Some error messages have extraneous information that does not help resolve the issue, such as error codes that wouldn't make sense to a beginner programmer.

In this paper, we present ReCodeZ, an online tool to improve the experience of beginner programmers in resolving errors. Similar to the programming environment, both the code and the output areas are part of the same application. This allows for more convenience as the programmer does not have to switch tabs or open another application to see the results of the code. However, our method of making debugging easier for beginners is to reword every error message into a more understandable form. Instead of having extra error codes that do not inform the programmer on what the problem is, we give a better response including pointing out where the bug is and several ideas on how to fix it. This is different from the standard programming environment because instead of keeping error messages in a hard to understand way, ReCodeZ outputs an improved version that helps the beginner programmer understand why they have that bug and what they can do to resolve the issue. Compared to other online resources that allow programmers to ask questions to other experienced programmers, our tool is more convenient. Instead of having to search online for help, a beginner programmer can stay on one site and find out how to debug their code.

We demonstrate how the above usage of both the coding and debugging environment on the same website can be both convenient and useful to the beginner programmer. To do so, we conducted a case study on several beginner programmers. We created several sample Python programs that all had an error. These were simple errors that a beginner would be expected to make. Then, by putting these errors into our tool, we determined whether our tool worked as expected. Additionally, we gave the ReCodeZ output to the beginners to see if they were more able to understand what the error was, comparing it to several of the other methods discussed previously.
The remainder of the paper is as follows. Section 2 gives the details on the challenges that we met in designing and testing our tool. Section 3 focuses on the details of our solution corresponding to the challenges stated in Section 2. Section 4 presents relevant details in the experimentation to analyze our solution. Section 5 gives an analysis of related work. Finally, Section 6 concludes and provides insight into future work related to this project.

2. CHALLENGES

2.1 Challenge 1: Detecting Errors

There are several types of errors that can be present in a piece of code, such as syntax errors and arithmetic errors. With all these different kinds of errors, detection becomes more difficult. In order to detect an error, we had to figure out what can cause that error. We then have to create an algorithm that can determine if the code has a certain piece that causes an error. We also have to determine other details about the bug, such as the specific location of the error. This process had to be repeated for every type of error that the Python programming language could possibly give.

2.2. Challenge 2: Describing Errors

For our tool to reach its goal of making error messages more understandable, the errors have to be described in a way that makes sense to a beginner. Some beginners may understand more about coding than others, and creating a one-size-fits-all description of the error is difficult. One person may be able to understand a certain description, but others may not. We had to find the balance between a long explanation and a summary of the error, creating an error message that is clear to any beginner who uses our tool.

For example, using the same error in Figure 1, we had to figure out how to describe that parentheses were missing. Some beginners may not understand that parentheses have to be balanced, so we had to explain what was missing from the code. However, we had to keep the messages short to appease the more experienced programmers, who would not want to read through a long error message.

2.3. Challenge 3: Possible Solutions to the Error

While there are a few errors that only have one solution, the vast majority of them are determined by the programmer's intent. Depending on what the user wanted to happen, the method to fix the error changes. This is a task that is difficult for humans, which means that many algorithms on a computer would not be able to predict what the user wants to happen. Because there can be many solutions to each error, a certain fix of the error may not be the one that results in their code behaving as they expected.

For example, continuing with Figure 1, there are several possible solutions to making the parentheses balanced. However, the possible solution displayed in the error message may not be the exact one that the programmer wanted. While this example is simple, in a more complex piece of code, such as an equation, the correct placement of parentheses depends on the desired function. In such cases, it may be best to leave it to the programmer to figure out what they need, but helping them understand the error as much as possible.

3. SOLUTION

ReCodeZ is an online Python programming environment that clarifies errors for beginner programmers [7]. ReCodeZ was designed to be flexible enough to handle all types of errors. Many Python errors can be handled by ReCodeZ, which has a modular method of implementing additional error messages. Error messages may be added, removed, or edited, allowing ReCodeZ to keep up with the improvement of programming languages. Additionally, other programming languages can be added to our tool. Therefore, it can be used for learning a variety of different programming languages and determining how to fix bugs in each language.

ReCodeZ implements many of the features of a standard programming environment while being online. Users can edit code, run programs, and see the outputs on one page. This allows more people to learn to program and improves convenience for people who do not have access to a computer.

An overview of the entire system is given in Figure 1. Our tool consists of four main parts: the website, the server, the interpreter, and the database. The website handles the user input and displays the output. The server is the main controller of ReCodeZ and combines the functions of the other components. The interpreter runs the code and gives output and an error message. The database holds all of the errors and can retrieve them. When a user runs a program on our website, it sends their code to our server. The server then runs that code using a Python interpreter and returns the output. However, if there is an error, the output is intercepted and an error message from our database is retrieved to replace it.

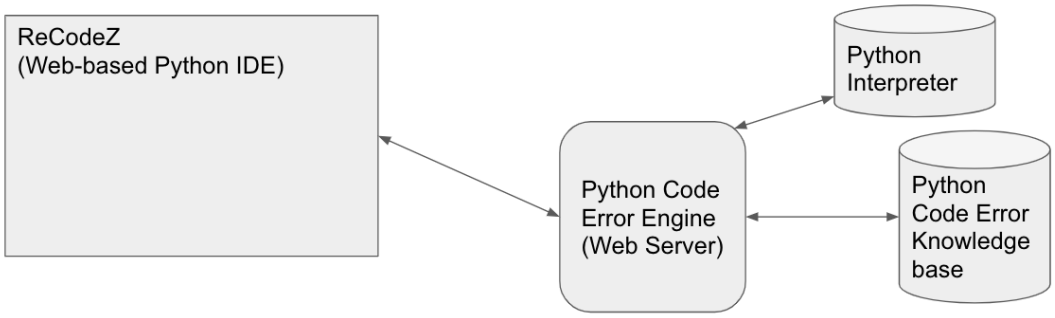


Figure 1. An overview of the system

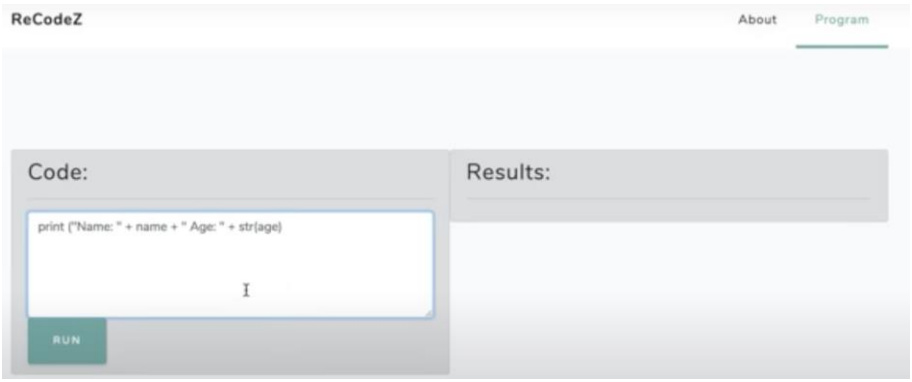


Figure 2. ReCodeZ interface

Our website is where the user can write code, submit it, and see the results. This is done through a text box located on the page that allows the user to type in their code or copy and paste it in. When the user clicks the run button, a JavaScript script submits the code to the server to check and run [15]. Once the server finishes processing the code and returns the output, the website updates the page and displays the output or error message.

The server takes the code passed to it by the website and passes it to the interpreter. After receiving the output, it creates a chain of error handlers. These error handlers are pulled from the database and linked together in a way that allows any errors to be handled in sequence.

As seen in Figure 3, there are three errors checked as a proof of concept. While more errors could be handled, ReCodeZ in the example checks if the code contains unbalanced parentheses, missing colons, or an undeclared variable. These error checkers traverse the code and return if they find an error or pass the code on to the next error checker. Finally, after all errors have been checked, the program gives the output back to the website. The server passes the user's code to the algorithm displayed. This runs another Python file, displayed that will compile and run the user's code. Finally, it splits the error messages into a separate output, allowing our program to capture and replace the errors.

```
private Result checkForErrors(String code) {
    TerminalAccess terminal = new TerminalAccess();
    TerminalOutput overallOutput = terminal.getTerminalOutput(code);

    // executed
    if (overallOutput.getOutput() != null) {
        return new Result(overallOutput.getOutput());
    }

    CodeFileTraverser fileTraverser = new CodeFileTraverser(file);
    ErrorChecker parentheses = new ParenthesesErrorChecker();
    ErrorChecker colons = new ColonErrorChecker();
    ErrorChecker varName = new VariableNameErrorChecker();

    parentheses.setSuccessor(colons);
    colons.setSuccessor(varName);

    fileTraverser.traverse(parentheses);

    int size = ErrorChecker.getMessages().size();
    String[] messages = new String[size];

    for (int i = 0; i < size; i++) {
        messages[i] = ErrorChecker.getMessages().remove();
    }

    return new Result(messages);
}
```

Figure 3. ReCodeZ error check

The database of ReCodeZ consists of several classes that each handle a different error. This allows for a modular design, as more classes can be added without needing to change the other classes. When the server constructs the chain of error handlers, each class can be added to the chain either on its own or with a reference to the next handler. When the error from running does not match the error ReCodeZ is checking, it moves on to the next error. This allows for the abstraction of how errors are checked. Each of the errors is a different class that takes in the output from running the code and checks it against a general form of the error. If the error matches, the program then gives the ReCodeZ error message back to the server to update the website for the user.

4. EXPERIMENT

To test ReCodeZ, we created 20 sample pieces of code. All of these samples were short with only a couple lines, and each had an error. This is so that we could test our tool in creating error messages. We submitted them through our website to get the 20 error messages corresponding to each error. To compare our solution to other programming environments, we also ran each piece of code in a different programming environment, PyCharm, to get 20 more error messages [14]. We used a small sample of 10 beginner programmers in the Southern Californian area. These are

programmers who have only been learning coding for less than a year. While this may not represent all beginner programmers, due to a lack of funding and access to these beginners, this small sample space was all of the volunteers for the study. Each of the participants received all of the pieces of code along with their corresponding error messages from ReCodeZ and PyCharm. However, the order of their list of the pieces was randomized, resulting in that any specific ordering would largely not affect the result. There were 81 total votes for PyCharm error messages and 119 total votes for ReCodeZ error messages. This results in a significant difference of 38 votes. Of the 20 samples, 12 had more people vote for ReCodeZ than for PyCharm error messages, with four being tied and four with the opposite.

The experiment results show that ReCodeZ probably creates better error messages than PyCharm for those 20 samples of code. With 16 of them having at least as good if not better error messages in ReCodeZ, it is likely that ReCodeZ is better for beginner programmers, in terms of errors. We did try to create representative code samples, many taken from common online questions, so the results may be applied to a more general population of errors often experienced by beginner programmers. Our research can also be generalized to beginner programmers similar to those in our study, with less than a year of experience.

5. RELATED WORK

PyLint is a Python helper tool that analyzes code [8]. The main features of PyLint include improving coding quality and error detection. This means that our work and PyLint are relatively similar. PyLint is robust and can handle most errors. However, ReCodeZ is a programming environment, while PyLint is a tool that can only be used on existing code. PyLint cannot be used to create new programs. Additionally, ReCodeZ is more convenient as it is online, while PyLint must be downloaded. However, a strength of PyLint is that it can improve the quality of code. It can format the code to fit programming standards.

Another related work is a StackOverflow integration called Seahawk [9]. StackOverflow is an online resource that allows programmers to submit questions and get answers from other programmers. Seahawk imports this into a Java programming environment. This combines the power of the online resource and the programming environment, making it useful for getting specific answers for beginners. However, it is somewhat not as convenient as ReCodeZ as it is not online. Additionally, Seahawk takes some time to receive answers, as it is based on having other people respond to a user's question. ReCodeZ does not have this issue as all of the error messages are preloaded.

Repl.it, an online programming environment, which allows users to program in many different programming languages [10]. Both ReCodeZ and Repl.it are online environments, making programming more convenient. However, Repl.it has the ability to save code and share files online, which ReCodeZ is currently incapable of doing. However, it does not change error messages like ReCodeZ. ReCodeZ is slightly more convenient as the error messages better explain the errors from a user's code. Users of Repl.it may still have to use another online resource to understand errors, while ReCodeZ has error explanations on the same site.

6. CONCLUSION AND FUTURE WORK

In this project, we proposed a method to improve the experience of debugging for beginner programmers through the use of explanatory error messages. A website has been developed for users to submit Python code and see the output as well as any errors that their code receives. Instead of having error messages that have little information or possible misinformation,

ReCodeZ clarifies error messages and gives possible solutions to help the beginner to debug their code. To test our solution, we conducted a study of 10 beginner programmers. We gave them each 20 pieces of code, along with error messages from ReCodeZ and PyCharm, and asked them to vote on which one they thought had a better description. The results showed that ReCodeZ had more pieces of code with the majority vote and more votes than PyCharm. This means that ReCodeZ was able to give clear error messages that allowed the beginner programmer to understand how to debug their code. Regarding the aforementioned challenges, ReCodeZ solves all three. ReCodeZ is able to detect errors using the Python interpreter in conjunction with our server. ReCodeZ can describe the errors using our database of error handlers. Additionally, while fixing errors to make the program do what the user wants it to do is impossible without knowing what the user wants to happen, ReCodeZ gives possible solutions that may help the user to fix their error themselves.

The limitations of ReCodeZ lie in error detection. There are many different errors even for one programming language. Implementing all of the errors would take a large number of resources. Each error has to have an algorithm to detect it and a template to describe the error. The error may also require another algorithm to create possible solutions. Additionally, handling all of the errors in a chain will result in slow response times and lead to the programmer using a different resource. Our solution would lose its convenience if we added more error handling linked by a chain structure.

Another limitation is that many programming languages, including Python, allow users to create errors. ReCodeZ currently does not have the capability to fix these user-created errors as it does not know what the error was made to do. Similarly, as mentioned in Challenge 3, ReCodeZ cannot always predict the correct solution as there are often multiple solutions that each result in different outputs.

In future works, these limitations can be solved. A possible way to implement all of the errors is to use machine learning to generate text to describe errors. Another method is to use online resource integration, like Seahawk with StackOverflow, but with preset questions and answers. The chain structure of error handling can be replaced with a lookup table that can access errors that would be farther down the chain with higher efficiency. Additionally, default error messages can be added so that user-created errors can still be described. Finally, with machine learning, a solution that fits the user may be more likely to be predicted.

REFERENCES

[1] Dahl, Ole-Johan, Edsger Wybe Dijkstra, and Charles Antony Richard Hoare, eds. Structured programming. Academic Press Ltd., 1972.
[2] Baker, Catherine M., Lauren R. Milne, and Richard E. Ladner. "Structjumper: A tool to help blind programmers navigate and understand the structure of code." Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015.
[3] Where Developers Learn, Share, & Build Careers. stackoverflow.com/.
[4] Hartmann, Björn, et al. "What would other programmers do: suggesting solutions to error messages." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2010.
[5] Guzdial, Mark. "Programming environments for novices." Computer science education research 2004 (2004): 127-154.
[6] Ramadhan, Haider, and Benedict du Boulay. "Programming environments for novices." Cognitive models and intelligent environments for learning programming. Springer, Berlin, Heidelberg, 1993. 125-134.
[7] Lutz, Mark. Programming python. " O'Reilly Media, Inc.", 2001.
[8] Thenault, Sylvain. "Pylint—Code Analysis for Python." (2006).

