COMBINING MACHINE LEARNING AND SEMANTIC ANALYSIS FOR EFFICIENT MISINFORMATION DETECTION OF ARABIC COVID-19 TWEETS

Abdulrahim Alhaizaey and Jawad Berri

Department of Information Systems, King Saud University, Riyadh, Saudi Arabia

ABSTRACT

With the spread of social media platforms and the proliferation of misleading news, misinformation detection within microblogging platforms has become a real challenge. During the Covid-19 pandemic, many fake news and rumors were broadcasted and shared daily on social media. In order to filter out these fake news, many works have been done on misinformation detection using machine learning and sentiment analysis in the English language. However, misinformation detection research in the Arabic language on social media is limited. This paper introduces a misinformation verification system for Arabic COVID-19 related news using an Arabic rumors dataset on Twitter. We explored the dataset and prepared it using multiple phases of preprocessing techniques before applying different machine learning classification algorithms combined with a semantic analysis method. The model was applied on 3.6k annotated tweets achieving 93% best overall accuracy of the model in detecting misinformation. We further build another dataset of Covid-19 related claims in Arabic to examine how our model performs with this new set of claims. Results show that the combination of machine learning techniques and linguistic analysis achieves the best scores reaching 92% best accuracy in detecting the veracity of sentences of the new dataset.

KEYWORDS

Misinformation, machine learning, Arabic NLP, contextual exploration, rumor detection.

1. INTRODUCTION

In recent years, the use of social media platforms has increased considerably, imposing itself as one of the most important sources for spreading news and broadcasting opinions. Social media platforms have become a rich source of information and have provided many researchers with an opportunity to analyze the vast amount of data that users of these platforms post on a daily basis. Within the Arab world, a recent study indicated that social media platforms had become the dominant source of news for Arab youth [1]. Twitter[1] is one of the most important of these platforms, as it is the second most used platform after Facebook[2] by Arabic speakers [2]. The Arabic language is the sixth largest language present on Twitter [3], making it an important source for data mining research and analysis.

With the emergence of the Covid-19 pandemic, Twitter, like many other social media platforms, has witnessed an increased interaction by people in expressing opinions and discussions related to the Coronavirus. Regardless of the different measures enforced by governments to contain the spread of this virus, the spread of misleading news about the Coronavirus represents a unique

[1] https://twitter.com/home
[2] https://www.facebook.com/

phenomenon, requiring many decision-makers to hold daily conferences to answer and refute many widespread fake news and rumors. Moreover, with millions of daily posted tweets and replays, it becomes nearly impossible for reliable fact-checking services to keep up checking claims with such a massive amount of data. We argue that data mining and machine learning tools can take some of the burden off fact-checking platforms that fundamentally rely on the role of experts in checking and refuting rumors and misleading news. A broad range of literature has studied claim verification using machine learning tools in the English language online content. Many new recent studies address the problem of Covid-19 rumors detection in English online content such as [4] and [5]. However, very few studies have addressed this problem in Arabic online content. For example, [6] suggested a model for rumor detection in Arabic tweets using semi-supervised and unsupervised expectation-maximization. Another study [7] introduced an Arabic corpus of fake news targeting YouTube. However, both of these works did not address the misinformation detection related to the Coronavirus pandemic in Arabic language. In general, studies on the verification of Arabic language claims and rumors, particularly those related to the pandemics, are rarely explored. To the best of our knowledge, this is one of the very few studies that address this topic.

This paper provides a model that uses machine learning tools combined with a semantic analysis method to detect misleading news related to the Coronavirus pandemic. We used the ArCOV19-Rumors dataset [8] in our study, which is an Arabic COVID-19 related dataset retrieved from Twitter. The dataset ArCOV19-Rumors is the only available Arabic dataset to support claims verification studies. In addition, the authors who released this dataset have intensively annotated the veracity of the tweets. This intensive annotation makes it a good resource to support research on misinformation detection as one of the major problems encountered during the pandemic. Furthermore, the dataset is balanced in terms of the distribution of true and false tweets, making it a sound resource for building and training a verification system. The dataset we used was analyzed using Sentiment Analysis approaches to provide instant detection of fake news. Despite that this model could arguably be a more affordable alternative than relying on expert editors or journalists, this model could also help in immediately notifying those who are interested in fact-checking such as professional journalists and editors running fact-checking services. We used the Ar-Covid-19 dataset [8] in our study. This dataset contains a total of 1831 Arabic tweets with a valid claim and 1660 Arabic tweets with a false claim related to the Coronavirus during the period from $27th$ January till the end of April 2020. Table 1. shows a statistical summary of the labeled tweets in ArCOV19-Rumors dataset. Based on this labeled dataset, we developed a claim verification system taking advantage of the capabilities provided by Python programming language with a suite of libraries for natural language processing and machine learning. Our model consists of multiple components: data collection, data preprocessing and cleansing, and machine learning classification. In the next section, we provide a background of the related works. Then, in Section 3, we present the methodology. Section 4 presents the results and discussion. Section 5 shows how the accuracy of the classifiers can be enhanced on our testing set of claims after applying some sematic based rules. The conclusion of the work is then provided.

