CUSTOMER OPINIONS EVALUATION: A CASE-STUDY ON ARABIC TWEETS

Manal Mostafa Ali

Al-Azhar University, Faculty of Engineering Computer&
System Engineering, Egypt

ABSTRACT

This paper presents an automatic method for extracting, processing, and analysis of customer opinions on Arabic social media. We present a four-step approach for mining of Arabic tweets. First, Natural Language Processing (NLP) with different types of analyses had performed. Second, we present an automatic and expandable lexicon for Arabic adjectives. The initial lexicon is built using 1350 adjectives as seeds from processing of different datasets in Arabic language. The lexicon is automatically expanded by collecting synonyms and morphemes of each word through Arabic resources and google translate. Third, emotional analysis was considered by two different methods; Machine Learning (ML) and rule-based method. Finally, Feature Selection (FS) is also considered to enhance the mining results. The experimental results reveal that the proposed method outperforms counterpart ones with an improvement margin of up to 4% using F-Measure.

KEYWORDS

Opinion Mining - Arabic - Bag of Words - Feature Selection - Emotions- Adjective Lexicons.

1. INTRODUCTION

Mining customer's opinions aid in gauging reactions, targeting advertising, and evaluation of public voters' opinions. Besides reputation management and public relations, one could perform trend prediction in sales or other relevant data. Hence, by polling of this information, we can give quantitative indications of customer's positive or negative attitude about products, services or business [1]. In general, extracting useful patterns and detecting customer feedback from natural language is challenging and could be time consuming for several reasons. It is difficult to distinguish between objective and subjective information. News itself can be generally classified as good or bad news without being subjective [2].Text Classification (TC) also requires deeper analysis and understanding of textual features [3]. In opinion texts, lexical content alone can be misleading. Furthermore, most of the current studies related to this topic focus mainly on English texts with very limited resources available for other languages such as Arabic [4] [5].

Most of the recent work in Arabic TC is not yet releasing their resources [6] [7]. Several studies share the same weak points such as using a few features for opinion mining. Some of these attempts are based on statistical approaches applied on Bag of Words (BoW) as in [4]. Most of them neglect semantic analysis as in as in [8], and [9]. Others consider semantic features but ignore morphological information as in [10]. Some systems don't handle negation that inverts the statement classification as in [4] [8]. Another type of researches do not pay attention to emotions such as [5] [10] and depends only on the linguistic information. As a result, they lack a common framework which combines genre types of features. Lexical features only capture

local information and ignore possible relations between terms. It also loses the order of the word and ignores grammatical structure. Furthermore, BoW representation suffers from huge feature vectors that should be carefully considered to avoid hardware limitation, software capabilities, and computational time complexity [10].

In contrast to these approaches, we provide an integrated framework to analyze the Arabic text lexically, morphologically, and semantically. Such framework can devise more accurate and reduced representations of Arabic texts. Next, we define new lexicons for positive and negative adjectives. Then, we pay particular attention to emotion recognition and classification that can help in automatic annotation. Finally, FS is also considered for high dimensional social media content. The rest of the paper is organized as follows. Section 2 describes previous work on customer opinion evaluation with a focus on classification of Arabic opinion. Section 3 provides the proposed methodology. Section 4 portrays the models used in the comparative evaluation and analyzes the final results. Concluding remarks and future work are referred in Section 5.

2. RELATED WORK

This section reviews and discusses some research areas that are closely related to customer opinion evaluation and pay particular attention to Arabic opinion classification.

a. Customer Opinion Evaluation

Yingcai et al [1] presented Opinion Seer; an interactive visualization system that visually analyzes a large collection of online hotel customer reviews. [11] presented a new framework for measuring customer satisfaction with mobile services that use Sentiment Analysis (SA) and VIKOR technologies.[12] developed a robust classification approach of customer reviews by applying a statistical approach on a self-annotated corpus. The authors first identified subjective words in each test sentence and then used a Mutual Information (MI) approach to find the Sentiment Orientation (SO). The algorithm has been tested in 24159 test sentences across the six domains with an accuracy of 75.27%.In other work [13] Minqing et al provide a feature-based summary of customer reviews of a product. They proposed a set of techniques for mining and summarizing product reviews based on data mining and NLP methods. The average values of recall and precision for opinion sentence extraction are 69.3 %, 64.2% respectively while the sentence orientation accuracy was 84.2%.The main limitation of the two previous researches is that, the authors only used adjectives as indicators of opinion orientations. There are also nouns, adverbs etc., which bear sentiment and should be involved in the classification process. Another study [14] proposes a tool for aiding the evaluation of customer satisfaction in a Brazilian online job search company through the use of SA. The authors stumbled with the problem of finding SA tools for processing texts in Portuguese. Therefore, they performed an extra translation step by employing google translate. But, the translation step may lead to semantic information loss.