[9] Ponzanelli, Luca, Alberto Bacchelli, and Michele Lanza. "Seahawk: Stack overflow in the ide." 2013 35th International Conference on Software Engineering (ICSE). IEEE, 2013.
[10] Repl.it - The collaborative browser based IDE. https://repl.it
[11] CodeProject - For those who code. https://www.codeproject.com/.
[12] Khaleel, Firas Layth, et al. "Programming Learning Requirements Based on Multi Perspectives." International Journal of Electrical and Computer Engineering 7.3 (2017): 1299.
[13] Butler, Matthew, and Michael Morgan. "Learning challenges faced by novice programming students studying high level and low feedback concepts." Proceedings ascilite Singapore 99-107 (2007).
[14] Brains, Jet. "PyCharm." URL https://www. jetbrains. com/pycharm (2018).
[15] Sullivan, Bryan. "Server-side JavaScript injection." Black Hat USA (2011).

DO8NOW: AN INTELLIGENT MOBILE PLATFORM FOR TIME MANAGEMENT USING SOCIAL COMPUTING AND MACHINE LEARNING

Ruichu (Eric) Xia[1], Yu Sun[2], Fangyan Zhang[3]

[1]Santa Margarita Catholic High School, Rancho Santa Margarita, CA 92688, USA
[2]California State Polytechnic University, Pomona, CA, 91768, USA
[3]ASML, San Jose, CA, 95131, USA

ABSTRACT

Many people today suffer from the negative effects of procrastination and poor time management which includes lower productivity, missing opportunities, lower self-esteem and increased levels of guilt, stress, frustration, and anxiety. Although people can often recognize their tendency to procrastinate and the need to change this bad habit, the majority of them still do not take meaningful actions to prevent themselves from procrastinating. To help people fix this problem, we created a goal tracking mobile application called iProgress that aims to assist and motivate people to better manage their time by allowing them to create short-term and long-term goals that they want to achieve, and encouraging them to complete those goals through a rank/reward system that provides them with the opportunity to compete with other users by completing more goals.

KEYWORDS

Procrastination, iProgreass, flutter, iOS, Android

1. INTRODUCTION

Procrastination is the intentional avoidance or delay of completing a certain task that is often stressful and unpleasant but also important. Although being a common experience for most people and not a mental health diagnosis by itself, procrastination can cause many negative emotions such as guilt, frustration, and anxiety, and is associated with some serious mental disorders such as attention-deficit hyperactivity disorder (ADHD), depression, and anxiety disorders according to various psychological research studies [1] [2][3]. In one of the first studies conducted on the effects of procrastination published in 1997, the researchers at Case Western Reserve University found that among the participants, students from the university, those who procrastinated generally displayed poorer performance at school, earned lower grades, and reported higher levels of stress than those who did not procrastinate [13]. Procrastination is also very prevalent today as shown in another study conducted in 2004, from which about 70% of university students reported that they would consider themselves to be procrastinators. Interestingly, the tendency to procrastinate is not only limited to humans. According to one study conducted on the behavioral patterns of pigeons, the researchers have found pigeons also display clear evidence of procrastination [4]. While most people may not experience some of the most

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 165-173, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101217

severe impacts of procrastinating and may sometimes even benefit from it as procrastination can act as a coping mechanism for stressful situations and mood regulations, chronic procrastinators can have their lives persistently disrupted and their success severely limited due to procrastination. To individuals that have the issue of chronic procrastination, the effect may include poor performance at school or work, financial difficulties, increasing levels of stress and anxiety, and poor physical health resulted from lack of exercise and nutrition.

In order to address the problem of procrastination, researchers have been trying to determine the causes of such behavior. Although earlier research has attributed procrastination to an individual's lack of self-regulation and time management skills, recent studies suggest another factor at play that is more important than the other factors: the emotional component of procrastination [5][6]. Studies have shown that people choose to delay or entirely avoid doing certain tasks despite knowing full well the deadline and the consequences of not completing the tasks because they tend to seek the immediate emotional improvements associated with delaying the task at hand. So instead of completing the task way ahead of the deadline, people who tend to procrastinate would choose to put it off for some time because they have the false belief that they will be better and more emotionally equipped at finishing the task in the future. Therefore, along with the lack of self-regulation, the immediate pleasure-seeking temptation causes people to procrastinate. There are some useful intervention methods that reduce procrastination [11][12]. This includes setting up a schedule for the deadlines, breaking down the goal into smaller tasks, and practice to be mindful of certain negative feelings associated with doing the task. Some of these techniques are certainly very helpful to people who tend to procrastinate. For example, setting up a timer for when everything needs to be done is a simple but also very effective way of keeping people on track. However, most of these techniques by themselves cannot actually effectively prevent people from procrastinating. While a simple timer might help a person be aware of what needs to be done before a certain time, the emotional aspect of the problem remains unanswered. Breaking down tasks into smaller pieces is certainly a good technique, but without any external demands and requirements of the procrastinator, it would still be difficult for that person to commit to finishing the entire task without being distracted by temporary pleasure temptations.

In this project, we aim to address the problem of procrastination by combining all the useful strategies in dealing with procrastination into a mobile application called iProgress. Although iProgress is also a habit/goal tracking app that is common on the app market, it has several unique features that can effectively reduce people's tendency to procrastinate, making it overall a better productivity tool than most of the other existing apps. First, an entirely customizable schedule is put in place for the users to set and keep track of their daily tasks and long-term goals. Second, by allowing users to consciously point out the benefit of not putting off a certain goal, the app targets the user's emotions and reward seeking behaviour, encouraging them to avoid temporary distractions. And lastly, inspired by various rank systems in video games, we added a similar rank system to the app that will further motivate its users to complete more goals on time as they compete with each other by completing more goals without procrastinating. Therefore, we believe that with the efficient methods incorporated into the app, iProgress can help people to effectively deal with the problem of procrastination and form a better and more productive lifestyle by developing good work ethics and more healthy habits.

To prove our results, we recruited 20 participants who are interested in using the app and conducted a close interview with each of the participants both before and after 30 days of using the app. Through the interview process, we were able to compare the participant's initial experience in procrastination with their experience after using the app for some time. Then we analyzed the changes in their experiences and determined whether or not the app has an overall positive effect on the participants in helping them avoid procrastination. Moreover, we also

collected evaluation of the app from the participants as well as any suggestions for potential changes for the app in the future.

The rest of the paper is organized as follows: Section 2 displays the details of the challenges we had in designing the app and implementing the methods and functions; Section 3 gives the details of our solutions regarding each challenge we faced as mentioned in Section 2; Following that, Section 4 presents the relevant details about the user experiences and evaluations; Section 5 lists out some of the related applications on the market, followed by Section 6 which provides the conclusion remarks as well as an overview on the potential future improvements of the application.

2. CHALLENGES

Building a mobile application for the first time with a new programming language is not an easy task. Throughout the course of the development, we ran into several challenges that needed overcoming. Here is a brief overview of some of the most difficult challenges that we faced when developing this app.

2.1. Challenge 1: Coming up with a Solution that Addresses All Aspects of the Problem is Not an Easy Task.

After doing some research on the issue of procrastination, we noticed that there are many different aspects of the problem that needed to be taken into consideration when we were designing the app. We found that there is no one single method that can effectively address the problem of procrastination. For example, a schedule system that allows users to set and complete goals is not sufficient as the only source of motivation to keep people on track. Therefore, coming up with a comprehensive solution that addresses both the time management aspect and the emotional aspect of procrastination was critical to the effectiveness and the success of the app. Moreover, the details of the rank system also needed to be carefully designed as we wanted to ensure that the system is both fair and rewarding.

2.2. Challenge 2: Implementing the Design Ideas can be Hard Because the Rank System Requires a Database Server.

Unlike other goal tracking apps, our design of the app includes a rank system that requires us to have a functional database server to store all the necessary information of the users including points, completed goals, rank position, and username, etc. Without a server, the app would not offer the opportunity for the users to compete with each other, which is an important motivation factor in reducing the tendency of procrastination. Besides implementing a functional server, other design ideas such as separating long-term and short-term goals, implementing a remind/notification system, and making the motivational statement associated with each goal set by the user useful in targeting the emotional aspect of procrastination were all challenges that needed us to address throughout the course of the development.

2.3. Challenge 3: Building a Cool User Interface and Adding New Functions and Modules are Necessary in Attracting More Users.

After finishing implementing the basic functions of the app, we began to focus all our attention on building a beautiful user interface with smooth transitions from modules to modules. We knew that having an UI that is both easy to view and easy to understand is crucial in attracting new users as well as keeping the current user base because no one wants to use an app every day that

is hard to look at and way too complicated to understand. However, connecting different modules and building a cool user interface with flutter for the first time was not easy. Since every page's design had to be written in code, building a good UI required us to learn the new language more thoroughly before knowing how to format each page, connecting them together, and making sure that the app runs smoothly on the user device.

3. SOLUTION

The goal of our application is to help the users reduce their tendency to procrastinate. An overview of the app is shown in the diagram below. To use the app, the user is first asked to create an account by providing a username, email and password. Once an account has been created, a new user id will be generated and all the user's information will be initialized in the firebase server, which is the server that we use to store the relevant user information. Then the user can login to his/her account and gain access to the main functions of the app. As seen from the diagram, the app consists of three main functions.

3.1. Overview of iProgress System

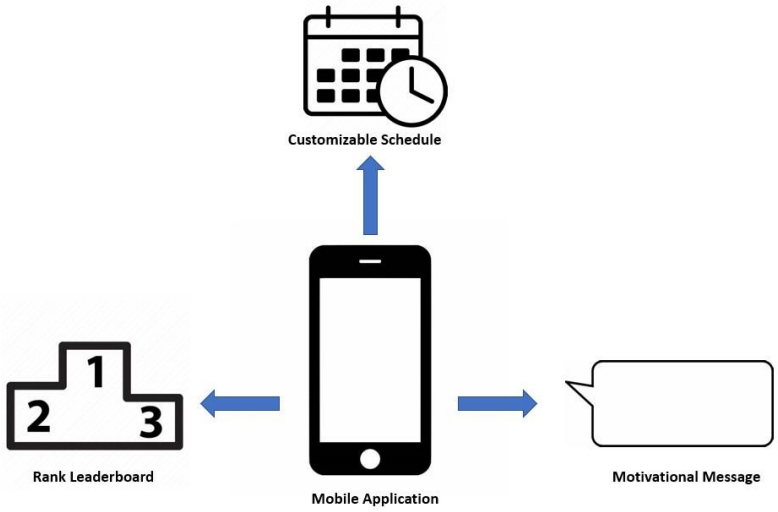


Figure 1: Overview of iProgress system

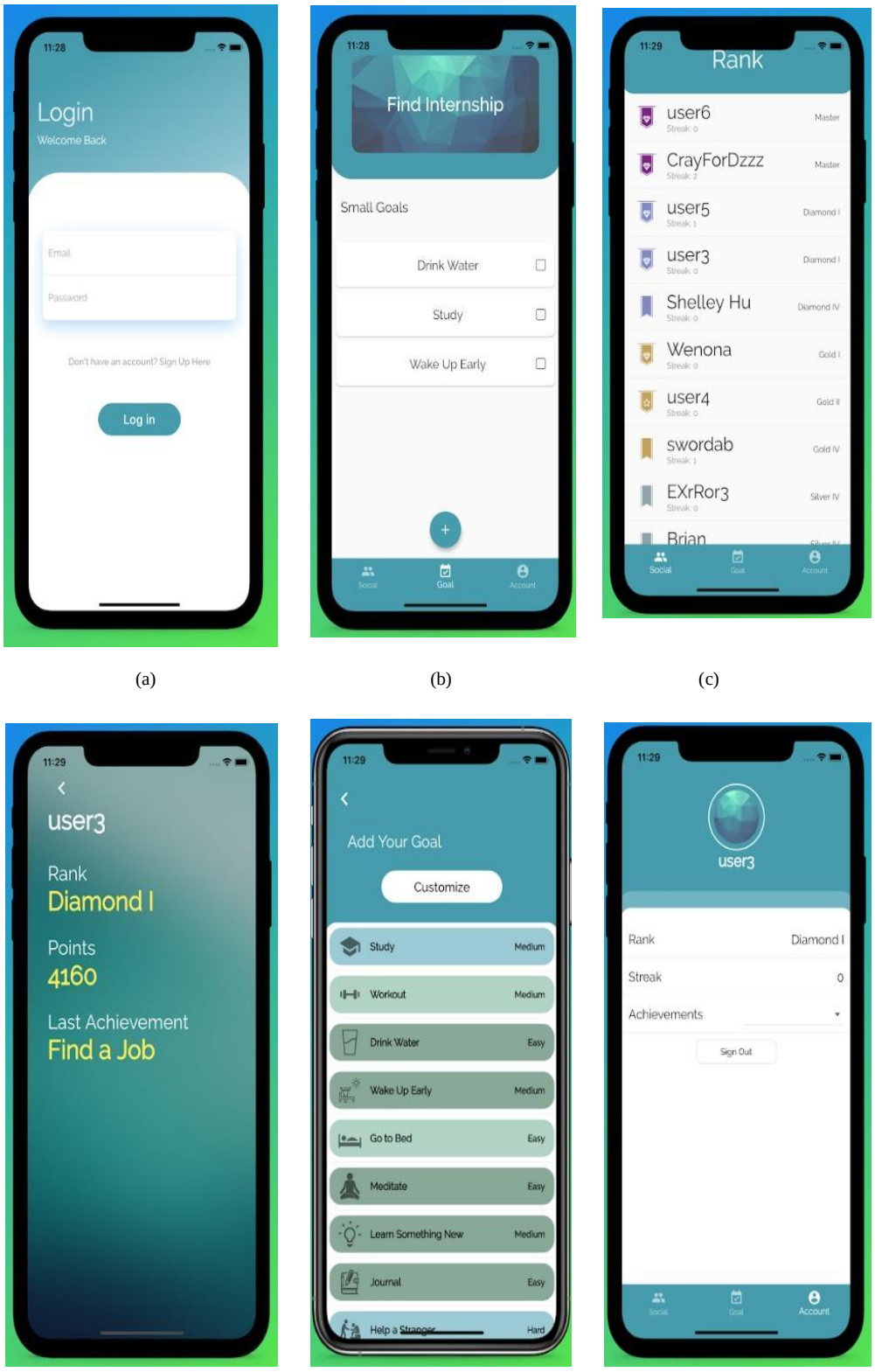
It allows the user to create a list of small goals and one big goal and helps them to keep track of those goals. This function aims to solve one aspect of procrastination, which is one's inability to efficiently manage schedule and time. The goals are clearly displayed on the homepage of the app with a big goal on top and a list of small goals at the bottom. A check box can be found next to each small task that the user has set to complete. The user has the ability to customize the schedule by adding, removing, and completing goals. Once a small goal/task is completed, the check box will automatically be checked that tells the user that the task has already been completed. When adding new small goals or changing the big goal, the user has the option to either choose from a list of predefined goals on the add goal page or customize a new goal by pressing the customize button on the top of the page. To customize a goal, the user can simply input the name and the perceived difficulty of the goal. After clicking confirm, the user will then be asked to set a reminder time for the goal and also provide a brief statement to why it is necessary to complete the goal on time. Then hit confirm again, the user will return to the homepage and the goal will be changed or added to the list.