Table 1. Statistics of Claims in ArCOV19-Rumors [8].

Label	**Number of tweets**	**Percentage**
False Tweets	1753	(48.9%)
True Tweets	1831	(51.1%)
Total	3,584	(100%)

2. LITERATURE REVIEW

In the internet age, the rise of fake news highlights the urgent need for organizational solutions for such a challenging problem. Fake news and manipulating people for different reasons have a long history and used to be addressed by journalistic norms of objectivity; however, social networks have created a context in which fake news can attract a mass audience [9]. With the emerging of Covid-19, fake news and claims spreading on social media platforms are immensely remarkable. With the increased transformation to social media as the new news sources, the traditional ones that had enjoyed good levels of public trust and reliability have lost their popularity. Hence, a new way of verifying news on social media is needed. Fact-checking websites can be seen as an attempt to alleviate this complex problem in the internet world. Very few Arabic platforms provide Fact-Checking services to evaluate factual claims of news stories in Arabic. For example, Misbar[3], AFP[4], and fatabyyano[5] are digital services that have been launched for the purpose of news verification. These fact-checking services rely heavily on their networks of expert editors and journalists to run their fact-check services. However, despite that the use of these fact-checking services is still not measured, the effectiveness of such services in the Arabic world needs investigation.

There is a broad body of literature that would suggest that humans are not good at predicting and decision making. For that reason, machine learning has been widely used in many real-life applications to improve and enhance human decision-making. For example, in August 2016, the Allegheny County Department of Human Services implemented a machine learning-based predictive risk modeling tool called AFST [10]. This tool was explicitly designed to improve child welfare call screening decisions. The tool was the result of a two-year process of exploration about how existing administrative data could be used more efficiently to improve decision-making at the time of a child welfare referral. Facebook also uses machine learning algorithms to identify and remove sensitive content such as violence and sexual contents automatically. In addition, realizing that it is impossible for human fact-checkers to review stories on an individual basis with around two billion Facebook registered users, Facebook recently started to use similar technologies to recognize false news and take action on a bigger scale [11]. Likewise, we believe that using the state of the art machine learning and artificial intelligence techniques could be an effective tool in combating misinformation in the near future. Instantaneous insights into fake news and rumors in social media could be done using sentimental analysis (SA) and machine learning (ML) techniques. SA is a methodology used for natural language processing. It is used to analyze human opinions, emotions, sentiments, and attitudes. The noticeable increase in employing SA techniques in opinion mining has been co-occurred with the increased use of social media, benefiting from the huge amounts of available data in social media platforms. SA can be implemented using machine learning models. So, could we take advantage of such techniques and methods to detect misinformation in social media? This is what we are trying to do in this research. We apply SA using ML to Ar-Rumor dataset [8] to see to what extent these techniques can be beneficial in detecting Arabic fake news. The main contribution of this work is, in one aspect, that it shows the feasibility of automating the detection of misinformation on the microblogging platforms such as Twitter. There must be automated systems for such purposes, given the uncontrollable vast amount of content and users. The second aspect is that misinformation detection in Arabic is rarely explored, which this paper has touched upon. The third aspect is the combination between machine learning techniques and semantic analysis which turned out to produce better results in detecting misinformation. The following section presents our methodology for tackling such a problem.

[3] https://misbar.com/
[4] https://factuel.afp.com/ar/list
[5] https://fatabyyano.net

3. METHODOLOGY

Once we retrieved and collected the dataset, we build our claim verification system, which contains multiple components. The first part is data preprocessing and normalization using text preprocessing techniques followed by an additional layer of data filtering and feature extraction. The next component is the classification using different machine learning classifiers. Our model is depicted in Figure 1. The details of the model components are described in the following paragraphs.