Additionally, the usage of customer opinions is not limited to the business organizations. It is important for many different applications such as government and business intelligence to track the public opinions [15]. It also can be used to understand the opinions of voters about political events [6] [14]. Therefore, Shulong et al [16] are interested in analyzing the latent reasons behind the public sentiment variations regarding a certain target. However, the aforementioned researches dealt broadly with SA for English language. Accordingly, there is a need to pay particular attention to Arabic customer reviews and develop new resources aid in measuring customer satisfaction.

b. Arabic Opinion Classification

[17] shed light onto the existing work on Arabic SA. They surveyed large number of studies, methods and the available Arabic sentiment resources.[10] demonstrated a comparative study of the different ATC algorithms. The study presents a comparison between the supervised ML algorithms found in the literature. It reveals that kNN and SVM are the best performing classifiers and score accuracy of 84.75 % and 84.5% respectively. A Similar comprehensive study of the different tools for Arabic text preprocessing, attribute selection, reduction and classification is presented by Khorsheed and Al-Thubaity [18]. The results show that the superiority of SVM followed by the Decision Tree (DT) and Naiive Bayes (NB).

Another study proposed by Hossam et al [5] who presented SA for Modern Standard Arabic (MSA) and Egyptian dialect with a corpus of different types of data. They employed a number of novel and rich features that include valence shifters, negation, intensifiers, questions and supplication terms to improve the classification performance. Also, [4] considered the multi-way SA problem for Arabic reviews and coupled BoW with the most popular classifiers. They used a dataset of more than 63,000 book reviews based on a 5-star rating system. Their results showed that Multinomial Naiive Bayes (MNB) had the best performance for both balanced and unbalanced datasets with weighted average accuracy reached 46.4%. However, the obtained accuracies are low confirming the intuition that the multi-way SA problem is difficult and needs further attention.

Abdul-Mageed et al [9] in their contribution propose SAMAR system, which adopts two-stage classification approach; subjectivity and SA for Arabic. A similar finding was reported by [8] who use a Genetic Algorithm (GA) for both English and Arabic Web forums sentiment detection on the document level. For this purpose, they provide different feature sets consisting of syntactic and stylistic features. Although the previous two papers [8] [9] used varieties of feature sets, they avoided semantic features because they are language dependent and need lexicon resources. There remains, however, the semantic problem that primarily emanating from the lack of Arabic semantic resources. A few types of research concerning ATC used Arabic Word Net (AWN) for improving classification results such as [19] [20]. Here, we expand AWN as a semantic resource for synonyms substitution and root extraction.

3. PROPOSED METHODOLOGY

Exploring NLP can achieve accurate results by resolving context of the words. Lexical, morphological and semantic analyses were performed. International Corpus of Arabic (ICA) [21] is fundamentally used to handle Arabic morphological pluralism. AWN [19] [20] is exploited to extract semantic relations for the lexical units. A system for automatic ED and mood recognition was built using a combination of ML and rule-based methods. The rule set can minimize the tedious effort required for the learning process. FS was included to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. We also introduce a Most Relevant Features (MRF) as a new algorithm for FS. Performance of MRF was compared with the classical CHI-Test algorithm. A schematic overview of the proposed methodology was graphically given in Figure 1.

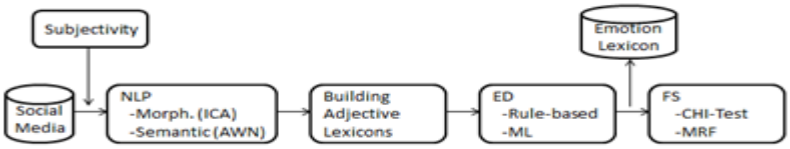

Figure 1- System Architecture

During this research, SVM [22] is conducted as the classifier because of its reported performance [3] [18]. Moreover, Random Forest (RF) [23] is tuned as another classification algorithm due to its superior performance over a single DT. Term presence - based weighting scheme [10] is adopted to measure the weights of attributes because it reveals an interesting difference [2]. We also developed a new lexicon to generate a set of sentiment words as a new dimension for Arabic language. To evaluate the proposed method, we conducted several experiments using different datasets about celebrities, companies and some products. We primarily focus on Arabic tweets. Therefore, we start by fetching Arabic datasets about some specific queries. Then, a prototype was developed to extract empty/null and repeated messages from the dataset. The process of subjectivity refers to the task of classifying texts as either objective or subjective. Therefore, the textual datasets have two primary configurations; either facts or opinions [6] [9]. Facts are known as objective information about elements, objects, occasion and their properties. "E.g.,سامسونج تقرر إطلاق تحديثًا لإيقاف جالاكسي نوت 7 عن العمل تمامًا". On the other hand, opinions are generally subjective expressions that illustrate individual's sentiments. "E.g., شبكة فودافون النهاردة ممتازة". In general, facts can be generalized by exploiting special dictionaries (so-called filters), containing synonyms for objective words. The main flaw of our approach is the necessity of manual selection of terms for the filters that can extract out facts, news, advertisements and exclude neutral concepts.