When entering each goal, the user is asked to offer a reason why he/she must complete that goal on time. This message is then displayed as a notification on the user's device at the specific reminder time that the user has set for the goal. By requiring the user to consciously give a reason to not procrastinate and reminding the user of that reason, the app aims to address the emotional aspect of procrastination as the user is reminded that he/she will ultimately feel better if the goal is completed on time.

The rank system is the last major function of the app. To the left of the homepage is the social page where the rankings of all users of the app are located. Inspired by common rank systems in video games, the rank system of this app consists of six different divisions ranging from bronze to master and each division has four tiers. Ranks are represented by different rank badges with different colors that are located on the left side of each user. In order to earn points, the user must complete goals on time without procrastinating. Points are rewarded based on the difficulty of the goal and more points are rewarded for long-term goals than small goals. We hope that with the incorporation of a rank system, the users will have more incentives to keep themselves on track of their work and schedule and avoid being distracted by things that will delay the completion of the goals.

3.2. The Implementation of iProgress System

To implement the app, we used flutter, an open-source UI software development kit made by Google [6]. Since flutter allows developers to write apps for both IOS and Android using the same language and code, we were able to write everything once in Android Studio and successfully published the app to both iTunes and Google Play store. For the server we used a firebase real-time database, which is a cloud database that stores and syncs data across all clients in real time. Using the firebase database ultimately helped us overcome the challenge of implementing more complicated functions into the app such as the rank system that required a cloud server that can store and distribute information across different devices [7]. Another challenge was building the user interface after we finished implementing the basic functions of the app. To make sure that we have a clean and smooth UI for the app, we built widgets and put them together using flutter material design, which has a bunch of tools for app developers to decorate their apps. The adaptability of the UI also creates a consistent user experience not only across different platforms, but also different devices. Moreover, because flutter supports both IOS and Android, it is very easy for us to make changes to the code and update the app in stores, which means that we can constantly add new functions and make improvements to the app. There were several new features that were added to the app after the initial publication, including an achievement tab that allows the user to see all of his past completed big goals, a streak system that keeps track of how many days in a row a person has been completing goals, and an user details page that displays the points, streak, and the last achievement of the user.



(a) (b) (c)

(d) (e) (f)

Figure 2: Interface of iProgress system (a) Log in (b) set a target (c) overall ranks (d) user details (e) summary of a user's goal (f) user rank

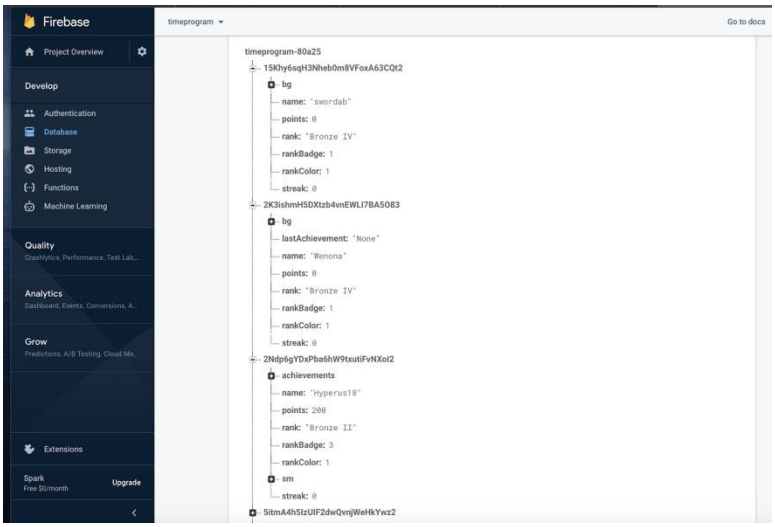


Figure 3: Firebase of iProgress system

4. EXPERIMENT

To evaluate the effectiveness of our approach in helping people to deal with procrastination, we collected user evaluation and feedback from 20 participants who used the app for 30 days in the form of close interviews both before and after they began using the app.

In order to determine if the app did actually help reduce people's tendency to procrastinate, the first thing that we did was to compare the initial experience of the participants in dealing with procrastination with their experience of procrastination after using the app. The participants were mainly high schoolers who have all experienced procrastination at some point in their high school career. This is shown clearly during the initial interview as a majority of the participants reported to have a tendency to procrastinate when doing assignments for school or studying for exams. However, when we compared the participants' post-experience with procrastination to their initial experience, the result was overall very positive: out of the 20 participants, 16 of them reported that they are less likely to procrastinate when they have entered the task to the small goal list in the app, while 4 of them reported no changes to their experience with procrastination, which suggests that a majority of the user base tend to benefit from app in terms of reducing procrastination.

We then collected evaluations from the participants by asking them to list out the things that they liked and/or disliked about the app. Some things that the participants liked about the app were features such as the rank system, the clear display of small goals along with the notification system, and the message that you can write to yourself when entering each goal. Some participants reported that the rank system really motivated them to finish the assignments that they set off to do since they could all see each other on the leaderboard and wanted to have a higher rank than the other participants. The clean display of small goals was also a favorable feature as many of the participants reported that they could easily create and manage a schedule using the app. Some things that people disliked about the app were features like the restriction of having only one big goal at a time and the predefined goals being too limited. Some of the participants reported that the app only allows them to create one big goal at a time, which really limits their aspiration to complete more goals. Some also complained that the list of predefined goals were far too limited and suggested more variations of small goals should be added to the

list. However, the dissatisfaction with other functions of the app was still very minimal, suggesting that the app was overall a very effective tool to help reduce procrastination.

Lastly, we asked the participants to give any feedback or suggestions for the app in terms of improvements of certain parts and the implementation of other features in the future. One of the most important feedback that we received was that the social feature of the app can be further expanded to include other forms of social interactions between the users. For example, some of the participants felt that it would be more exciting if they could make friends with each other and check on each other's progress by chatting with them while using the app. Moreover, some people also suggested that we should implement a calendar system as a better form of visualization for the user's schedule. We believe that all of these suggestions are extremely valuable to the future improvements of the app, and we are currently planning on implementing these features while addressing the complaints by making changes to some of the existing parts of the app.

5. RELATED WORK

Forest is a popular productivity app made by SEEKRTECH CO., LTD. that helps people manage their time and stay focused on the tasks at hand by letting them plant trees in the app when they need to complete a task for a certain time without being distracted by the phone [10]. Similar to iProgress, Forest also has a reward system in which users can spend the coins they obtained to unlock new types of trees and bushes, or even plant real trees on Earth. This is an incredibly interesting feature as it creates a sense of responsibility and achievement for the users as they can protect the environment while becoming more productive. However, unlike iProgress, Forest does not have a rank feature or a customizable schedule.

Another related app is Momentum Habit Tracker developed by Mathias Maehlum [9]. Momentum is an ideal app for keeping track of habits and routines as it offers a calendar feature that allows users to create weekly targets and take notes on their progress, making it very easy to build new habits. Although the app is similar to iProgress in that they both let users create goals and keep track of them, Momentum is more like a calendar that helps people with developing new habits, while iProgress focuses more on addressing the problem of procrastination by providing incentive.

Habitica is another interesting app that helps people track and maintain good habits [8]. Similar to iProgress, Habitica also has a creative way to give people extra incentive to stay motivated. Inspired by RPG video games, the users of Habitica can rank up their characters as they complete more tasks, which makes habit building more fun and exciting. Compared to Habitica, iProgress does not have a complex RPG game reward system, but it does have a more goal oriented system that targets the emotions of the users, making it more effective in helping people beat procrastination.

6. CONCLUSION AND FUTURE WORK

To summarize, procrastination is a common issue that many people face today, and it can often have many negative effects on a person such as increased guilt and frustration, poor performance at work or school, and lower productivity. While most people do not experience the most severe consequences of procrastination, this issue is still prevalent as it can often limit the success of those who do procrastinate. Therefore, in this project, our goal is to address the problem of procrastination by creating a mobile application that combines all the tools and strategies that are effective dealing with this issue. With unique features such as a rank system that will motivate

users to complete more goals without procrastinating by creating a competitive environment, and a customizable message attached to each goal that allows the users to consciously note the importance of avoiding procrastination, iProgress can help people to effectively deal with the problem of procrastination and form a better and more productive lifestyle by developing good work ethics and more healthy habits. Through the evaluations and feedback from some of the users, we were able to determine that the app has an overall positive effect on people who struggle with procrastination. We found that the majority of people who use the app tend to benefit from it in terms of reducing their tendency to procrastinate. As for the app's limitation, there are definitely a few changes that need to be made in order to improve the overall user experience. One of the limitations is that iProgress's social feature is not very comprehensive as it does not allow users to interact with each other in the forms of chatting or making friends. However, we are currently planning on expanding the social feature in the near future. The first step would be to add a message system where the users can leave each other messages as reminders of the goals that they need to complete. Other changes and features that we heard from user suggestions are also going to be implemented. For example, we will add a calendar page to the app so that the user can create and manage schedules more easily. We will also expand the restriction on the number of big goals so that people can have multiple big goals at a time. In the meantime, we will continue to collect feedback and suggestions from users and make changes to the app so that more people will benefit from it by being more productive and freer of procrastination.

REFERENCES

[1] Klingsieck, Katrin B. "Procrastination." European Psychologist (2013).

[2] Solomon, Laura J., and Esther D. Rothblum. "Academic procrastination: Frequency and cognitive-behavioral correlates." Journal of counseling psychology 31.4 (1984): 503.

[3] Lay, Clarry H. "At last, my research article on procrastination." Journal of research in personality 20.4 (1986): 474-495.

[4] Mazur, James E. "Procrastination by pigeons: Preference for larger, more delayed work requirements." Journal of the Experimental Analysis of Behavior 65.1 (1996): 159-171.

[5] Lay, Clarry H., and Henri C. Schouwenburg. "Trait Procrastination, Time Management." Journal of social Behavior and personality 8.4 (1993): 647-662.

[6] Kuzmin, Nikita, Konstantin Ignatiev, and Denis Grafov. "Experience of Developing a Mobile Application Using Flutter." Information Science and Applications. Springer, Singapore, 2020. 571-575.

[7] Moroney, Laurence. "The firebase realtime database." The Definitive Guide to Firebase. Apress, Berkeley, CA, 2017. 51-71.

[8] Habitica: Gamified Taskmanager. HabitRPG, Inc,2020. Vers.2.7. Google Play Store. https://play.google.com/store/apps/details?id=com.habitrpg.android.habitica&hl=en_US

[9] Mathias Maehlum. Momentum Habit Tracker - Routines, Goals & Rituals. Momentum.cc, 2020. Vers. 3.5. Apple App Store. https://apps.apple.com/us/app/momentum-habit-tracker-routines-goalsrituals/id946923599

[10] Forest: Stay focused. SEEKRTECH CO., LTD,2020. Vers.4.20.0. Google Play Store. https://play.google.com/store/apps/details?id=cc.forestapp&hl=en_US

[11] Paden, Nita, and Roxanne Stell. "Reducing procrastination through assignment and course design." Marketing Education Review 7.2 (1997): 17-25.

[12] Lamwers, Linda L., and Christine H. Jazwinski. "A comparison of three strategies to reduce student procrastination in PSI." Teaching of Psychology 16.1 (1989): 8-12.

[13] Tice, Dianne M., and Roy F. Baumeister. "Longitudinal study of procrastination, performance, stress, and health: The costs and benefits of dawdling." Psychological science 8.6 (1997): 454-458.

AUTOMATION OF POLITICAL BIASES DETECTION USING MACHINE LEARNING

Bill Zheng

Webb School of California, USA

ABSTRACT

In the current political climate, mass media was depicted as highly divisive and inaccurate while many cannot efficiently identify its bias presented in the news. Using research regarding keywords in the current political environment, we have designed an algorithm that detects and quantifies political, opinion, and satirical biases present in current day articles. Our algorithm makes use of scipy'ssklearn linear regression model and multiple regression model to automatically identify the bias of a news article based on a scale of 0 to 3 (-3 to 3 in political bias detection) to automatically detect the bias presented in a news source. The usage of this algorithm on all three segments, politics, opinion, and satire has been proven effective, and it enables an average reader to accurately evaluate the bias in a news source.

1. INTRODUCTION

Media-processed news has become the predominant means in which Americans consume current information. Despite its presence in almost every facet of American life, many Americans today have expressed their discontent with the contemporary news service. With more than 62% of the Americans expressing that the mainstream media is biased and more than 80% of the Americans expressing that their news sources are biased, many set out to combat this polarization of news media in politics. Biases within a news source can be extremely harmful to the audience as misformation can produce harm at both a physical and a mental level. Millions of dollars have been spent on keeping the news impartial or detect any possible bias within the news, and this topic became more significant as we entered the present due to the polarization of news resources. An automated bias detection algorithm can not only save time for an average reader to identify the underlying biases of a news source, but it can also benefit the public by giving them more clarity in acquiring information and making rational decisions in the future.

While the issue of bias within the mainstream media was pervasive, many of the times readers were asked to identify the biases themselves as the media does not provide any insights on its own biases. Sometimes third-party fact-checkers such as Associated Press will conduct fact-checking by themselves, but that method was limited in its usage. One study done by Budak, Goel, and Rao [1] indicates that the bias of each news source can be classified as an n-dimensional vector and the computer can learn the bias of the vector by checking the attitude of the article on a certain topic. However, this study did not conclusively determine how an individual article presents its bias, instead, it gives us a holistic perspective on what biased is a news source. In the study, each of the data points gathered by the researchers was used to calculate the overall bias of a company instead of the bias of a given article, and that could be proven ineffective. Another practical problem is that this application does not provide a causal or correlational relationship between individual articles and its publishers and its context, which, if not used accurately, could produce misinformation and cannot be more complex than just indicating the general direction of the research.

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 175-181, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101218

In this paper, we follow the same line of research by Hamborg, Donnay, and Gipp [2]. Our goal is to use the semantics given in the news article to formulate the bias of each article. Our method is inspired by Hamborg, Donnay, and Gipp as they introduced both the source of the news and the semantics of the news together. First, we analyzed the source of the news through the means of identifying each keyword -- that is both a topic revolving around politics and adjectives in a specific manner in which positively and negatively describing things in one manner would result in a specific result from the multiplier. Using linear regression, one would get the results of bias/opinion bias/satire bias in the form of the slope of the best-fit regression, which will then be compared with the consensus to adjust the final value. This training will be iterated through many samples, and this ensures the accuracy of detecting biases that are subtle or satirical contents. Both the source of the news and the semantics will be combined in this method, which will also provide a holistic perspective on the biases presented in the news.