3.1. Retrieving Dataset

We retrieved the dataset using Hydrator [12], a free tool for hydrating Twitter ID datasets. We collected 9.4k labeled relevant tweets. A total of 3.6k of the collected tweets are labeled either true or false. A total of 1,753 tweets were labeled false, and a total of 1,831 tweets were labeled true. The distribution of true tweets versus false tweets is balanced, making it a good resource for training verification systems.

3.2. Data Preprocessing and Normalization

After collecting the Twitter dataset, the data were prepared for analysis. Due to the noisy nature of the microblog data such as Twitter tweets, text preprocessing and cleaning is an essential step to eliminate noise and unnecessary data that do not represent any value about the sentiment of the text. Data preprocessing plays a significant role in improving the accuracy of classification algorithms and speeding up the training process to enhance the models' learning efficiency [13]. For this purpose, we used a natural language toolkit (NLTK) [14] of Python, which has a suite of text processing libraries.

Preprocessing and cleansing include the following steps: removing repeated letters, removing stop words, which are provided in [14], removing words that have less than two letters, removing numbers and additional spaces, removing any non-Arabic letters, removing punctuation marks and emojis, removing Arabic diacritics, and normalizing the data by converting multiple forms of a letter into one uniform letter since some Arabic letters could be represented in different forms.

3.3. Additional Filtering

Besides eliminating noise data, and since the dataset might contain duplicate tweets and what could be considered spam tweets, we performed an additional filtering step. In this step, tweets with URLs, phone numbers, more than four hashtags, and duplicated tweets are classified as spam, so we deleted them. Then, we removed hashtags, URLs, and mentions. After that, the duplicated tweets are dropped to keep only one representation of each tweet. Finally, after preprocessing and cleaning, some tweets resulted in a very short text of only a few words, so we removed any tweet with less than two words because we believe such tweets do not contribute towards the classification and need to be removed.

3.4. Feature Extraction

Predictive modeling requires text data to be customized and prepared before using it for prediction purposes. The textual data have to be parsed and tokenized such that the words are converted into numbers or to be used as input for the machine, which is referred to as vectorization. For this purpose, the scikit-learn library [15] provides convenient tools to perform feature extraction of text data. With this tool, we used a popular technique called "Term

Frequency and Inverse Document Frequency" (TF-IDF) to represent the data as a feature where words are given frequency scores. This helps in determining the words that are more interesting for classification. Also, Arabic stop words imported from NLTK [14] were used to drop specific words from data. These words do not contribute towards classification.

3.5. Machine Learning Classification

Many supervised ML classifiers have been widely used in the literature for text analysis. We have used four different common ML algorithms to analyze and compare the results, namely: Logistic Regression (LR) [16], Random Forest (RF) [17], Multinomial Naive Bayes (MNB) [18], and Support Vector Machine (SVM) [19]. Although many ML classifiers are available in text classification problems, we adopted the four mentioned classifiers because these classifiers have been widely applied and used in this context. The ones involved here are just chosen as a proof of concept, although other classifiers could be chosen. These four ML algorithms have been popularly used for text and sentiment analysis, so we found them proper to use in our study context. After preprocessing data, we supplied the dataset to different ML algorithms to build a classifier that learns from training labeled tweets. In order to build our classifier, we used a Scikit-learn library, an open-source machine learning package in Python that provides various classification algorithms. We have randomly set our ML algorithms to choose 20% of the dataset as test sets. Then we have used the remaining 80% of the dataset to train our machine learning models. The trained models can then predict whether another claim from the test set is true or false claim. In our analysis, the dataset we used was classified into two classes, true and false.