3.1. Datasets Pre-Processing

Tokenization [24], normalization [25] [26], stemming, stop words[1] filtering [10] [27] and spelling correction were performed on the underlying dataset. In addition, Arabic tweets may contain "Arabizi", where Arabic words are written using Latin characters [28].Therefore, we also handle franco words and convert these word to its Arabic equivalent by examining google's API. We expand ICA [21] to provide varieties of morphological information as well as to overcome the high dimensional datasets. ICA is exploited for two purposes; firstly to diminish data sparseness by converting multiple variants of the same words to their common lemma. For example, the verbs (نشترى –اشترت - اشتروا-يشترى) are converted to one lemma which is (اشترى). Secondly, ICA tags every token with a code indicating its basic PoS tagging, gender and number. For example:

"من يريد ان يتعلم الصبر يشترى تليفون سامسونج ايفون افضل من سامسونج صحيح سعره غالى
لكن صوته جميل وشكله رائع"

PoS tagger associates every word with the typical PoS tag as following:

"PPN/ من , VER /يريد , VER / يتعلم , NOU / الصبر , VER / يشترى , NOU / تليفون , ADJ / صحيح , NOU / سعره , PTL / غالى, NOU / صوته , ADJجميل /, NOUوشكله / , ADJ / رائع "

Every term has been attached with a relevant tag indicating its role in the sentence, such as VER(verb), NOU(noun), ADJ (adjective), PTL (participle), PPN (proper noun), etc. The entire list of tags and their meaning is based on the ICA Tagset[2]. It should be considering that the

original form of a term is returned if it is unknown to the PoS tagger. Therefore, the word "افضل - سامسونج - ايفون"will be returned without annotation. In addition, some researchers suggest that there is a relationship between gender and sentiment expression [9]. So, ICA is extended to provide the gender types of the underlying feature set. Those types are masculine, female and unknown corresponding to the set {MCL, FEM, UNKNOWN).It also can determine the number classes; singular, dual, and plural corresponding to the set {SNG, DUA, PLR}. Sample is given in Figure 2.