We have tested three distinct scenarios using a total of 8 different articles from 8 distinct news sources that are known to have a certain bias, opinion, and satirical tone. We dissect them into two segments with one of them being less opinionated/ biased/ satirical tone than the other. We have done this method two times, one with 100 pieces of 15-word segments and one with 200 pieces of 15-word segments in training. Both of these experiments have shown varying degrees of success in detecting various forms of biases in news reports compared to the consensus. After training with 200 pieces of 15-word segments, our accuracy in terms of convergence of a specific piece of text has improved by 15% in evaluating the total range of the articles compared to the consensus.

Section 2 presents the challenges met in the processing of experimenting with the algorithm and ways of circumventing around it. Section 3 focuses on the solution and walks through its process of creating it while corresponding to each section with section 2. Section 4 presents details, results, and analysis of the experiment in section three, which is followed by presenting works that helped the creation of this paper in section 5. The section concludes the paper while giving future insights into similar projects.

2. CHALLENGES

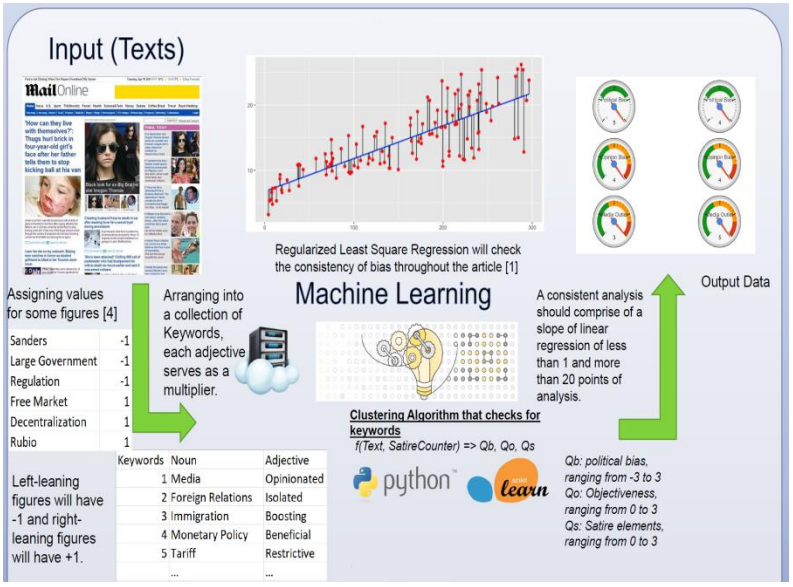
The biggest challenge of working on this project entails the inclusion of internal biases. The politics involved within the media culture is tremendous, and it is completely subjective and it is up to the interpretation of the audience. One of the more significant examples is the difference between the worldview between European politics and American politics. In general, Americans are more right-leaning than Europeans given the same ideological tenets of a politician. It should also be noted that defining biases is purely subjective as a right-winger is more tolerant towards right-wing content than a left-winger. Therefore, it is hard to find the consensus and to check the accuracy of a given article based on the information produced.

The second challenge regards dealing with the erratic nature of news reporting. Given that there are multiple news reporters and journalists with a wide range of ideological beliefs, it is hard to keep a track of everything that they are reporting. For example, the Wall Street Journal has a wide variety of reporters and journalists with some leaning to the right and some leaning to the left, and it has a high discrepancy within the own firm. Therefore, one of the major challenges will be to classify the distribution of the biases within a single firm and to make it as accurate as possible.

The final challenge regarding this paper is finding the appropriate sources. While there were some papers automating the process of political bias detection, they are mostly limited to finding

the biases of a certain keyword of a certain news source. It should also be noted that not many would attempt to write a thesis based on the two challenges presented above. Therefore, it would be harder than expected to find any credible resources to prove the effectiveness of this thesis and its experiment.

3. METHODOLOGY/SOLUTION



The system will first break the word segment into smaller samples based on the keyword dictionary provided. It will then use the dictionary to identify a numerical value for all of the word samples by giving a left-wing word a negative value and a right-wing word a positive value to correspond with their respective positions on a cartesian plane. It will then find any adjectives associated with such keywords, and it will multiply it with the adjective given with a positive value for a positive adjective and a negative value for a negative adjective. Using that, it will use regularized least square regression to check the apparent biases of the news by inserting the keywords into their original position as indicated by the numerical value, then it will compare itself with the sample already learned by the computer that has undergone the same procedure and has been given a numerical value for its biases to ensure its accuracy.

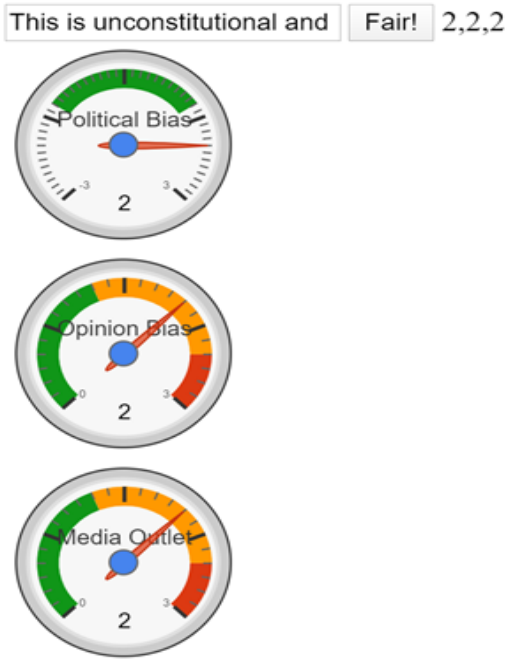
```
input_Data = []
for news in input_files:
  with open (news, 'r') as file:
    input_Data.append(file.read().replace('\n', ''))
cv = CountVectorizer()
input_ = cv.fit_transform(input_files).toarray()
output_ = [3, -3, 0, -1, 2, 2, -1, 1]
model = svm.SVC()
model.fit(input_, output_)

filename = 'model_source_bias.sav'
pickle.dump(model, open(filename, 'wb'))
pickle.dump(cv, open("cv_model_vectorizer.pickle_bias", "wb"))
```

```
@app.route("/politicalbias/<news_text>")
def make_bias_prediction(news_text):
  test = [news_text]
  loaded_model = pickle.load(open('model_source_bias.sav', 'rb'))
  cv = pickle.load(open("cv_model_vectorizer.pickle_bias", "rb"))

  politicalbiasresult = loaded_model.predict(cv.transform(test)
  .toarray())
  print(politicalbiasresult)
  return (str(politicalbiasresult[0]))
```

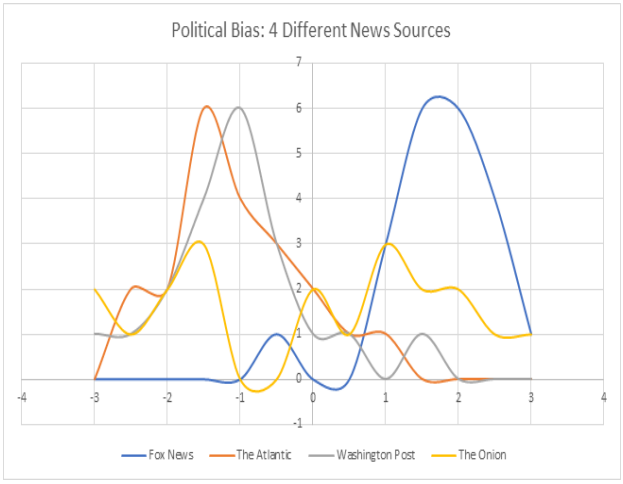
The code follows the same patterns as the steps discussed above: the computer will first detect relevant keywords and use a method of least regularized square regression to calculate the theoretical biases as described above, then it will match with the texts that have a similar set of keywords and compare their values. Data will be saved through pickle, and the real value will be sent out through methods of count vectorizer transforms. The final results will be printed and displayed in the dials below.



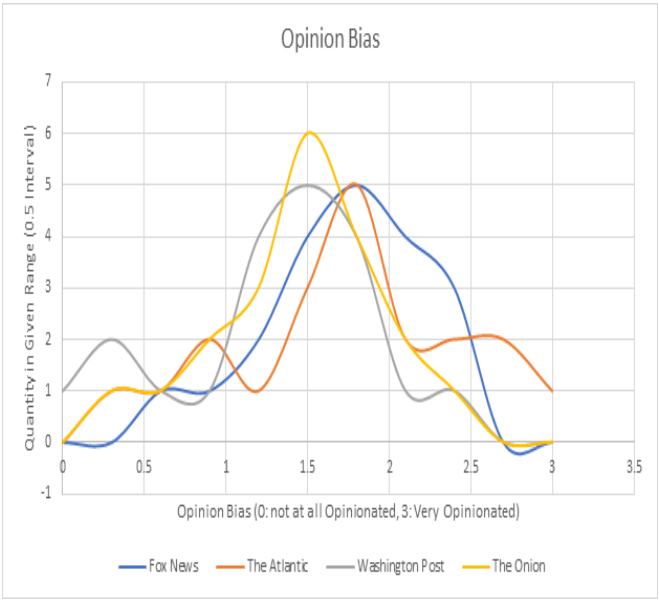
The first part of the algorithm attempts to quantify the biases using the linear regression model, then the linear regression will give an output and correlate it to the given value in the list. They will be fitted together and used in machine learning through a count vectorizer.

4. EXPERIMENTS/EVALUATION

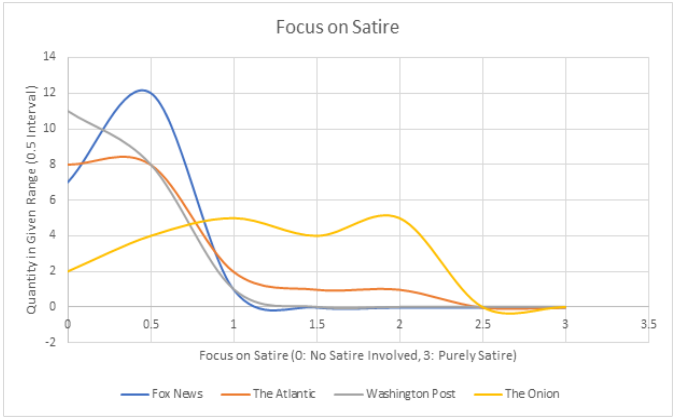
My solution strives to automate the detection of political bias within a piece of text, and the algorithm has guaranteed that it will generate an accurate result. My experiment includes using eight different articles with varying degrees of political, opinion, and satirical biases from 4 different sources. It totals up to around 2000 words, which will be fitting for the learning process.



Numerical Average:
Fox News: 1.92
The Atlantic: -1.52
Washington Post: -0.98
The Onion: 0.22

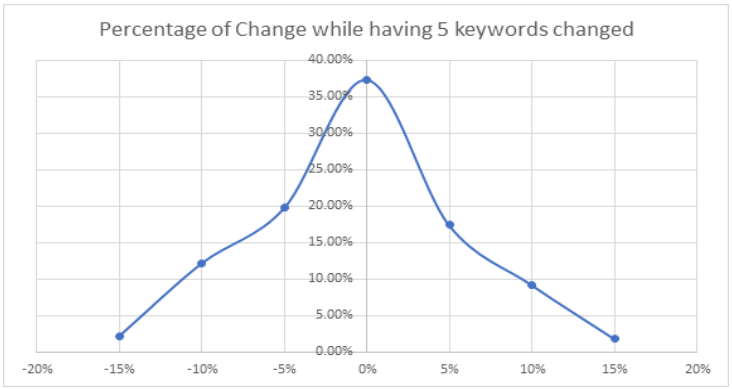


Numerical Average:
Fox News: 2.02
The Atlantic: 1.88
Washington Post: 1.51
The Onion: 1.44



Numerical Average:
Fox News: 0.42
The Atlantic: 0.88
Washington Post: 0.30
The Onion: 1.88

The second experiment is designed to test the algorithm's susceptibility to change, and more specifically the susceptibility of change in keywords. In each of the 30 words segment, we had 5 keywords change throughout the entire passage randomly, and we tested the susceptibility to keyword change without changing the general meaning of the article.



Compared to the general consensus, my results have shown that the twenty pieces gathered at random have shown great accuracy with each of them having less than 15% error. As of the changing keywords, despite it causing some major change in the results of the algorithm, calculations have shown that the standard deviation is around 4%, which shows that this algorithm while facing some changes in the keywords section, are in fact less susceptible to the changes in them.

5. Related Work & References

Related work 1

https://www8.gsb.columbia.edu/media/sites/media/files/JustinRaoMediaBias.pdf
This work is used to identify political biases within each news source by analyzing the semantics of each news source. My work expanded on this work by using the semantics and given values of

a text to fit the biases of each text instead of each news source. I improved the ways this study went as I used a more holistic way to approach this issue and provided a more stable and accurate result.

Related work 2

https://link.springer.com/article/10.1007/s00799-018-0261-y
This source discussed the potential role that the news publisher could have on emanating biases, and it also discussed how to use machine learning to detect biases by giving each way of using different semantics a fixed value. I believe this is a great foundation for my work as it also provided much of the materials needed for constructing my machine learning algorithm. My algorithm focuses less on the source, and it is more efficient as it simply combines two algorithms which have been laid foundations on.

Related work 3

Hainmueller, J., & Hazlett, C. (2014). Kernel Regularized Least Squares:
Reducing Misspecification Bias with a Flexible and Interpretable
Machine Learning Approach.
Political Analysis, 22 (2), 143 168.
doi:10.1093/pan/mpt019

Compared to the two algorithms discussed above, the algorithm introduced in this study focuses on a specific way of regression: Kernel regularized least squares. It helped me to provide a more accurate reading on the regression, and I used his work as a part of my algorithm. However, with the combined algorithm, I can create more applications for this mean of regression and put it into great use with great accuracy.

6. CONCLUSION AND FUTURE WORK (400+ WORDS)

In this paper, we develop a system which uses the semantics given in the news article to detect and quantify political, opinion, and satirical biases present in current day articles. Firstly, the system will first break the word segment into smaller samples based on the keyword dictionary provided. Then, it will identify a numerical value for all of the word samples by giving a left- wing word a negative value and a right-wing word a positive value to correspond with their respective positions. The results of bias/opinion bias/satire bias will be obtained in the form of the slope of the best-fit regression through linear regression. For the data training, it will be iterated through many samples, and this ensures the accuracy of detecting biases that are subtle or satirical contents. Both the source of the news and the semantics will be combined in this method, which will also provide a holistic perspective on the biases presented in the news.

In future, we will improve our model and make sure that the bias/opinion bias/satire bias can reflect the real case precisely.

FITABLE: A FREE CONVENIENT SOLUTION TO YOUR HEALTH GOALS

Wesley Fan[1], Eric Wasserman[1], Eiffel Vuong[1], Dylan Lazar[1], Matthew Haase[1], and Yu Sun[2]

[1]Portola High School, 1001 Cadence, Irvine, CA 92618
[2]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

In the recent decades, an increasing number of people become overweight, ranging from children to elders. Consequently, a series of diseases come along with obesity. How to control weight effectively is a big concern for most people. In order to improve the awareness of people's diets and calorie intake, this paper develops an application–Fitable, which can help users by calculating calories burned in a particular workout. The foods that Fitable recommends are all based on the lifestyle the user is aiming to achieve. Until now, the app is accessible to Android users.