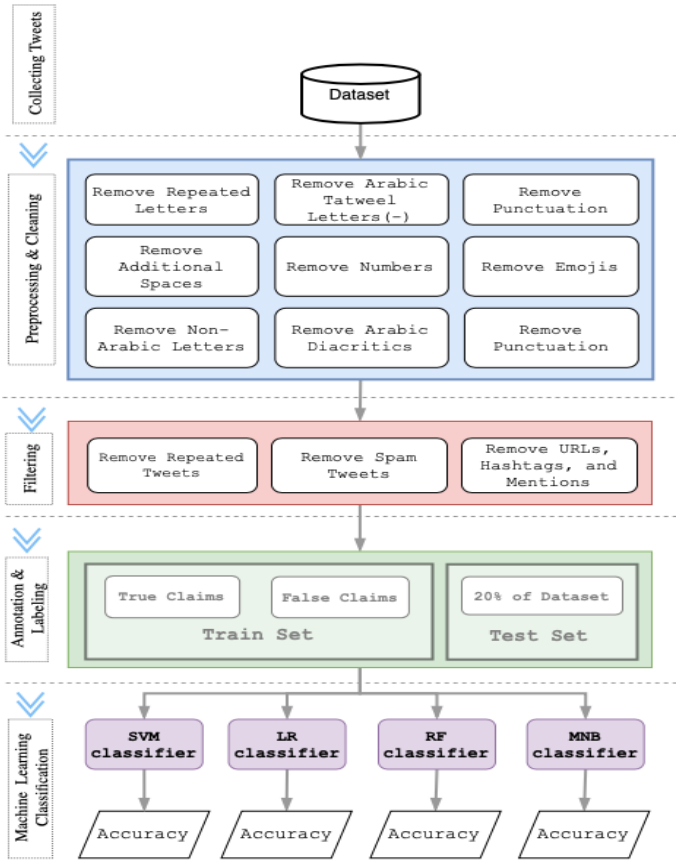

Figure 1. Verification Model Architecture

4. EXPERIMENT AND RESULTS ANALYSIS

After building our verification prediction model, we have carried out our experiments to test and evaluate the accuracy of this model. In order to evaluate our model, we used the accuracy metric (1), which is a common metric to measure performance in machine learning classification and pattern recognition. A true positive (TP) is an outcome where the machine learning model correctly predicts the true claim. Likewise, a true negative (TN) is an outcome where the model correctly predicts the false claim. Similarly, a false positive (FP) is a result where the model incorrectly predicts the true class. And a false negative (FN) is a result where the model incorrectly predicts the false claim. We further build another dataset of Covid-19 related claims in Arabic to examine how our model performs with this new set of claims. Figure 2 shows the analysis of the results of our model for the two datasets and the machine learning models chosen in the experiments.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Our first observation on the accuracy of our model is that the SVM model achieved 93% accuracy compared to the remaining machine learning approaches, where the accuracy of LR, MNB, RF are 91%, 91%, and 89%, respectively. On the other hand, when testing our model against our dataset to measure its performance, we see that our current system performs well for the new dataset that we built for the purpose of measuring the performance of the model. Our accuracy analysis shows that our dataset results in a bit lower accuracy. The accuracy of the machine learning algorithms in detecting veracity in our dataset is 79%, 75%, 85%, and 87% for LR, RF, MNB, and SVM approaches, respectively. It can be observed that SVM and MNB significantly outperform LR and RF in our dataset. This is because our dataset is relatively small. In addition, the proper learning of our system with more data can increase the accuracy of the system as a whole.