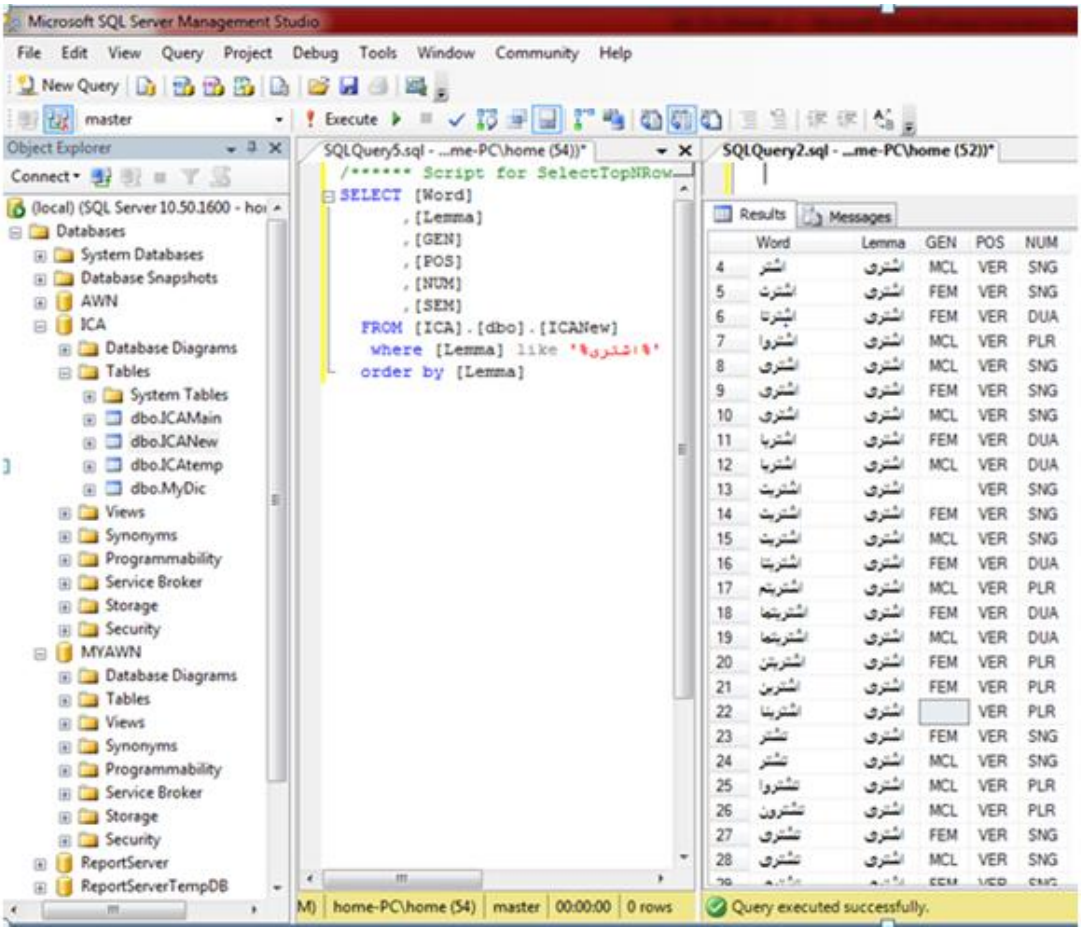

Figure 2- Morphological Properties for Some Verbs

BoW contains lots of items that have various syntactical forms but have similar meanings. For example, the words (متفوق – ممتاز - رائع) are different vocabularies but have similar meaning. Classifiers cannot deal with such words as correlated words that provide similar semantic interpretations. So AWN semantic relations are used to group those phrases into one synset. For example "♥اعلان فودافون الجديد متميز و جميل اووووي" has two synonyms "متميز"and "جميل" that would be replaced with "جميل". The model aggregates synonymous words into one cluster and each cluster is represented by a single word; the one that is commonly used in that context.

3.2. Adjectives

The presence of adjectives is useful for predicting whether a sentence is subjective. Hence, we also focus on extracting large number of adjectives relying on the fact that adjectives can mirror

most of subjective information in a given text. Adjectives can be extracted in the following manner:

```
for each Tk in the dataset {                    // Tk refers to tweet
        if (Tk doesn't contain negation article){
                if (Tk contains positive adjective)
                        classifyTk as positive
                else if (Tk contains negative adjective)
                        classify Tk as negative
}
}
```

Again using ICA, we can define a new dictionary of adjectives "AdjLex". The AdjLex contains about 111000 adjectives which is manually created and automatically expanded. Starting with the gold-standard 1350 adjectives collected manually as a seed from different datasets in Arabic language. Then, the initial lexicon is expanded by collecting synonyms and morphemes of each word. Finally, this lexicon was extended through google translate to get more acronyms for Arabic adjectives.

3.3. Emoticons Detection

A number of frequently adopted emotion labels had used to construct an emotion lexicon. All these labels are finally mapped to positive or negative category. We define a set of new hand-coded rules in a manner that accurately classifies tweets as positive or negative according to the type of emotions. However, some emotions tend to appear multiple times in a single sentence. They may even overlap between contradicting instances leading to contradictory sentiments within the tweet. Hence, depending only on emoticons as tweet's labels can introduce noise and adversely affect the performance. In such case, emotion is replaced with its typical meaning and added as a normal token. Then, all verbal information should be considered with the emotion label during classification. The following algorithm addresses the emotions classification.

```
for each Tk in the corpus{
        if (emotion has clear meaning){
                 If (Tk contains positive emotion)
                        classify as positive
                else If (Tk contains negative emotion)
                        classify as negative
        }
                Else            //emotion appears in positive and negative Tk
                replace the emotion with its meaning.
}
```

3.4. Most Relevant Features (MRF)

Clearly, using the whole tweets is confusing and misleading to be used for training [31]. MRF defines the usefulness of a feature by both its relevancy and redundancy. That is, a feature is said to be relevant if it is predictive of the decision features, otherwise, it is irrelevant. A feature is considered to be redundant if it is highly correlated with other instances. Hence, the search for a good feature subset involves finding those features that are highly correlated with the decision features but are uncorrelated with each other. MRF has two primary stages; performing Hapax Legomenon and applying probabilities laws.

3.4.1. Hapax Legomenon

The algorithm starts by counting the number of times a word occurs overall the dataset and arranging these frequencies in a descending order. Hapax legomenon [29] helps remove less common features.

3.4.2. Probabilities

Depending on probability laws, we can define the following rules:

1- If the feature is distributed with equal probabilities along positive and negative instances as in equation 1 or almost equal probabilities as in equation 2 then, it can be removed.

$$P(P) = P(N) \quad (1)$$

or

$$P(P) \approx P(N) \quad (2)$$

Where P(P) refers to probability of frequent words with positive T_k while P(N) denotes the probability of words come with negative T_k. This rule decreases the probability of mistaking important terms as redundant ones in the searching process. Eventually, the function tries to capture the intuition that the best features for T_k are the ones distributed most differently in the sets of positive and negative T_k. There are many words have low probabilities and almost equally distributed along each category as indicated by equation 2. Such terms within some pre-defined range can be ignored. In our research, we ignore all terms that has probability 0.51 with positive instances and 0.49% with negative ones and vice versa.