KEYWORDS

Android, flutter, firebase, machine learning

1. INTRODUCTION

In 1995, only 28% of the adult population was considered obese [1]. In 2016, roughly two-thirds of adults and nearly 20% of children were overweight [2]. Today, approximately 160 million people are obese [3]. That means roughly one-in-forty people on Earth are overweight. Numerous people know that not exercising and genetics correlate to obesity, however, dieting is by far the most crucial aspect in maintaining a healthy lifestyle. Many people exercise and do not know what to eat to supplement their exercise routine. Some eat unhealthy foods after a heavy workout. After a few weeks, they do not see any improvement and thus give up, therefore, becoming overweight. Obesity tends to lead to Type 2 diabetes, high blood pressure, strokes, and other heart and health issues [4][5]. The Fitable team set out to reduce these issues by spreading awareness using a dieting app, thus creating Fitable. Today, people recommend playing sports, having a daily exercise routine, and counting their calorie intake. This app takes all of these into account and recommends healthy foods to encourage a healthy lifestyle. Our app combines the convenience of popular products, such as Fitbit and Apple watches with the nutritional benefits of calorie calculators to give users the best experience.

The remainder of this paper is organized as follows: in Section II, we provide our challenges for the app development, and in Section III, we present the app and our solution to these challenges. We introduce the app and describe how it works. In Section IV, we provide insight into future features for Fitable and conclude our paper with a summary.

Exercise and dieting are essential aspects of human life. Many people are suffering from obesity, sparking the development of Fitable. The popular products, Fitbit and Apple Watch, calculate the steps and the time spent on a workout. However, these products do not calculate the foods

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 183-190, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101219

people should eat to support their exercise, intensity, and the specific type of workout [6][7]. Fitable provides users with what they should be consuming to reach their goal (i.e., general, bulking, slimming) based on their workout and exercises. Fitable's competition, Fitbit and Apple Watch, are considered luxuries, and not everyone can afford these luxuries. However, Fitable is free, which increases the potential number of users. Technology has been incorporated into the world more rapidly, and thus, we wanted to create an app that uses technology to help others. There are few, if any, apps that calculate the calories burned in a particular workout, and they do not suit the needs of potential users as accurately as Fitable. Additionally, the foods recommended by some websites and apps are absurdly high in sodium, cholesterol, and carbohydrates. The foods that Fitable recommends are all based on the lifestyle the user is aiming to achieve, no matter the goal. Fitable does not recommend foods that are high in sodium, cholesterol, or carbohydrates to any of their users.

The following map is a graph of obesity [8]. This graph shows the obesity population as recent as 2017. Since then, this number has gone up and is at an all-time high in 2019 [9].

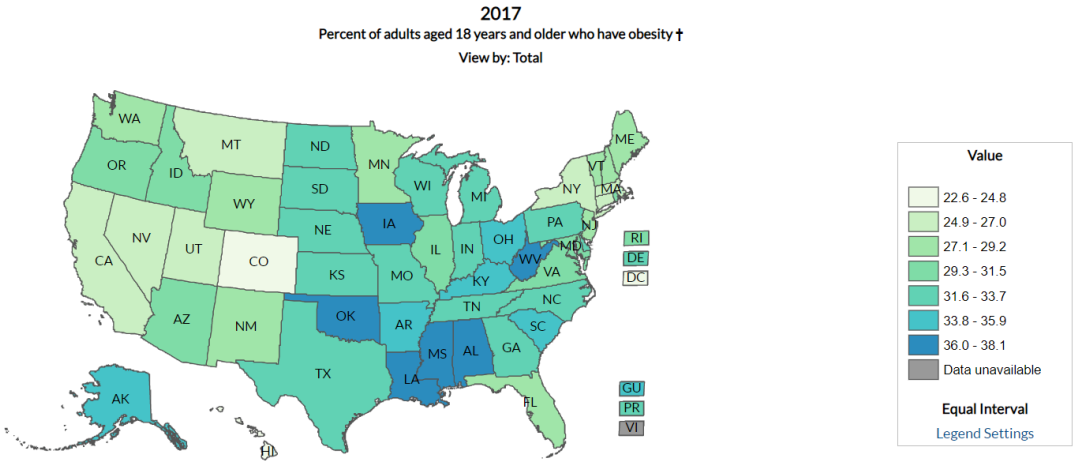


Table 1. Fitable vs. Competitors

Benefits	Apple Watch	Fitbit	Calorie Calculators [10]	Fitable
Recommended Calories Intake			✓	✓
Food Recommendation			✓	✓
Convenient	✓	✓		✓
Specific Exercises		✓		✓
Cheap			✓	✓
Private	✓	✓	✓	✓

In the customization of Fitable, the color scheme and aesthetics were an essential part of the process. We wanted a color scheme that would be suitable for all users, regardless of gender or race. The logo is an integral part of the development of any app, along with something that immediately catches the user's eye. It also has to relate to the primary purpose and intertwine with the color scheme; so, therefore, we designed a logo with a character running.

2. CHALLENGES

2.1. Challenge 1

Through the development of the app, we collected nutrition information on food items. Nutrition information is recorded per serving size, and serving size does not have universal metrics for all foods, which created the tedious job of determining the next most common metric. For example, the NLEA serving size -- or the amount of food that is generally consumed in a sitting -- for apples was present, whereas in some other foods (coconut, steak, etc.,) the NLEA serving size was absent, thus creating the challenge of determining the next most common metric [11]. This made the nutrition calculation, and therefore, food recommendation, very difficult.

2.2. Challenge 2

Calculating the number of calories burned depends on the time played and the intensity of the sport. Playing golf at an intensity of five for one hour would burn fewer calories than if you played soccer at an intensity of five for one hour. For each sport, a different formula had to be created, dependent on the intensity of that sport.

2.3. Challenge 3

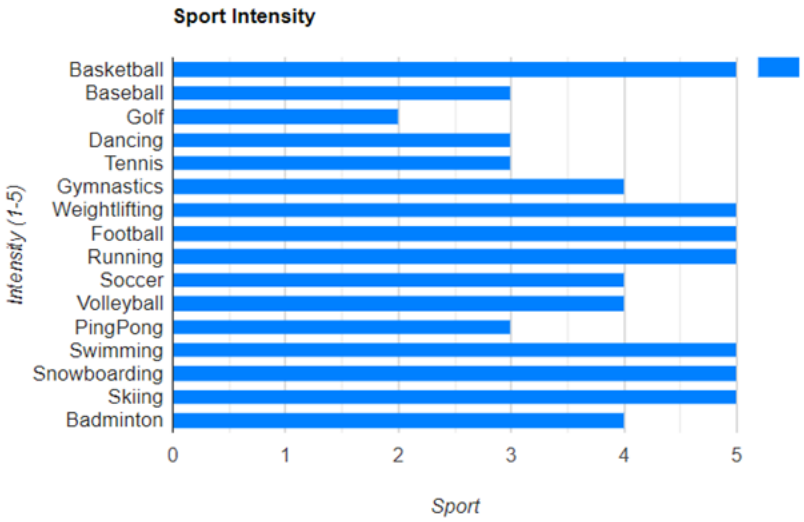
Author names are to be written in 13 pt. Times New Roman format, centered and followed by a 12pt. paragraph spacing. If necessary, use superscripts to link individual authors with institutions as shown above. Author affiliations are to be written in 12 pt. Times New Roman, centered, with email addresses, in 10 pt. Courier New, on the line following. The last email address will have an 18 pt. (paragraph) spacing following.

Deciding which foods to recommend in which categories. Along with the nutrition information already collected, we also needed to collect vitamin content, protein content, and carbohydrate content. These factors contributed to placing the foods in their respective categories. For example, a lot of meat and eggs were put in bulking, since meat is full of protein, which is crucial to the bulking process. However, foods with fewer calories and protein content were placed in slimming.

3. SOLUTIONS

These are the solutions to the previous challenges, respectively. The solution of not having a universal metric was to find the next most common metric (i.e., 100 grams, or 1 cup). This works fine now as a placeholder, but ideally, there will become an NLEA for all foods, so that Fitable can become consistent throughout. We solved the challenge of normalizing calories dependent on the sport. We recorded calories burned dependent on each sport; for example, we found that athletes burn more calories in tennis than in golf. Using this data, we calculated the number of calories burned per minute, which we would then be able to convert into the amount of time the user inputs into the app. In the diagram below, you can see we rated each sport based on the intensity and difficulty of the sport.

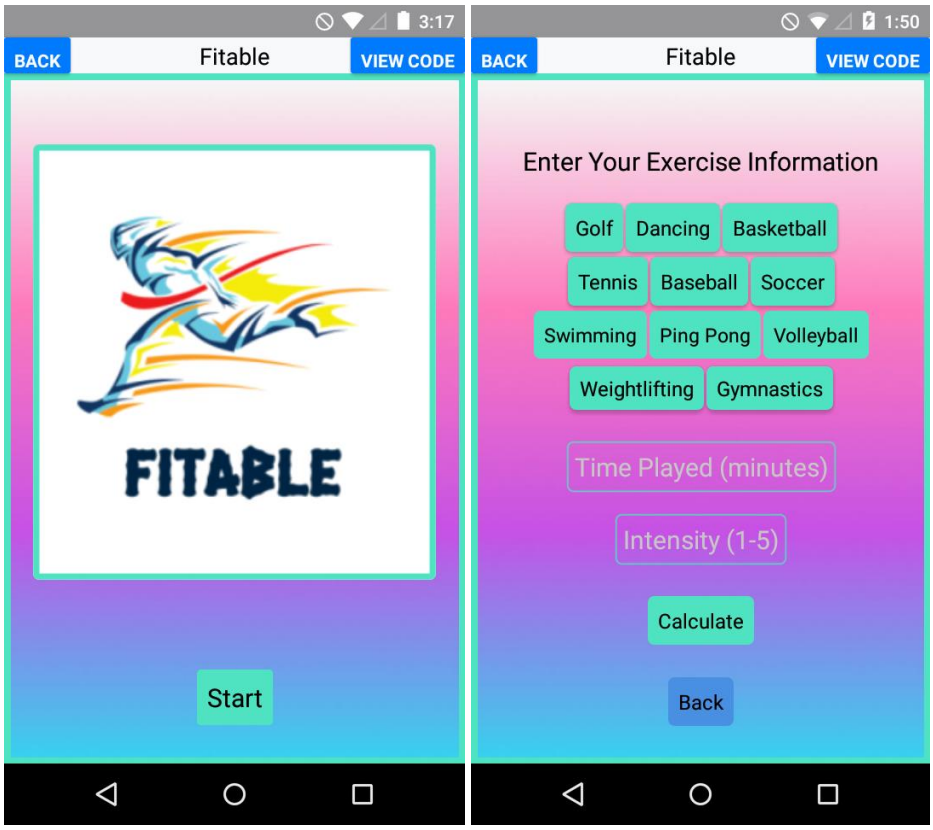
Table 2. Sports Intensity, as referenced in the Fitable app.

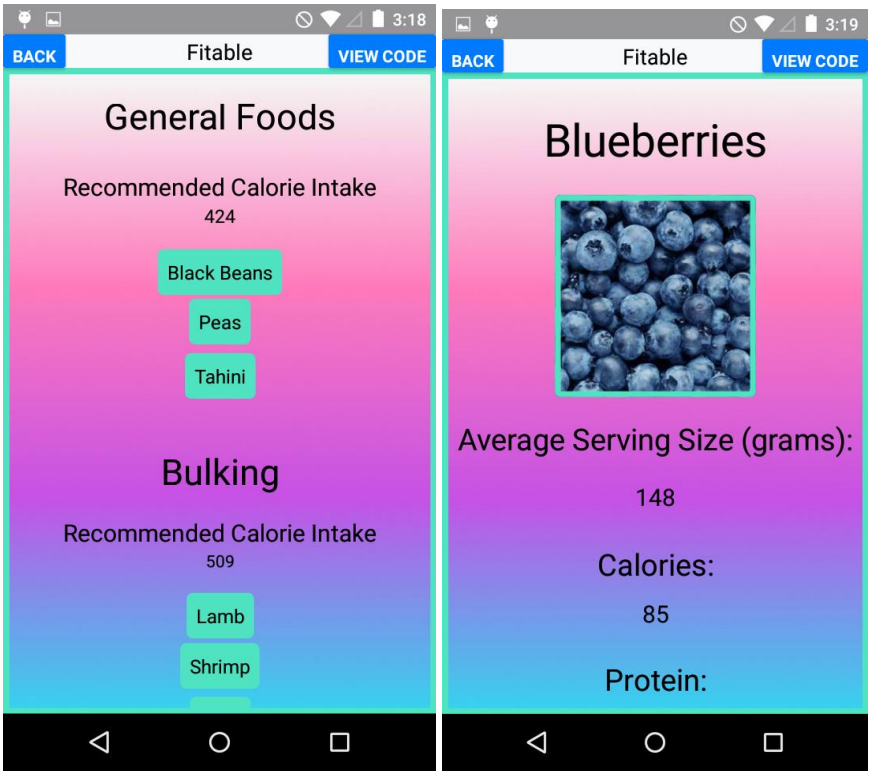


To place foods in the correct categories, we looked at their nutrition information. As stated earlier, foods with higher calories, carbohydrates, and protein content were placed into the bulking category. On the other hand, foods with fewer calories and carbohydrates were placed into the slimming category.

On the screen on the left, is the logo.

The screen on the right is where users enter their exercise information, including time played, sport, and intensity. Intensity is on a scale of one to five, where one is the weakest, and five is the strongest. The intensity of the exercises factor into the food's users receive. Users click calculate to collect their results. *The screen on the right has been updated, and more sports have been added.





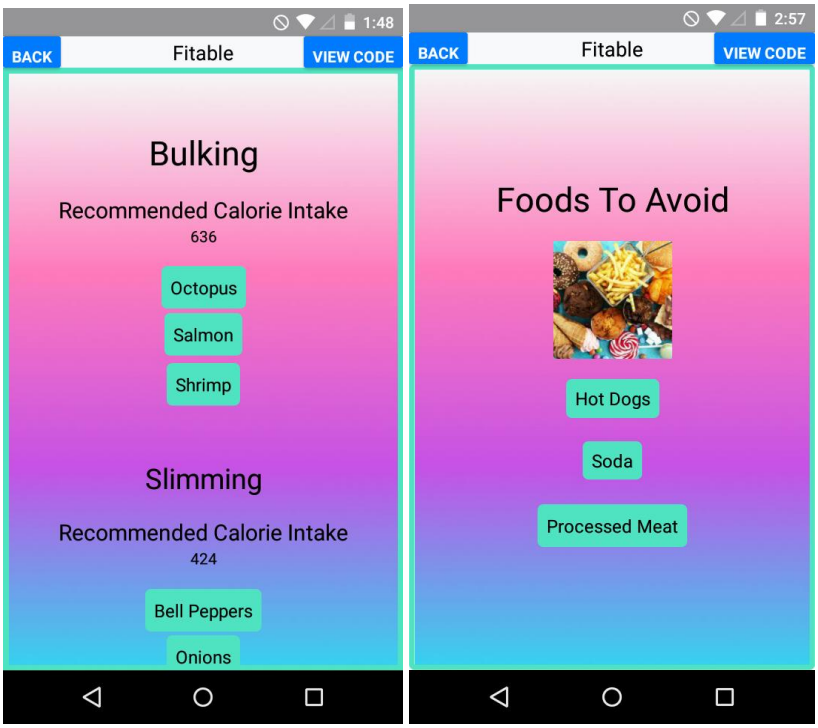
The screen on the left is where users receive recommended foods based on the information they enter. There are three categories, general, bulking, and slimming. The recommended calorie intake is the number of calories we suggest that users consume to reach the desired results. Once

they click a food, they are shown the screen on the right, where there is an image of the food, and its respective nutrition, such as average serving size, calories, protein, and carbohydrates.