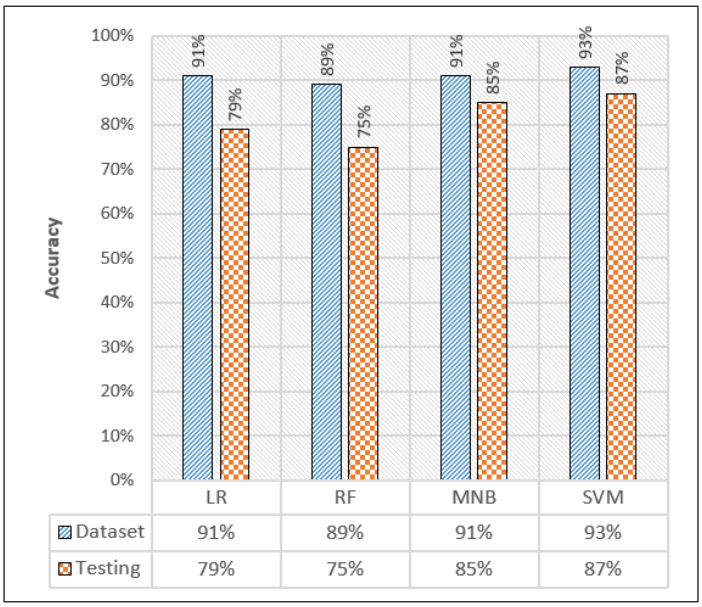

Figure 2. Results Analysis of the two datasets

5. SEMANTIC TWEETS ANALYSIS

In order to improve the accuracy of the results obtained in the previous step with machine learning algorithms, we used a linguistic analysis method which aim to analyze the results of the learning algorithms semantically. Machine learning algorithms are efficient in dealing with huge data and can learn and improve. However, these algorithms may not classify the tweets in the dataset correctly when it comes to very fine semantic analysis that is necessary to define the meaning of sentences. On the other hand, semantic analysis lacks performance when dealing with big datasets but is efficient in discovering fine semantic nuances in the meaning of tweets which can falsify the results. A combination of both methods is crucial to improve the accuracy of misinformation detection in the dataset used in the experiment. Semantic analysis should explore the tweet linguistic context to be able to detect the right semantic value of the tweet. We used a linguistic method, namely the Contextual Exploration (CE) method [22], which can analyze the context and take the right decision for labeling sentences semantically. CE is used here to compute the semantic value representing the meaning of a sentence. CE does not require huge linguistic resources and fine semantic descriptions of language units. Instead, it requires only the linguistic grammatical and lexical know-how used by a decision maker when assigning the semantic value to a sentence [23]. A close look at the dataset shows that many linguistic markers can falsely classify tweets' semantic value. We focused in this research on four semantic labels, namely: the *interrogation*, the *negation*, the *claim* and the *reported statement*. Table 2 provides illustrative examples of tweets from the dataset. Using interrogation sentences is common in news headlines to grab the audience's attention. On Twitter, it is usually posted with an attached URL to the news webpage. Interrogation is expressed in Arabic by using the interrogation mark "؟" such as in the tweet example S1. This sentence questions the veracity of the Hanta virus that appeared in China without stating any truth value. CE detects these sentences by the interrogation mark and accordingly labels them as an interrogation sentence and reclassify them as Neutral. Negation in standard Arabic can be expressed in different ways depending on the verb tense and using the five negation forms: لا، لم، ما، لن، ليس. For instance, sentence S2 in Table 2 is labeled as a negation sentence since it includes the word "لا". These sentences are reclassified to the opposite class assigned by the classifier.

Table 2. Examples of Tweet sentences.

Text	Sentence (Tweet)
S1	بعد كورونا ظهور فيروس هانتا في الصين... ما الحقيقة؟ *After Corona, the emergence of the Hanta virus in China ... What is the truth?*
S2	"S Gene not detected" لا يعني أنك بمنأى عن كوفيد-19! *"S Gene not detected" doesn't mean you are immune to COVID-19!*
S3	فيديو يدعي أن: لقاح كورونا موجود حتى قبل بدأ جائحة كورونا (...) *Video claims that: Corona vaccine existed even before the Corona pandemic began (...)*
S4	صحيفة ديلي ميل البريطانية: لن يعود الدوري الإنجليزي مرة أخرى هذا الموسم (...) *The British Daily Mail: The English Premier League will not resume this season (...)*

A claim is a statement conveyed to assert something either by the speaker or it is reported as a claim of another party such in sentence S3, where the claim is reported from a video (third person) using the verb "يدعي" (claims) and the colon mark (:) which is explicitly used in the sentence. In general, when the speaker claims something (by using the pronoun I, or we) this means that the claim is most likely definite and hence it is labeled as a positive sentence. On the other hand, when the claim is stated by another party (not the speaker) the sentence is classified as neutral. A reported statement is frequently used by news reporters who convey news statements from other news agencies or information sources such in sentence S4. In such case the sentence is labeled as neutral since the veracity of the reported sentence is referred to the claimer

without any responsibility on the speaker. It is worth mentioning that some sentences can be associated with more than one semantic label. For example, sentence S4 in Table 3 is labeled as a negation claim and a conveyed claim. For such sentences another semantic analysis stage is necessary to assign dominant semantic label and then reclassify it. Sentences falling in these four linguistic categories in the test dataset are handled properly and reclassified during the post-semantic analysis. Then the accuracy of the dataset is recalculated.