2- The attributes with a high probability of belongingness to dominantly one class are a candidate to identify the category of the unknown instance. The positive tokens can be identified by equation 3and the negative tokens can be identified by equation 4 respectively.

$$P(P) >> P(N) \quad (3)$$

or

$$P(N) >> P(P) \quad (4)$$

Equations 3 and 4 are important in the rule-based method, as they pick up the most frequently-occurring features in the corpus as polarity indicators. If a feature is highly encountered in positive T_k with probability nearly one, the unknown $T_k$that have this feature could be tagged as positive. On the other hand, if a feature is frequented purely in negative T_k, the new T_k would be elected to be negative. By applying these rules, a new lexicon was generated that can save a time consumed not only in learning but also in classification process. This lexicon is confirmed by two word lists. The first list is confirmed by "positive words" that suggest a positive opinion in an opinionated context (e.g "ممتاز"). The second list is confirmed by "negative words". These words suggest a negative opinion (e.g. "سيء").

4. EXPERIMENT, RESULT, AND EVALUATION

The algorithm has been tested on 8423 test sentences across two different domains; celebrities (D_1), communication companies and some products (D_2). $D_1$is about (Preachers, Announcers, Artists, and Politicians),D_2 includes (Vodafone, Mobinil and Etisalat, Rolls Royce, Samsong). Table1 shows the characteristics of the test data with their corresponding test sentence numbers. Balanced datasets are almost exploited to avoid the creation of models biased toward a specific class.

Table 1– Datasets Characteristics

Dataset	# Tweets	#Words	#Unique Words
D_1	4696	72579	18096
D_2	3727	37383	9786
Total	8423	109962	27882

As elaborated in Table 1, D_1 andD_2 are tokenized into 72579 and 37383unique attributes respectively. The corpus contains 109962words with about 27882 unique words. During all experiments, those attributes are considered as a baseline before applying any reduction methodology. The datasets are selected and annotated manually as positive and negative sentiment. The inter-annotator agreement in terms of Kappa (K) is measured. Kappa determines the quality of a set of annotations by evaluating the agreement between annotators. The message annotation task by the independent coders had a Kappa (k) value of 79.84% for D_1, 78.53% for D_2. It is worth mentioning that Weka must be run providing the utf-8 encoding to force correct handling for the Arabic language. Also, due to massive volumes of the dataset, Weka requires being executed with allocating considerably large amount of Java Virtual Machines Heap memory. All experiments were performed with the presence vectors. In each segment vector, the value of each dimension is binary regardless of how many times a feature occurs.

Several series of experiments were conducted to evaluate the effectiveness of the proposed methodology. First, performance of the classical BoW representation is measured against using of lexical, morphological and semantic analysis. Then, rule-based method for emotion detection is compared with ML approach. Finally, three experiments to measure the impact of FS methods on massive volume of social media datasets. A comparison is made among the BoW representation, reduced feature using CHI-Test and the reduced features using MRF algorithm. During all stages, the popular accuracy measures were examined.

4.1. Arabic Tweets Processing

The experiment is performed individually on two different stages. Firstly, dataset is tokenized and constructed into its initial vector (F_1) that represents BoW. Secondly, the underlying feature vector is passed through ICA then AWN to generate the candidate feature subset (F_2). For each testbed, SVM is conducted as the classification algorithm with different kernel types. Linear, Radial Basis Functions (RBF), and the sigmoidal kernels are used during experimentations. The results revealed that linear kernel promotes the best performance.

Moreover, Random Forest (RF) is tuned as another classification algorithm due to its superior performance over a single DT. RF constructs many DTs that will be used to classify a new instance by the majority vote. Each decision tree node uses a subset of attributes randomly selected from the whole original set of attributes. Additionally, each tree uses a different bootstrap sample data in the same manner as bagging. Classification scores using SVM and RF are tabulated in Table 2.

Table 2 - Precision, Recall and F-Measure Rates Using SVM& RF

Datasets	Feature set	# Feature	Precision%		Recall%		F-Measure%		Feature Reduction %
			SVM	RF	SVM	RF	SVM	RF	
D_1	F_1	18096	84.8	83.6	84.7	83.3	84.7	83.3	35.3
	F_2	11708	89	86.5	89	86.2	89	86.2	
D_2	F_1	9786	83.5	81.3	83.6	81.5	83.5	81.4	32.7
	F_2	6588	87.3	85.2	87.2	85.4	87.2	85.3	

The proposed method is effective in raising the F- Measure, which increased from 84.7% to 89% for D1. The method also raises F- Measure by 3.6% for D_2. On the other hand, F2 provides a considerable Dimension Reduction (DR) in comparison to BoW representations. For D_1, the feature vector was reduced from 18096 features to 11708 features with reduction percentage 35.3%. Whereas the feature vector for D_2 was diminished from 9786 to 6588 with reduction percentage 32.7%. These reductions are attributed to the fact that considerable numbers of tokens are converted to a common lemma by performing ICA analysis. In addition, AWN contributes this reduction by performing synonyms substitution. Therefore, the presented approach avoids the sparseness problem presented by word-based feature representations. As a result, F1 requires a huge amount of memory, CPU resource and extra time which perturbs the operation of the classifiers.