*These screens are scrollable, meaning you cannot see the bottom half of the screen. On the left, along with the categories general and bulk, there is the slimming category. On the right, along with average serving size and calories, protein and carbohydrates are included.

Example:

Jimmy is a user. He has a slim body type and wants to start bulking. He just finished a workout of weightlifting for sixty minutes without stopping, thus achieving an intensity of five. He wants to know what type of foods and meals he should eat to achieve his goal of bulking. Jimmy opens Fitable and enters his information and receives a number and some suggested foods. The figure represents the recommended amount of calorie intake for him to achieve his goal.



We made a ‘Foods to Avoid Page,’ where no matter what goal users might have, they want to avoid these foods. This is a screenshot of what users might see. The foods listed here are high in sugar, cholesterol, such as soda, fast food, and donuts [12].

This formula was used to calculate the calories burned for a particular intensity, time, and difficulty of the sport. For example, if a user played soccer for sixty minutes and at an intensity of four, they would want to consume 271 calories (0.3533*4*60*4*0.8) to reach their slimming goal. Fitable then uses this number to recommend foods to the user.

cal = 0.35330.8intetimediff

- *inte* = intensity played at
- *time* = time played
- *diff* = difficulty of the sport
- *cal* = recommended number of calories to consume

4. CONCLUSION AND FUTURE WORK

In this paper, we have outlined and shown the purposes of Fitable. We have also displayed some features of our app and how potential users would use the app. Our app is motivated by the increasing number of people in the nation that are obese, a number that is close to 30% of the nation today [17]. The solutions to some of the challenges listed above have been implemented into our app. As future work, we hope to add more features, and hopefully, Fitable can be used to reduce the number of people who are obese in the nation.

The app's objective is to improve the awareness of people's diets and calorie intake. For this app to be more successful, we would like to add many aspects, some of which are recorded here.

As of now, the app is accessible to Android users but not Apple users. The goal is to publish the app into the Apple App Store, so both Android and Apple users can use it [11]. It will be free to the public, increasing the potential number of users. As of right now, we have it on Google, found here:

Personalize the user information, such as adding gender, height, weight, and body mass index or BMI, along with the sport played and intensity [15]. These features will allow our calculations to be more accurate, and therefore, recommend the best food suited for the user. This will also create the best rate of success for the user, regardless of their goals.

Another future feature is to add whole meals, rather than individual foods. This will allow the app to be much more user-friendly, giving users recommendations for easy to prepare meals, rather than forcing the user to think of meals by themselves. However, there are a variety of ways to make a specific meal, such as spaghetti, it is hard to track the number of calories in the meal [16]. Certain noodles will be healthier than others, and each noodle has different nutrition values.

Adding more sports, exercise routines, and food/meal options. This will allow a more diverse customer base, helping more people reach their exercise and weight goals.

In the future, we would recommend drinks, such as smoothies as another way to give users an idea of how to consume the recommended foods.

REFERENCES

[1] M, A. (2005, December 02). Emerging epidemic of obesity in developing countries. Retrieved from https://academic.oup.com/ije/article/35/1/93/849975
[2] FastStats - Overweight Prevalence. (n.d.). Retrieved from https://www.cdc.gov/nchs/fastats/obesity-overweight.htm
[3] The vast majority of American adults are overweight or obese, and weight is a growing problem among US children. (2018, November 27). Retrieved from http://www.healthdata.org/news-release/vast-majority-american-adults-are-overweight-or-obese-and-weight-growing-problem-among
[4] Type 2. (n.d.). Retrieved from http://www.diabetes.org/diabetes-basics/type-2/
[5] Health Risks of Being Overweight. (2015, February 01). Retrieved from https://www.niddk.nih.gov/health-information/weight-management/health-risks-overweight
[6] Fitbit. (n.d.). Retrieved from https://www.fitbit.com/home
[7] Apple Watch Series 4. (n.d.). Retrieved from https://www.apple.com/apple-watch-series-4/?afid=p238|sJKznVXu8-dc_mtid_20925qtb42335_pcrid_358346459887&cid=wwa-us-kwgo-watch-slid---apple watch-e

[8] Nutrition, P. A., & Data, O. (2015). Trends and Maps web site. US Department of Health and Human Services, Centers for Disease Control and Prevention (CDC). National Center for Chronic Disease Prevention and Health Promotion, Division of Nutrition, Physical Activity and Obesity, Atlanta, GA.
[9] Adult obesity rates rise in 6 states, exceed 35% in 7. (n.d.). Retrieved from https://www.ama-assn.org/delivering-care/public-health/adult-obesity-rates-rise-6-states-exceed-35-7
[10] (n.d.). Retrieved from https://www.calculator.net/calorie-calculator.html
[11] Frey, M. (2019, June 24). How to Use NLEA Serving Sizes to Outsmart Food Labels and Lose Weight. Retrieved from https://www.verywellfit.com/what-is-serving-size-3496390
[12] 20 Foods That Are Bad for Your Health. (n.d.). Retrieved from https://www.healthline.com/nutrition/20-foods-to-avoid-like-the-plague
[13] Fitable (n.d.). Retrieved from https://wesleyfan2015.wixsite.com/fitable
[14] Apple. (n.d.). Retrieved from https://www.apple.com/
[15] About Adult BMI | Healthy Weight | CDC. (n.d.). Retrieved from https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html
[16] Livermore, S. (2019, June 07). 93 Spaghettis That Are Total Pasta Goals. Retrieved from https://www.delish.com/cooking/g3086/spaghetti/
[17] Adult Obesity in the United States. (n.d.). Retrieved from https://www.stateofobesity.org/adult-obesity/

FUSION OF MULTI-FOCUS IMAGES WITH NEIGHBOUR LOCAL DISTANCE

Ias Sri Wahyuni[1] and Rachid Sabre[2]

[1]Universitas Gunadarma, Jl. Margonda Raya No. 100 Depok 16424, Indonesia
[2]Laboratory Biogéosciences CNRS,
University of Burgundy/Agrosup Dijon, France

ABSTRACT

The aim of multi-focus image fusion is to integrate images with different objects in focus so that we obtained a single image with all objects in focus. In this paper, we present a novel multi-focus image fusion method based on neighbour local variability (NLV). This method takes into consideration the information in the surrounding region of pixels. Indeed, at each pixel, the method exploits the local variability calculated from quadratic difference between the value of pixel and the value of all pixels that belong to its neighbourhood. It expresses the behaviour of pixel relative to all pixels belonging to its neighbourhood. The variability preserves edge feature because it detects the abrupt image intensity. The fusion of each pixel is performed by weighting each pixel by the exponential of the local variability. The precision of this fusion depends on the largenumberof the neighbourhood where the largenumber depends on the blurring characterized by the variance and its size of blurring filter. We constructed a model that gives the value of the large….. from the variance and the size of blurring filter. Comparing our method with other methods, it shows the best result.

KEYWORDS

Neighbour Local Variability; Multi-focus image fusion; Root Mean Square Error (RMSE)

1. INTRODUCTION

Due to the limited depth-of-focus of optical lenses, it is often difficult to capture an image that contains all relevant objects in focus. Only the objects within the depth-of-field are in focus, while other objects are blurred. Multi-focus image fusion is developed to solve this problem. There are various approaches that have been performed in literature. These approaches can be divided into two types: the spatial domain method and the multi-scale fusion method. The spatial domain fusion method is performed directly on the source images. In spatial domain techniques, we directly deal with the image pixels. The pixel values are manipulated to achieve the desired result. The fusion methods such as averaging, Principal Component Analysis (PCA) [1], maximum selection rule, bilateral gradient-based methods [2] and Guided Image Filter (GIF)-based method [3] and maximum selection rule fall under spatial domain approaches. The disadvantage of spatial domain approaches is that they produce spatial distortion in the fused image. Spatial distortion can be very well- handled by multi-scale approaches on image fusion. In multi-scale fusion methods, the fusion process is performed on the source images after decomposing them into multiple-scales. The discrete wavelet transform (DWT) [4]-[9], Laplacian pyramid image fusion [10]-[17], Discrete cosine transform with variance calculation (DCT+var) [18], saliency detection based method (SD)[19] are examples of image fusion techniques under transformdomain.

David C. Wyld et al. (Eds): MLNLP, BDIoT, ITCCMA, CSITY, DTMN, AIFZ, SIGPRO - 2020
pp. 191-205, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101220

In this paper, we propose pixel level multi-focus image fusion based on the neighbour local variability (NLV). This method takes into consideration the information of the surrounding region of pixels. Indeed, at each pixel, the method exploits the local variability calculated from quadratic difference between the value of pixel and the value of all pixels that belong to its neighbourhood. It expresses the behaviour of pixel relative to all pixels belonging to its neighbourhood. The variability preserves edge feature because it detects the abrupt image intensity. The fusion of each pixel is performed by weighting each pixel by the exponential of the local variability. The precision of this fusion is depending on the width of region of pixels considered in the neighbourhood. Firstly, we studied the optimal width of region for having the minimum error. Hence, we showed that the width of region depends on the blurring characterized by the variance and its size of blurring filter. We constructed a model that gives the value of the large from the variance and the size of the blurring filter.

While comparing our method with other methods existed in literature (DWT and LP-DWT), it was shown that our method gave the best result by using Root Mean Square Error (RMSE). In this work, the experimental for fusion image and compare to other methods.

This paper is organized as follows: The first section reveals the steps of the fusion process of the proposed method and a model giving the size of neighbourhood. In section 3, we studied the experimental resultsand compared our method to some recent methods. Section 4 gives conclusion of this work. In section 5, we give mathematical details for showing a propriety of the local variability.

2. THE PROPOSED METHOD

Consider the fusion of two images, I_1 and I_2 that have respectively blurred parts B_1 and B_2. These images have the same size:N_1 x N_2. We study the case where B_1 and B_2 are disjoint. The idea of the NLV fusion method consists of summing the pixel values of the two images weighted by local variability in each picture. This local variability at (x, y) is calculated from the exponential of average of the square difference between the value of the pixel (x, y) and the value of its neighbors. The NLV at (x, y) is defined as follows:

$$v_{a,k}(x,y) = \sqrt{\frac{1}{R}\sum_{m=-a}^{a}\sum_{n=-a}^{a}\left|I_k(x,y) - I_k'(x+m,y+n)\right|^2} \quad (1)$$

where k is the index of k^{th} source image (k = 1, 2), a is the size of neighborhood

$$I_k'(x+m,y+n) = \begin{cases} I_k(x+m,y+n), \ if \ 1 \le x+m \le N_1 \ and \ 1 \le y+n \le N_2 \\ I_k(x,y), \ otherwise \end{cases},$$

$$R = (2a+1)^2 - card(S),$$

$$S = \left\{(m,n) \in \left([-a,a]^2 - \{(0,0)\}\right) \middle| I_k'(x+m,y+n) = I_k(x,y)\right\}.$$

In the annex 1,it is shown that this local variability is small enoughwherethe location is on theblurred area (B_1 or B_2).

In this paper, we develop a novel fusion method that consists of weighting each pixel of each image by exponential of neighbour local variability. This neighbour local variability is calculated from the quadratic difference between the value of the pixel and the all pixel values of its neighbours. The idea came from the fact that the variation of the value in blurred region is smaller than the variation of the value in focused region. We used the neighbour, with the size "a", of a pixel defined as follows:

$(x+i, y+j)$ where $i=-a,-a+1,\ldots,a-1,a$ and $j=-a,-a+1,\ldots,a-1,a$.

For example, the neighbor with the small size ("a" = 1) contains: $(x-1,y-1)$, $(x-1,y)$, $(x-1,y+1)$, $(x,y-1)$, $(x,y+1)$, $(x+1,y-1)$, $(x+1,y)$, $(x+1,y+1)$.

	(x-1, y-1)	(x-1,y)	(x-1,y+1)	
	(x,y-1)	(*x*,*y*)	(x,y+1)	
	(x+1,y-1)	(x+1,y)	(x+1,y+1)	

Fig. 2. Pixel at (x,y) within its neighborhood, a = 1.

Then, the steps of image fusion with size "a" are as follows:

Suppose there are M original source images, I_1, ..., I_M, with different focus to be fused. The images here have the same size $(N_1 \text{x} N_2)$. The general principle of making fusion rules are:

Step 1:For each pixel of each image, we calculated the neighbor local variability (NLV) of every source image, $v_{a,k}(x,y)$ defined in (1).

Step 2: The fused image proposed, F is calculated in the following model:

$$F(x,y)=\frac{\sum_{k=1}^{M}\exp\left(v_{a,k}(x,y)\right)I_k(x,y)}{\sum_{i=1}^{M}\exp\left(v_{a,k}(x,y)\right)} \quad (17)$$

Obviously, this method depends on the size "a". First, we tried with a small size (a = 1). Hence, the NLV method is better than DWT method. To improve this method and to compare it with all other methods, we optimized the value of "a"for having the minimum Root Mean Square Error (RMSE), where RMSE is defined in section 4. For that, we showed that the value of "a" depends on the blurred area.

The choice of the size of the neighborhood "a" used in NLV method depends on variance (v) and the size(s) of the blurring filter. Our problem is to have a model that gives the value of the "a" according to the "v" and "s"; we take a sample of 1000 images that we blurred using Gaussian filter with different values of v and s (v=1,2,3,...,35 and s=1,2,3,...,20).

After that, for each image we blurred with parameters "v" and "s", we applied our fusion method with different values of "a"("a=1,2,...,17")and determined the value of "a" that gives the minimum RMSE, denoted by $a_I(v,s)$. Then, we took the mean of the $a_I(v,s)$ for 1000 images, denoted $a(v,s)$, because the coefficient of variation is smaller than 0.1.