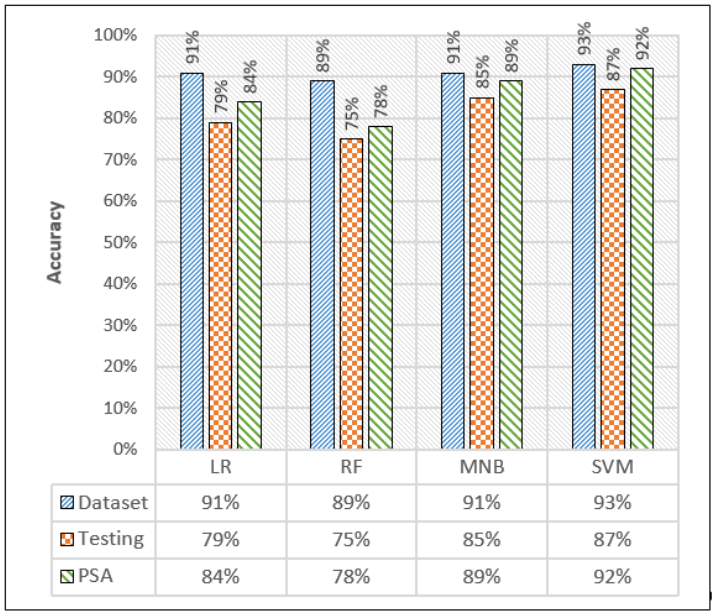

Figure 3. Results with Post Semantic Analysis

Figure 3 shows the results after the post-semantic analysis. The post semantic analysis has clearly improved accuracy for all algorithms for the test dataset. The proposed approach, which combines machine learning algorithms with post semantic analysis has better accuracy then the machine learning approach alone. Although the linguistic approach currently uses only four semantic labels in the classification process, it may be further enhanced by including additional semantic labels to capture the fine semantic meanings carried by natural language sentences.

6. DISCUSSION

Social media networks have introduced a new era for the public to share ideas, opinions, and debates. These shared data establish a significant source to detect opinion trends regarding different topics. In Parallel, different methods are becoming efficient in analyzing attitudes of complex data in social media. For example, artificial intelligence, sentiment analysis, data mining, and machine learning would help by providing different means of information extraction and knowledge acquisition regarding different topics. With people's tendency to exchange and broadcast their opinions on social media platforms, and with the emergence of the COVID-19 pandemic, many legislative and health authorities have found themselves in the face of pandemic-related rumors. These rumors affect and slow down the efforts to combat the pandemic, as it consumes much time and effort to refute them. For that scenario, researchers and decision-makers can benefit from this data in analyzing and studying several social, behavioral, or organizational phenomena. Unlike the traditional knowledge management practices that rely on human experts as a primary source of knowledge acquisition, harvesting and analyzing social media data with the help of machine learning could represent the ultimate source from which the knowledge is extracted. Harvesting and mining the COVID-19 Twitter data could provide

innovative insights and ideas on how to better manage the implications and apprehensions associated with pandemic-related rumors. Such studies could provide some roadmap for potential future alternative sources of knowledge and suggest future machine learning and system involvement. In addition, media practitioners can benefit from such affordable technical solutions to automate the process of refuting rumors and misinformation.

Figure 4. False tweets word clouds

Figure 5. True tweets word clouds

In terms of results accuracy, the experiment shows that combining machine learning techniques with semantic analysis can increase accuracy in detecting rumors in a dataset and avoid claims that may false classify tweets. Text analytics could be enhanced using data visualization techniques such as words cloud. Though a simple concept, word clouds are a good visualization technique to communicate an overall picture of text contents [20]. A study [21] concluded that word clouds are certainly an effective tool for text analysis if enhanced with further information and powerful interaction techniques. Figure 4 and Figure 5 display word clouds generated from the dataset [8] based on its class label. Yet, there are some limitations of this work. First, the focus of this work was on Twitter content only due to the availability of the annotated dataset. However, Twitter is not the only source that misinformation can spread. Second, the dataset is relatively small, which impacts the splitting decision of the training and testing sets. Thus, the proper and sufficient learning dataset of the system with more data can increase the system's accuracy as a whole.

7. CONCLUSION

Analyzing Twitter news data using natural language processing, sentimental analysis, machine learning, and artificial intelligence techniques could serve as critical tools to help protect people from misinformation. With millions of contents posted every day on social media platforms, relying on human fact-checkers to review stories on an individual basis is almost impossible. Alternatively, there is an urgent need for a new way to quickly detect posts that may contain false claims and automatically send them to fact-checkers to focus their time and expertise on fact-checking new content. English text analysis had had a great deal of work on detecting rumors. However, in Arabic, this field is still in an early stage where the lack of resources and language nature confront huge challenges to the research. This study presents an Arabic misinformation detection model that mines related Arabic texts from Twitter to detect fake news. For this study, we preprocessed a dataset of 3.6k tweets related to Covid-19 claims. This dataset was then fed into different machine learning classifiers to measure classifier accuracy in detecting misinformation. The results showed that the best overall accuracy of the model in detecting misinformation was 93%. Another dataset of Covid-19 related claims in Arabic was built to examine the performance of our model. The results showed that the model's overall accuracy in detecting the veracity of the new dataset was 87% at best. The results provided by the classifiers

were analyzed further by a semantic analysis method to detect neutral sentences which can false classification of tweets. The results show that this post semantic analysis phase has improved accuracy of the results of the top result in the first phase to 92%.