RF is experimented with a different number of trees to beget the highest accuracy. F-Measure was 83.3% and 81.4% at 10 trees for D_1 and D_2 respectively. The accuracy incrementally increased up to 86.2% for D_1 and 85.3 % for D_2 at 50 trees. At this point, F-Measure remains stable during the range from 50 to 120 trees and starts to slightly increase at 130 trees. In general, we can set the number of trees in a trial and error basis. Growing a larger forest will improve predictive accuracy, although there are usually diminishing returns once we get up to several hundreds of trees and only increase the computational cost. Increasing number of trees does not always mean the performance of the forest is significantly better than previous forests (fewer trees). In contrast, doubling the number of trees may be worthless. It is also possible to state there is a threshold beyond which there is no significant gain, unless a huge computational environment is available. In our case, increasing number of trees more than 130 trees do not achieve more significant performance gain, unless huge computational resources are available for large datasets.

4.2. Emotions

We evaluate our approach on 1136 tweets, which all contain emoticons. These tweets have been filtered out from varieties of corpora and are divided into two sets.The first set contains 508 tweets that have clear emotionsuch as (:-*,:*,♥,(y), (n), 3:), 3:-)). Customers used these emotions to stress the actual sentiment that already conveyed by the actual text. Such tweets can be classified directly by examining the rule set with ahigher classification accuracy reaches 99%.The second set contains 628 tweets that have mixed emotions where the emotion of customer is not the same as the opinion conveyed in the text. In this case, emoticons can be considered as ambiguous symbols that introduce noise by carrying contradict opinions. Rules are not sufficientto classify such tweets. So, the underlying tweets are tokenized into 2364 unique tokens. Then, all verbal information with the emotional labels are classified using SVM. Accuracies greater than 88 % were reported for label prediction.

4.3. FS Techniques

Here, we examine the performance of MRF against the performance of the CHI-Test algorithm. Chi Square has proved to record high accuracy in classifying both English [30] and Arabic text [18]. The result achieved by BoW (F_1) was used as a baseline. F_2 represents dataset after performing CHI-Test and F_3denotes dataset after reduction by MRF. As elaborated in Table 3,F_1 achieved 84.7% for F-Measure. CHI-Test slightly alters the quality of the results, instead of bringing up an added value. However, the classification accuracy for F_2 drops by 2.4 % compared to BoW. Finally,F_3 has the highestaccuracy and achieves 88.3% for F- Measure.

Table 3- Accuracies for all modules

Classifier	SVM		
Metrics	Precision%	Recall %	F-Measure%
F_1	84.8	84.7	84.7
F_2	82.6	82.5	82.4
F_3	88.6	88.3	88.3

As can be seen from the Table 3, the highest value of precision was achieved for F3 that register 88.6 %. The worst recall value was obtained in the case of F_2 while the highest kvalue was for F_3. In simple terms, high precision of MRF means that the algorithm returned substantially more relevant results than irrelevant ones. A high recall of MRF means that it returned most of the relevant results. As a result, CHI-Test (F3) has the lowest performance in terms of precision, recall, and F-Measure.

Moreover, F1 has the highest number of features as can be seen from Figure 3

Figure . Chi statistics successfully minimizes attributes to a very high extent reaches 66.2%. Although, CHI-Test achieves the highest reduction percentage, it adversely affects the accuracy and causes information loss. On the other hand, MRF reduces the feature space by 46.7% and simultaneously raises the classification measures by 3.6%.

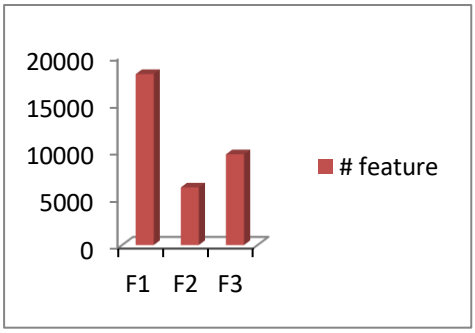

Figure 3 – Attribute Reduction Percentage

Attribute weighting time is proportional to the number of attributes in a given dataset. The more features the higher the preprocessing time. Figure 4 illustrates the runtimes (hours) for the mode formation in the three different stages. It is very interesting to note the big gap between the time

needed to compute the initial feature set (F_1) and the various FS methods. Obviously, the training time sharply increased with the larger sample sizes.

The lowest preprocessing time was achieved in the case of MRF when compared with the other two vector types.