To propose a model, firstly, we have studied the variation of "a" in according to variance "v" for each fixed size of blurring filter "s". We noted that this variation is logarithmic.For example, "s=8" on Fig. 4.By using nonlinear regression, we obtained the model:

$$a = 2.1096 \ln v + 2.8689 .$$

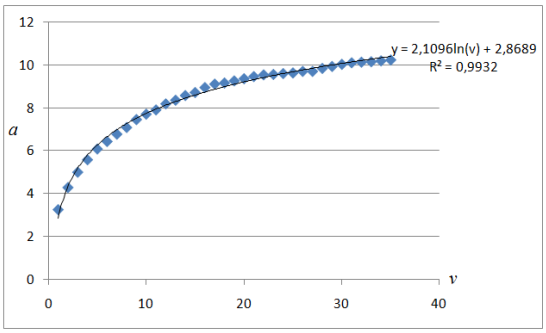
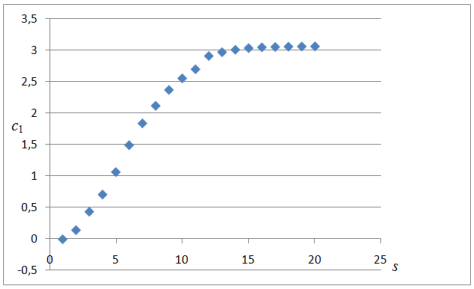


Fig. 4. Graph between "a" and variance of blurring filter where "s"=8.

In general, the model is:

$$a = c_1(s) \ln v + c_2(s) \qquad (18)$$

where the c_1 and c_2 are functions that depend on "s". The graphs that describe c_1 and c_2, respectively, are revealed in Fig. 5. and Fig. 6.



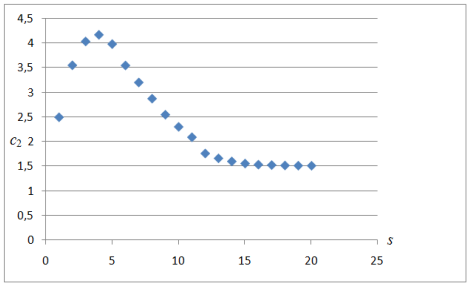


Fig. 5. graph of $c_1(s)$ **Fig. 6.** graph of $c_2(s)$.

By giving a model of c_1 and a model of c_2 and introducing these models in (19), we get the general following model:

$$a(v,s) = \frac{3.0348761}{1+29.0909139\exp(-0.5324955s)} \ln(v) + 0.434\left(\frac{75.062269}{1.225175s}\right)\exp\left(-0.5\left(\frac{\log(s)-2.655551}{1.225175}\right)^2\right) \qquad (19)$$

As "*a"* is integer, we have two choices of *a*. *It* is either the floor of $a(v,s)$, denoted by $\lfloor a(v,s) \rfloor$ or the ceiling of *a*(v,s) , denoted by $\lceil a(v,s) \rceil$ where $\lfloor x \rfloor = \min\{n \in \square \mid n \geq x\}$ and $\lceil x \rceil = \max\{m \in \square \mid m \leq x\}$. Since the RMSE values of both "*a*" are very slightlydifferent, then we can choose any "*a*" of them. We use "*a*" = $\lfloor a(v,s) \rfloor$ in the remaining part of this paper.

We validated our model by applying it to 100 images (we generated 100 pairs multi-focus images with various values of variance and size of blurring filter) and the result is as good as it was expected. Thus, our method is better than DWT and LP-DWT methods. To use this NLV method, we must firstly estimate the variance and the size of blurring filter. For that,there exists some papers giving the methods to estimate variance of blurring filter and the blur detection as in [23]-[27].We also proposed another method wherein we combined Laplacian pyramid method and NLV method. Indeed, we used Laplacian pyramid with NLV as a selection rule, denoted by LP-NLV.

3. EXPERIMENTAL RESULT

The NLV method is performed on a datasets of images [26] using Matlab2013a. We blur these images using Gaussian filter with many values of variance and size. To lighten the reading of the paper, we presented only two examples with the size 256x256 (N1 = N2 = 256). The first, image 'bird' Fig.1 and the second image 'bottle' Fig.2, all images consist of two images with different focus and one reference image.

For comparison purposes, we performed fusion using methods: PCA method [1], Discrete Wavelet Transform (DWT) method [6], Laplacian Pyramid LP_PCA [15], LP_DWT [17] and Bilateral gradient (BG) [2].

In order to compare these methods, we used the following four evaluation criteria frequently used:

Root Mean Square Error (RMSE)

RMSE finds out the difference between the reference image R and the fused image F. It gives the information how the pixel values of fused image deviate from the reference image. RMSE between the reference image and the fused image is computed as:

$$RMSE = \sqrt{\frac{1}{rc}\sum_{i=1}^{c}\sum_{j=1}^{c}\left[R(x,y) - F(x,y)\right]^2} \qquad (20)$$

where*R* is a reference image, *F* is a fused image,*r*x *c* is the size of the input image, and *x, y* represents to the pixel locations. A smaller value of RMSE shows a good fusion result. If the value of RMSE is 0 then it means the fused image is as exactly the same as the reference image.

For two images that are presented in this paper and blurred with variance = 10 and size of blurring filter = 5, the model (20) gives the neighbour size "a" = 5 and "a" = 6. Here, we use "a" = 6 because it results the smaller RMSE compared to "a" = 5 however the RMSE values of "a" = 5 and "a" = 6 are very slightly different.

Blurred image 1 Blurred image 2

Figure1. Fusion by proposed method NLV

We have found that the NLV method better fusion compared to other methods, see Fig.1.

Table 1. Performance evaluation of image 'bird'

	PCA	DWT	LP-DWT	LP-PCA	DCT+var	Bilateral gradient	GIF	SD	NLV	LP-NLV
RMSE	6.9205	3.5678	1.5190	1.4681	2.6860	8.8378	2.2792	10.4547	**0.5466**	0.8431

From the value of RMSE calculated for ten methods on Table 1, for image'bird': the smallest is NLV method, the second smallest is LP-NLV, the third is LP-PCA, as we can see on the Table 1. NLV method is the best method among the above methods and LP-NLV is better than LP-PCA and LP-DWT.

Blurred image 1 Blurred image 2

Figure2. Fusion by proposed method (NLV)

We have found that the NLV method performs better compared to other methods, see Fig.2.
To confirm our visually result, we calculated the evaluation metrics: RMSEsee Table 2. From the value of RMSE calculated for ten methods in Table 2, we can classify these methods from the smaller value of RMSE. The smallest value is NLV, the second smallest is LP-NLV, the third smallest is LP-PCA.

Table 2. Performance evaluation of images of 'the bottle'

	PCA	DWT	LP-DWT	LP-PCA	DCT+var	Bilateral gradient	GIF	SD	NLV	LP-NLV
RMSE	15.005	5.384	2.528	2.485	2.642	20.380	3.681	16.919	**0.902**	1.584

According to the evaluation measure RMSE, the Table 3 gives the mean and standard deviation of RMSE for the considered methods applied on 150 images.

Table 3. Statistic parameters of the sample (150 images)

Methods	PCA	DWT	LP_DWT	LP_PCA	DCT_var	BG	NLV	LP_NLV
Mean	8,713	4,194	2,049	1,995	2,839	11,044	0,591	1,344
Standard deviation	3,866	1,381	0,756	0,743	1,308	4,859	0,204	0,697
Time of execution by image	7s	5s	7s	7s	6s	6s	5s	7s

The results show that the proposed method (NLV) has a smaller mean of the RMSE. The histograms of RMSE for 150 images by different methods (Figures 3, 4, 5, 6, 7, 8 and 9) show for almost all methods that the values of RMSE are almost symmetrically centred around the mean value.

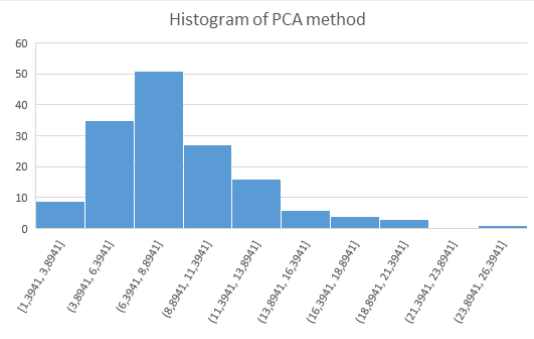


Figure 3. Histogram of PCA method

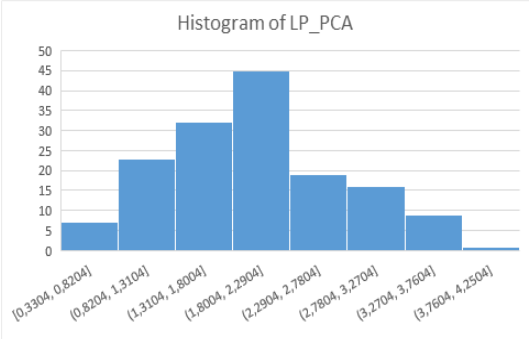


Figure 4. Histogram of LP_PCA method

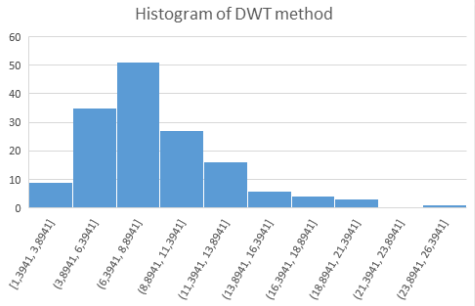


Figure 5. The Histogram of DWT method

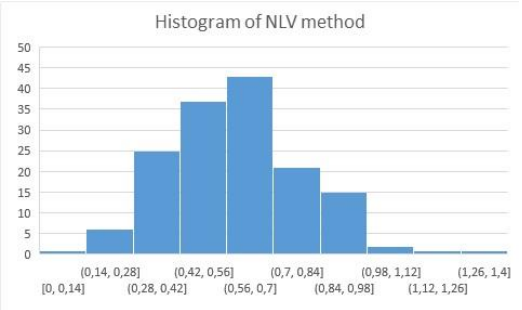


Figure 6. Histogram of LP_DWT method

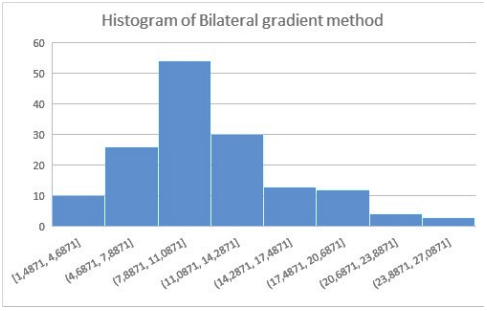


Figure 7. Histogram of Bilateral gradient method

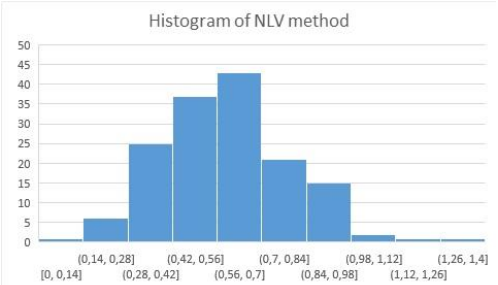


Figure 8. Histogram of NLV method

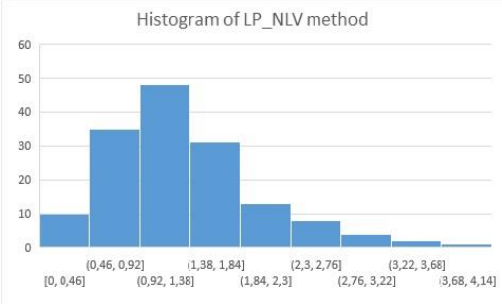


Figure 9. Histogram of LP_NLV method

To compare analytically the proposed method to other methods, we used the Analysis of variance (ANOVA) with dependent samples (dependence by image). The software R gives the following Anova table:

Table 4. Anova table with one factor: method

	Df	Sum Sq	Mean Sq	F value Pr(>F)
Method	9	25467	2829.6	742 <2e-16 ***
Residuals	1341	5114	3.8	

As Pr(>F) is smaller than 1% , in table 4., the methods are significantly different. We used thenthe Newman Keuls test to compare the methods two-by-two and made groups having significantly the same mean. The software R shows the results of the test as follows:

Table 5. Test of Newman Keuls

$groups

	RMSE	groups
SD	12.6072900	a
BG	11.0447767	b
PCA	8.7139600	c
DWT	4.1941660	d
DCT_var	2.8395233	e
GIF	2.5146353	e
LP_DWT	2.0496413	f
LP_PCA	1.9954953	f
LP_NLV	1.3446913	g
NLV	0.5921593	h

From table 5., we have the means of RMSE of methods which are significantly different except the methods DCT_var and GIF form the group “e” and the methods LP_DWT and LP_PCA form the group “f”.

The proposed method NLV has a smaller mean and significantly different of the all methods. We conclude that the proposed method is better than other methods.

4. CONCLUSION

This paper presents the image fusion method based on neighbour local variability (NLV). The principal method of fusion is described in details. The result of the experiment shows that the NLV method gives a significant improvement result in both visual and quantitative image fusion comparing to other fusion methods which are respectively DWT and LP-DWT. Laplacian pyramid with NLV as a selection rule was also applied, LP-NLV. Based on the experiment result, LP-NLV is better than LP-DWT and DWT.

The advantage of the proposed method lies in the fact that it takes into account the variability between each pixel and its neighbours. This gives a power to the coefficient of the pixel located in the focus part.This method can be extended to multimodal images used in particular in medicine (scanner, echography, X-ray, etc.) to give the presence of certain cancer cells seen in one image and not visible in another image.

The method proposed can be used inmany applications such as:

1) Drone: it is a new technology in digital imaging, it hasopened up unlimited possibilities for enhancingphotography. Drone can capture images on the samescene that zooms in on different objects, and at variousaltitudes. It produces several images on the samescene but with different objects in-focus.

2) For quality control of food industry: cameras areused to take pictures. Each camera targets one ofseveral objects to detect an anomaly. The objects are on atreadmill. To have a photo containing all the objectsclearly, we can use the proposed method of fusion whichgives more details.

The perspectives of this work:

- As many work on image fusion, implementing grayscale images, all proposedmethods in this paper are performed on the grayscale image.However, these proposed methods can be extendedto color images as color conveys significantinformation.

- We are also encouraged to fuse more than twoimages by taking into account the local variabilityin each image (intra-variability) and variabilitybetween image (inter-variability). Inter-variability can detect the 'abnormal pixels' amongthe images.