REFERENCES

[1] Radcliffe and H. Abuhmaid, "Social media in the middle east: 2019 in review," *SSRN Electron. J.*, 2020.

[2] ExtraDigital Ltd, www. extradigital. co. uk, "Prominent Arabic Social Media," *Extradigital.co.uk*. [Online]. Available: https://www.extradigital.co.uk/articles/arabic/social-media.html. [Accessed: 02-May-2021].

[3] "Twitter: most-used languages 2013," *Statista.com*. [Online]. Available: https://www.statista.com/statistics/267129/most-used-languages-on-twitter/. [Accessed: 02-May-2021].

[4] Alam *et al.*, "Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society," *arXiv [cs.CL]*, 2020.

[5] P. Patwa *et al.*, "Fighting an Infodemic: COVID-19 Fake News Dataset," *arXiv [cs.CL]*, 2020.

[6] S. M. Alzanin and A. M. Azmi, "Rumor detection in Arabic tweets using semi-supervised and unsupervised expectation–maximization," *Knowl. Based Syst.*, vol. 185, no. 104945, p. 104945, 2019.

[7] M. Alkhair, K. Meftouh, K. Smaïli, and N. Othman, "An Arabic corpus of fake news: Collection, analysis and classification," in *Communications in Computer and Information Science*, Cham: Springer International Publishing, 2019, pp. 292–302.

[8] Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV19-Rumors: Arabic COVID-19 Twitter dataset for misinformation detection," *arXiv [cs.CL]*, 2020.

[9] M. J. Lazer *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[10] R. Vaithianathan, N. Jiang, T. Maloney, P. Nand, and E. Putnam-Hornstein, *Developing predictive risk models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation [PDF*. Auckland: Centre for Social Data Analytics, 2017.

[11] "Working to stop misinformation and false news," *Facebook.com*. [Online]. Available: https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news. [Accessed: 02-May-2021].

[12] Documenting the Now, *Hydrator [Computer Software]. Retrieved from https://github.com/docnow/hydrator*. Accessed March, 2021.

[13] S. Larabi Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali, and I. Abunadi, "Arabic natural language processing and machine learning-based systems," *IEEE Access*, vol. 7, pp. 7011–7020, 2019.

[14] S. Bird, E. Klein, and E. Loper, *Natural language processing with python: Analyzing text with the natural language toolkit*. O'Reilly Media, 2009.

[15] Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *arXiv [cs.LG]*, 2012.

[16] I. Webb *et al.*, "Logistic Regression," in *Encyclopedia of Machine Learning*, Boston, MA: Springer US, 2011, pp. 631–631.

[17] M. D. Buhmann *et al.*, "Random Forests," in *Encyclopedia of Machine Learning*, Boston, MA: Springer US, 2011, pp. 828–828.

[18] I. Webb, E. Keogh, R. Miikkulainen, R. Miikkulainen, and M. Sebag, "Naïve Bayes," in *Encyclopedia of Machine Learning*, Boston, MA: Springer US, 2011, pp. 713–714.

[19] Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[20] B. Y.-L. Kuo, T. Hentrich, B. M. Good, and M. D. Wilkinson, "Tag clouds for summarizing web search results," in *Proceedings of the 16th international conference on World Wide Web - WWW '07*, 2007.

[21] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," in *47th Hawaii International Conference on System Sciences*, 2014.

[22] J. Berri, M. Al-Khamis, Information Exploration Using Mobile Agents, *WSEAS Transactions on Computers*, vol. 3, no. 3, 706-712, 2004.

[23] J. Berri, R. Benlamri, Y Atif, H. Khallouki, Web Hypermedia Resources Reuse and Integration for On-Demand M-Learning, *International Journal of Computer Science and Network Security*, vol. 21, no. 1, pp. 125-136, 2021