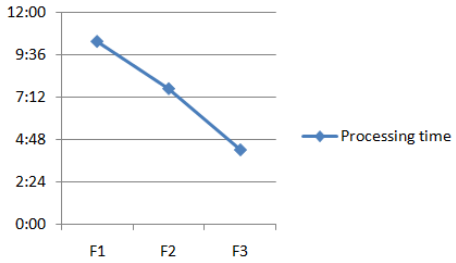

Figure 4 - Processing Time for All Modes

Therefore, FS is a preprocessing step carried out prior to classification where the classifiers can be designed in an easy way to compute. However while doing so, it must also retain the accuracy and must not lead to loss of information. In words, using all features without applying any reduction methodologies result in the classifier being too sensitive to the minor syntactical variations of the same feature. Therefore, these features are considered not correlated and adversely affect the exactness of the classifier.

5. CONCLUSION

The study presents a method to evaluate Arabic opinions and measure customers' loyalty about certain popular characters, companies and products. The results revealed that incorporation of morphological characteristics, semantic knowledge and emotional state description in feature vector outperformed the classical BoW representation. We also developed different types of lexicons and define a set of new rules to classify the Arabic text automatically. Such a system has the ability to rapidly adapt to new domains with minimal supervision. For FS, experimentsconfirm thatMRFoutperformsCHI-Test in terms of classification accuracy and running time. To further enhance the readability of the mined segments, we examined the results for reflecting positive or negative opinions of customer about different product and services. For example, if we interested to comprehend the latent reasons for increasing positive opinion about "Vodafone" at some point; we can analyze the tweets with positive sentiment in the changing period and mirror the revolving events with these positive opinions.

We believe that the accuracy should still be improved. Semantics will help in this direction. The polarity of a tweet may depend on the perspective you are interpreting the tweet from. For example, in the tweet "موبينيل تسبق فودافون فى سرعة النت", the sentiment is positive for " موبينيل" and negative for "فودافون". Using a semantic role labeler may indicate which noun is mainly associated with the verb and correctly the classification would take place. This may allow "فودافون تسبق موبينيل فى سرعة النت"to be classified differently from "موبينيل تسبق فودافون فى سرعة النت"

REFERENCES

[1] Yingcai Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, HuaminQu, "OpinionSeer: Interactive Visualization of Hotel Customer Feedback" IEEE Trans. Visualization and Computer Graphics. Vol. 16, No. 6, pp. 1109 – 1118, 2010.

[2] Erik Cambria, BjörnSchuller, Yunqing Xia, Catherine Havasi "New Avenues in Opinion Mining and Sentiment Analysis" IEEE Intelligent System, Vol.28, No. 2, pp. 15 – 21, 2013.

[3] Anuj Sharma, ShubhamoyDey "Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis" Special Issue of I.J. of Computer Applications on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, pp. 15- 20, June 2012.

[4] Bashar Al Shboul, Mahmoud Al-Ayyoub and YaserJararweh "Multi-Way Sentiment Classification of Arabic Reviews" 2015 6th Int. Conf. On Information and Communication Systems (ICICS). IEEE 2015.

[5] Hossam S. Ibrahim, Sherif M. Abdou, MervatGheith "Sentiment Analysis for Modern Standard Arabic and Colloquial" I. J. on Natural Language Computing (IJNLC), Vol. 4, No.2, pp. 95 – 109, April 2015.

[6] Walaa M., Ahmed H., Hoda K. "Sentiment analysis algorithms and applications: A survey" Ain Shams Engineering Journal, ElSevier, Vol. 5, No. 4, pp. 1093–1113, Dec 2014.

[7] Nizar A. Ahmed, Mohammed A. Shehab, Mahmoud Al-Ayyoub and Ismail Hmeidi "Scalable Multi-Label Arabic Text Classification" 6th Int. Conf. On Information and Communication Systems (ICICS), IEEE, pp. 212 – 217, 2015.

[8] A. Abbasi, H. Chen, A. Salem "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums" ACM Trans. Information Systems, Vol. 26, No. 3, Article 12, June 2008.

[9] Muhammad Abdul-Mageed, Mona Diab, Sandra Kübler "SAMAR: Subjectivity and sentiment analysis for Arabic social media" Computer Speech & Language. ACM, Vol. 28, No. 1, pp. 20–37, Jan 2014.

[10] Fawaz S. Al-Anzi, Dia Abu Zeina "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing" Elsevier, April 2016.

[11] Daekook Kang, Yongtae Park, "Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach", Expert Systems with Applications, pp. 1041–1050, Elsevier, 2014.

[12] FaragSaad, Brigitte Mathiak "Revised Mutual Information Approach for German Text Sentiment Classification", Int. World Wide Web Conference Committee (IW3C2), Rio de Janeiro, Brazil, ACM, pp. 579 – 586, May 2013.