REFERENCES

[1] Naidu, V.P.S. and. Raol, J.R. (2008) "Pixel-level Image Fusion using Wavelets and Principal Component Analysis", Defence Science Journal, Vol. 58, No. 3, pp. 338-352.
[2] Tian, J., Chen, L., Ma, L., Yu, W., (2011) "Multi-focus image fusion using a bilateral gradient-based sharpness criterion", Optic Communications, 284, pp 80-87.
[3] Zhan, K., Teng, J., Li, Q., Shi, J. (2015) "A novel explicit multi-focus image fusion method", Journal of Information Hiding and Multimedia Signal Processing, vol. 6, no. 3, pp.600-612.
[4] Mallat, S.G. (1989) "A Theory for multiresolution signal decomposition: The wavelet representation", IEEE Trans. Pattern Anal. Mach. Intel., 11(7), 674-93.
[5] Pajares, G., Cruz, J.M. (2004) "A Wavelet-Based Image Fusion Tutorial", Pattern Recognition 37. Science Direct.
[6] Guihong, Q., Dali, Z., Pingfan, Y. (2001) "Medical image fusion by wavelet transform modulus maxima". Opt. Express 9, pp. 184-190.
[7] Indhumadhi, N., Padmavathi, G., (2011) "Enhanced Image Fusion Algorithm Using Laplacian Pyramid and Spatial Frequency Based Wavelet Algorithm", International Journal of Soft Computing and Engineering (IJSCE). ISSN: 2231-2307, Vol. 1, Issue 5.
[8] Sabre, R. Wahyuni, I.S, (2020) "Wavelet Decomposition and Alpha Stable", Signal and Image Processing (SIPIJ), Vol. 11, No. 1. pp. 11-24.
[9] Jinjiang Li , Genji Yuan and Hui Fan (2019) "Multifocus Image Fusion Using Wavelet-Domain-Based Deep CNN", Computational Intelligence and Neuroscience, Vol. 2019 Article ID 4179397 | https://doi.org/10.1155/2019/4179397
[10] Burt, P.J., Adelson, E.H. (1983) "The Laplacian Pyramid as a Compact Image Code", IEEE Transactions on communication, Vol.Com-31, No 40.
[11] Burt, P.J. (1984) "The Pyramid as a Structure for Efficient Computation. Multiresolution Image Processing and Analysis", A. Rosenfeld, Ed., Springer-Verlag. New York.
[12] Burt, P.J., Kolezynski, R.J. (1993) "Enhanced Image Capture Through Fusion", in: International Conference on Computer Vision, pp. 173-182.
[13] Wang, W., Chang, F. (2011) "A Multi-focus Image Fusion Method Based on Laplacian Pyramid", Journal of Computers, Vol.6, No 12.

[14] Zhao, P., Liu, G., Hu, C., Hu, and Huang, H. (2013) "Medical image fusion algorithm on the Laplace-PCA". Proc. 2013 Chinese Intelligent Automation Conference, pp. 787-794.
[15] Verma, S. Kaur K.,, Kumar M., R..(2016) "Hybrid image fusion algorithm using Laplacian Pyramid and PCA method", proceeding of the Second International Conference on Information and Communication Technology for Competitive Strategies.
[16] Wahyuni, I.S, Sabre, R. (2019) "Multifocus Image Fusion Using Laplacian Pyramid Technique Based on Alpha Stable Filter", CRASE Vol. 5, No. 2. pp. 58-62.
[17] Wahyuni, I.S, Sabre, R. (2016) "Wavelet Decomposition in Laplacian Pyramid for Image Fusion", International Journal of Signal Processing Systems Vol. 4, No. 1. pp. 37-44.
[18] Haghighat, M.B.A, Aghagolzadeh, A., Seyedarabi, H. (2010)"Real-time fusion of multifocus images for visual sensor networks". Machine vision and image processing (MVIP), 2010 6th Iranian. 2010.
[19] Bavirisetti, D.P., and Dhuli, R.(2016) ''Multi-focus image fusion using multi-scale image decomposition and saliency detection'', Ain Shams Eng. J., to be published. [Online]. Available: http://dx.doi.org/ 10.1016/j.asej.2016.06.011.
[20] Petland, A.(1984) "A new sense for depth of field", IEEE Transactions on Pattern Analysis and Machine Intelligent, Vol. 9, No. 4, pp. 523-531.
[21] Nayar, S.K.(1992) "Shape from Focus System", Proc. of IEEE Conf. Computer Vision and Pattern Recognition, pp. 302-308.
[22] Gonzales, R.C., Woods, R.E. (2002) "Ditigal Image Processing" 2nd edition. Prentice Hall.
[23] Liu,R., Li, Z., Jia, J.(2008) "Image Partial Blur Detection and Classification", Computer Vision and Pattern Recognition, CVPR 2008. IEEE Conference DOI: 10.1109/CVPR.2008.4587465
[24] Aslantas, V. (2007) "A depth estimation algorithm with a single image." Optic Express, Vol. 15, Issue 8. OSA Publishing.
[25] Elder, J.H., Zucker, S.W.(1998) " Local Scale Control for Edge Detection and BlurEstimation." IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol.20, No.7.
[26] www.rawsamples.ch. Accessed: 15 November 2017.
[27] Kumar, A., Paramesran, R., Lim, C. L., and Dass, S.C. (2016) "Tchebichef moment based restoration of Gaussian blurred images". Applied Optics, Vol. 55, Issue 32, pp. 9006-9016.

5. ANNEX 1

Consider, without loss the generality that we have a focus pixel (x, y) in image I1 and blurred in image I2 as in Fig. 1.

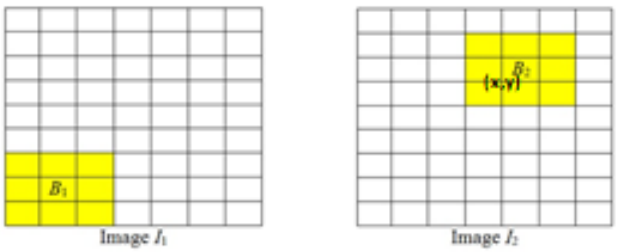


Fig. 1. Two multi-focus images, the yellow part is blurred area and the white part is clear (focus) area.

The neighbor local variability of images I1 and I2, respectively is defined in (1) by:

$$v_{a,1}(x,y) = \exp\left(\sqrt{\frac{1}{R}r_1(x,y)}\right) \text{and } v_{a,2}(x,y) = \exp\left(\sqrt{\frac{1}{R}r_2(x,y)}\right) \text{where} r_1(x,y) \text{and } r_2(x,y)$$

can be written as follows:

$$r_1(x,y) = \sum_{m=0}^{2a} \sum_{n=0}^{2a} \left| I_1(x,y) - I_1\left(x+(m-a), y+(n-a)\right) \right|^2 \quad (2)$$

$$r_2(x,y) = \sum_{m=0}^{2a} \sum_{n=0}^{2a} \left| I_2(x,y) - I_2\left(x+(m-a), y+(n-a)\right) \right|^2 \quad (3)$$

Let I_R is the reference image of multi-focus images I_1 and I_2. Moreover, it is shown in [20] and [21] that the blurred image can be seen as the product convolution between the reference image and Gaussian filter:

$$I_1(x,y) = \begin{cases} w_1 * I_R(x,y), & (x,y) \in B_1 \\ I_R(x,y), & (x,y) \notin B_1 \end{cases} \quad I_2(x,y) = \begin{cases} w_2 * I_R(x,y), & (x,y) \in B_2 \\ I_R(x,y), & (x,y) \notin B_2 \end{cases}, \quad (4)$$

where w_1 and w_2 are Gaussian filter defined by:

$$w_1(k,l) = w_1(k,l) = \frac{\exp\left(-\frac{k^2+l^2}{2\sigma_1{}^2}\right)}{\sum_{k=-s_1}^{s_1} \sum_{l=-s_1}^{s_1} \exp\left(-\frac{k^2+l^2}{2\sigma_1{}^2}\right)}, \ (\mathrm{k},\mathrm{l}) \in [-s_1, s_1]^2,$$

$$w_2(k,l) = \frac{\exp\left(-\frac{k^2+l^2}{2\sigma_2{}^2}\right)}{\sum_{k=-s_2}^{s_2} \sum_{l=-s_2}^{s_2} \exp\left(-\frac{k^2+l^2}{2\sigma_2{}^2}\right)}, (k,l) \in [-s_2, s_2]^2$$

The product convolution is defined as follows:

$$w_1 * I_R(x,y) = \sum_{k=-s_1}^{s_1} \sum_{l=-s_1}^{s_1} w_1(k,l) I_R(x-k, y-l), \ w_2 * I_R(x,y) = \sum_{k=-s_2}^{s_2} \sum_{l=-s_2}^{s_2} w_2(k,l) I_R(x-k, y-l),$$

$$\text{Put} \quad r_1(x,y) = \sum_{m=0}^{2a} \sum_{n=0}^{2a} \left| D^1_{(m,n)}(x,y) \right|^2 \quad \text{and} \quad r_2(x,y) = \sum_{m=0}^{2a} \sum_{n=0}^{2a} \left| D^2_{(m,n)}(x,y) \right|^2 \quad (5)$$

$$\text{where} \quad D^1_{(m,n)}(x,y) = I_1(x,y) - I_1\left(x+(m-a), y+(n-a)\right) \quad (6)$$

$$D^2_{(m,n)}(x,y) = I_2(x,y) - I_2\left(x+(m-a), y+(n-a)\right) \quad (7)$$

Proposition:

The local variability on blurred part is smaller than the local variability on focused part. Let $(x,y) \in B_2$ (the blurred part of I_2) and $(x,y) \notin B_1$ (focus par of I_1), then $(r_2(x,y) \le r_1(x,y))$.

Proof:

For that, we use Plancherel theorem:

$$\sum_{m=0}^{2a}\sum_{n=0}^{2a}\left|D^1_{(m,n)}(x,y)\right|^2 = \frac{1}{(2a+1)^2}\sum_{m=0}^{2a}\sum_{n=0}^{2a}\left|\widehat{D}^1_{(n,m)}(x,y)\right|^2 (8)$$

where $\widehat{D}^1_{(n,m)}(x,y)$is Fourier transform of$D^1_{(m,n)}(x,y)$.

$$\widehat{D}^1_{(n,m)}(x,y) = FT\left[D^1_{(m,n)}(x,y)\right] = FT\left[I_1(x,y) - I_1\left(x+(m-a), y+(n-a)\right)\right] \quad (9)$$

As $(x,y)\in B_2$ therefore $(x,y)\notin B_1$, from (4), equation (9) can be written as follows:

$$\widehat{D}^1_{(n,m)}(x,y) = FT\left[I_R(x,y) - I_R\left(x+(m-a), y+(n-a)\right)\right] \quad (10)$$

and

$$I_2(x,y) = \sum_{k=-s_2}^{s_2}\sum_{l=-s_2}^{s_2} w_2(k,l) * I_R(x-k, y-l) \quad (11)$$

By using the definition of convolution, equation (11) can be written as:

$$I_2(x,y) = \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} w_2(k,l)1_{[-s_2,s_2]^2} I_R(x-k, y-l) (12)$$

and

$$I_2(x,y) = \left(w_2 1_{[-s_2,s_2]^2}\right) * I_R(x,y) (13)$$

Where

$$1_{[-s_2,s_2]^2}(k,l) = \begin{cases} 1, & if\ \ (k,l) \in [-s_2,s_2]^2 \\ 0, & otherwise \end{cases}$$

.

The Fourier transform of $D^2_{(m,n)}(x,y)$ is

$$\widehat{D}^2_{(n,m)}(x,y) = FT\left[w_2 1_{[s_2,s_2]^2} * I_R(x,y) - w_2 1_{[s_2,s_2]^2} * I_R\left(x+(m-a), y+(n-a)\right)\right]$$

$$= FT\left[w_2 1_{[s_2,s_2]^2} * \left(I_R(x,y) - I_R\left(x+(m-a), y+(n-a)\right)\right)\right]$$

$$= FT\left[w_2 1_{[s_2,s_2]^2}\right] FT\left[I_R(x,y) - I_R(x+(m-a), y+(n-a))\right] \quad (14)$$

Substitute (10) into (14), we get

$$\widehat{D}^2_{(n,m)}(x,y) = FT\left[w_2 1_{[s_2,s_2]^2}\right]\widehat{D}^1_{(p,q)}(x,y)$$

$$= \left(\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} w_2(k,l)1_{[s_2,s_2]^2}(k,l)e^{-i2(kp+lq)}\right)\widehat{D}^1_{(n,m)}(x,y) \quad (15)$$

Hence, from equation (15), we can obtain

$$\left|\widehat{D}^2_{(n,m)}(x,y)\right| = \left|\sum_{k=-s_2}^{s_2}\sum_{l=-s_2}^{s_2}\frac{e^{-\left(\frac{k^2+l^2}{2\sigma_2{}^2}\right)}}{\sum_{k'=-s_2}^{s_2}\sum_{l'=-s_2}^{s_2} e^{-\left(\frac{k'^2+l'^2}{2\sigma_2{}^2}\right)}} e^{-i2(kn+lm)}\,\widehat{D}^1_{(n,m)}(x,y)\right|$$

$$\leq \sum_{k=-s_2}^{s_2}\sum_{l=-s_2}^{s_2}\left|\frac{e^{-\left(\frac{k^2+l^2}{2\sigma_2{}^2}\right)}}{\sum_{k'=-s_2}^{s_2}\sum_{l'=-s_2}^{s_2} e^{-\left(\frac{k'^2+l'^2}{2\sigma_2{}^2}\right)}}\right|\left|\widehat{D}^1_{(n,m)}(x,y)\right| \leq \left|\widehat{D}^1_{(n,m)}(x,y)\right| \quad (16)$$

On the other hand, from equation (5) and Plancherel-Parseval's theorem, we have

$$r_2(x,y) = \sum_{m=0}^{2a}\sum_{n=0}^{2a}\left|D^2_{(m,n)}(x,y)\right|^2 = \frac{1}{(2a+1)^2}\sum_{m=0}^{2a}\sum_{n=0}^{2a}\left|\widehat{D}^2_{(n,m)}(x,y)\right|^2$$

From (16), we get

$$r_2(x,y) \leq \frac{1}{(2a+1)^2}\sum_{m=0}^{2a}\sum_{n=0}^{2a}\left|\widehat{D}^1_{(p,q)}(x,y)\right|^2 \leq \sum_{m=0}^{2a}\sum_{n=0}^{2a}\left|\widehat{D}^1_{(m,n)}(x,y)\right|^2$$

$$r_2(x,y) \leq r_1(x,y)$$

This proves that the local variability in blurred part is smaller than local variability in focused part.

AUTHORS

Rachid Sabre received the PhD degree in statistics from the University of Rouen, Rouen, France, in 1993 and Habilitation (HdR) from the University of Burgundy, Dijon, France, in 2003.He joined Agrosup Dijon, Dijon, France, in 1995, where he is an Associate Professor. From 1998 through 2010, he served as a member of "Institut de Mathématiques de Bourgogne", France. He was a member of the Scientific Council AgroSup Dijon from 2009 to 2013. In 2012, he has been a member of " Laboratoire Electronique, Informatique, et Image" (Le2i), France. Since 2019 has been a member of Laboratory Biogeosciences UMR CNRS, University of Burgundy. He is author/co-author of numerous papers in scientific and

technical journals and conference proceedings. His research interests lie in areas of statistical process and spectral analysis for signal and image processing.

Ias Sri Wahyuni was born in Jakarta, Indonesia, in 1986. She earned the B.Sc. and M.Sc. degrees in mathematics from the University of Indonesia, Depok, Indonesia, in 2008 and 2011, respectively.In 2009, she joined the Department of Informatics(COMPUTING) System, Gunadarma University, Depok, Indonesia, as a Lecturer. She is currently a PhD student at University of Burgundy, Dijon, France. Her current research interests include statistics and image processing.