[13] Minqing Hu, Bing Liu, "Mining and Summarizing Customer Reviews" KDD"04 ACM, Seattle, Washington, USA, ACM, Aug 2004.

[14] Marcelo Drudi Miranda, Renato José Sassi "Using Sentiment Analysis to Assess Customer Satisfaction in an Online Job Search Company" Springer Int. Publishing Switzerland, BIS 2014 Workshops, LNBIP 183, pp. 17–27, 2014.

[15] Yingcai Wu, Shixia Liu, Kai Yan, Mengchen Liu, Fangzhao Wu "OpinionFlow: Visual Analysis of Opinion Diffusion on Social Media" IEEE Trans. Visualization and Computer Graphics, Vol. 20, No. 12, pp. 1763 – 1772, Dec 2014.

[16] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen, Xiaofei He "Interpreting the Public Sentiment Variations on Twitter" IEEE Trans. Knowledge and Data Engineering, Vol. 6, No. 1, pp. 1 – 14, Sep 2012.

[17] Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, YaserJararweh, Mohammed N. Al-Kabib "A comprehensive survey of Arabic sentiment analysis" Elsevier, Sep. 2018.

[18] Mohammad S. Khorsheed, Abdulmohsen O. Al-Thubaity, "Comparative evaluation of text classification techniques using a large diverse Arabic dataset," Language Resources and Evaluation, Springer, Vol. 47, No. 2, pp. 513 – 538, March 2013.

[19] ViolettaCavalli-Sforza, Hind Saddiki, KarimBouzoubaa, LahsenAbouenour "Bootstrapping a WordNet for an Arabic Dialect from Other WordNets and Dictionary Resources" 10th ACS/IEEE Int. Conf. On Computer Systems and Applications (AICCSA 2013), Fes/Ifrane, Morocco, May 2013.

[20] ZakariaElberrichi, KarimaAbidi "Arabic Text Categorization: a Comparative Study of Different Representation Modes" Int. Arab Journal of Information Technology, Vol. 9, No. 5, pp. 465 – 470, Sep 2012.

[21] SamehAlansary, MagdyNagi "The International Corpus of Arabic: Compilation, Analysis and Evaluation" Proc. of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 8–17, Doha, Qatar, Oct. 2014.

[22] Keng-Pei Lin, Ming-Syan Chen "On the Design and Analysis of the Privacy-Preserving SVM Classifier" IEEE Trans. Knowledge and Data Engineering, Vol. 23, No. 11, pp. 1704 - 1717, Nov 2011.

[23] Le Zhang, PonnuthuraiNagaratnamSuganthan "Random Forests with ensemble of feature spaces" Pattern Recognition, Elsevier, pp. 3429–3437, 2014.

[24] NizarHabash, Owen Rambow "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop" In Proc. of the 43rd annual Meeting of the Association for Computational Linguistics (ACL05), pp. 573-580, June 2005.

[25] JaberAlwedyan; Wa'el Musa Hadi; Ma'an Salam; Hussein Y.Mansour, "Categorize Arabic Data Sets Using Multi-Class Classification Based on Association Rule Approach", Int. Conf. On Intelligent Semantic Web-Services and Applications (ISWSA"11), Amman, Jordan, ACM, April 2011.

[26] Ali Farghaly, KhaledShaalan "Arabic Natural Language Processing: Challenges and Solutions" ACM Trans. Asian Language Information Processing, Vol. 8, No. 4, Article 14, pp. 14 – 22, Dec 2009.

[27] A. Alajmi, E. M. Saad, R. R. Darwish "Toward an Arabic Stop-Words List Generation" I.J. of Computer Applications, Vol. 46, No.8, pp. 8 – 13, May 2012.

[28] RamyBaly, Georges El-Khoury, RawanMoukalled, RitaAoun, HazemHajj, KhaledBashirShaban, WassimEl-Hajj "Comparative Evaluation of Sentiment Analysis Methods Across Arabic Dialects", 3rd International Conference on Arabic Computational Linguistics, ACLing. in Arabic Computational Linguistics. Elsevier, Vol. 117, Pp. 266-273, Nov 2017.

[29] Daniel Jurafsky, James H. Martin "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition". 2ndEdition. Prentice-Hall, 2009.

[30] Joseph D. Prusa, Taghi M. Khoshgoftaar, David J. Dittman "Impact of Feature Selection Techniques for Tweet Sentiment Classification" Proc. of the 28th Int. Florida Artificial Intelligence Research Society Conference. pp. 299 – 304, 2015.

[31] Emad E. Abdallah , Sarah A. Abo-Suaileek "Feature-based Sentiment Analysis for Slang Arabic Text " I. J. of Advanced Computer Science and Applications (IJACSA), Vol. 10, No. 4,pp. 298-304, 2019.