Arabic Hands-on Analysis, Clustering and Classification of Large Arabic Twitter Data set on Covid-19

Abdelrahman Hamdy [*], Ayman Youssef [1, 2], Conor Ryan [2]

[*] *Information Technology and Computing*, Arab Open University Egypt-Branch
Email: Abdelrahmanhamdy33@gmail.com

[1] Electronics Research Institute, *Department of Computer and Systems*
Bio-Computing and Developmental Systems (BDS)
Email: aiman.mahgoub@ul.ie

[2] *Department of Compute Science and Information Systems*, University of Limerick,
Limerick, Ireland
Email: conor.ryan@ul.ie

***Abstract* - The novel coronavirus has had a huge impact on the world, not only for those infected, but for the population at large in the context of the work, money incoming, economy. In this paper, our objective is to study different types of tweets collected from Twitter from different perspectives of analysis, and machine learning classification. Previous work extracted a large corpus of tweets based on keywords such as [كورونا ، السفر ، تفشي ، الصين], which translate to China, Corona, outbreak, travel; however, we hypothesis that not all these are genuinely relevant to Covid-19. In this work, we combine different machine learning models to classify tweets into those that do discuss the Coronavirus and those that do not. In our result, based on the different analysis and models, we have seen that more than 55% of these tweets were talking about topics other than the Coronavirus. These indicates the care that must be taken when extracting tweets, particularly in a language like Arabic, which has many nuances that make simple keyword approaches prone to error.**

Keywords - Arabic Covid-19 data set, analysis, machine learning classification, & clustering.

I. INTRODUCTION

Covid-19 is a worldwide pandemic that hit the entire globe unexpectedly. This pandemic has many effects on society and economy. The worldwide pandemic has caused massive loss lives and a huge number of still-increasing infections. The pandemic has had a high impact on societies and economies all over the world. The spread of pandemic, which now in these days have its Second wave with 76,250,431 confirmed cases around the world and 1,699,230 deaths on December 23, 2020, as reported by World Health Organization (WHO). Social media has been a huge source of information and careful analysis can yield insights into events, ideally helping improve our response to similar events in the future. This analysis must be done automatically using natural language processing (NLP) techniques due to the massive data available on social media platforms. There has been a lot of work on Covid-19 social media data but little done on Arabic data sets to extract information, analyze and classify these data sets. In this current work we choose to work with Arabic data which contain many features which make it particularly challenging to work with. Arabic is the fifth most spoken language in the entire world and the fourth most used language on internet. It is challenging language to work with because it contains diacritics that can change the meaning of a sentence, and has a difficult grammar structure compare to non-Arabic languages. There are other challenges when dealing with Arabic language for example sentences written different grammatically has different meaning despite using similar words.

The rest of the paper is organized as follows. Section II has a short literature survey on work related to the paper. Section III discusses the data set we used in details, section IV describes the data processing we did to the data set and the steps we followed to make the tweets ready for the classification problem. Section VI discusses the classification we did to the data set and results in details and section VI gives the conclusion of the paper. The implementation is accessible on GitHub at this address:

https://github.com/Abdelrahmanrezk/Arabic-Hands-on-Analysis-Clustering-and-Classification-of-Large-Arabic-Twitter-Data-set-on-COVID19

II. LITERATURE SURVEY

There is much work done on analysis of Covid-19 tweets in English and other languages. In [1] a content analysis of Persian/Farsi Tweets during pandemic in Iran was presented. The authors analyze the content of 530,000 original tweets

after doing some prepossessing steps. They annotated random tweets to classify the data set and get more analytical information from it. In [2] an analysis for tweets in North America was done to a tweet data set. That analysis was done using human in the loop experts. In [3] machine learning models were used to perform to sentiment analysis for fear in tweets it using naive Bayes and achieving accuracy of 91% for sentiment analysis. In [4] the authors perform sentiment analysis on the data set to answer the question if people are depressed during the pandemic. They use NLP (natural language processing) to study the effect of the pandemic on our mental health. However, there has been some work done to analyze Arabic tweets during the pandemic, for example, in [5] an NLP analysis for Arabic tweets in the early days of Covid-19 was done. The authors divide the tweets according to the subject they tweet about. They used machine learning models to classify and analysis the data set. In [6] the authors used Arabic tweets data set for ranking common Covid-19 symptoms. In [7] used Arabic data set to classify the tweet as either rumor or non-rumor. The author used three machine learning models to classify the tweets and was able to reach 84% accuracy using feature extraction such as Word2Vec.

Data Set Description

The large Arabic data set is Arabic tweets collected from the Twitter streaming API by [8], the referenced paper has their own way of collecting the tweets and they have the full tweet object, including ID, username, hashtag, and geolocation of each tweet, but, using the IDs of the tweets we created our own pipeline for collecting these tweets, as our problem just depends on the tweets Arabic text, so we retrieve only the text of the tweets based on the IDs of the tweets. The collected tweets are from January 1, 2020 to April 15, 2020 and the overall number of tweets is more than 3,934,610 tweets. Our concern with the data set was on the choice of keywords, which we felt could generate a data set too general to base many conclusions on. We will prove by machine learning algorithms that using only these keywords in gathering the data set resulted in large portion of the data set are tweets that are not related to Covid-19.

III. DATA PROCESSING

We manually labeled more than 15,000 tweets to class 0, for general tweets, and 1 for Coronavirus tweets. This data was used to train used machine learning models which were subsequently used to classify the entire data set. The data comes in different forms, not all of which are useful to us. For example, URLs are generally not helpful in this case, and diacritics, noted below. While many NLP processes remove simple stop-words such as “not”, this is not possible in such as Sentiment Analysis as these words can change the meaning of the tweet.

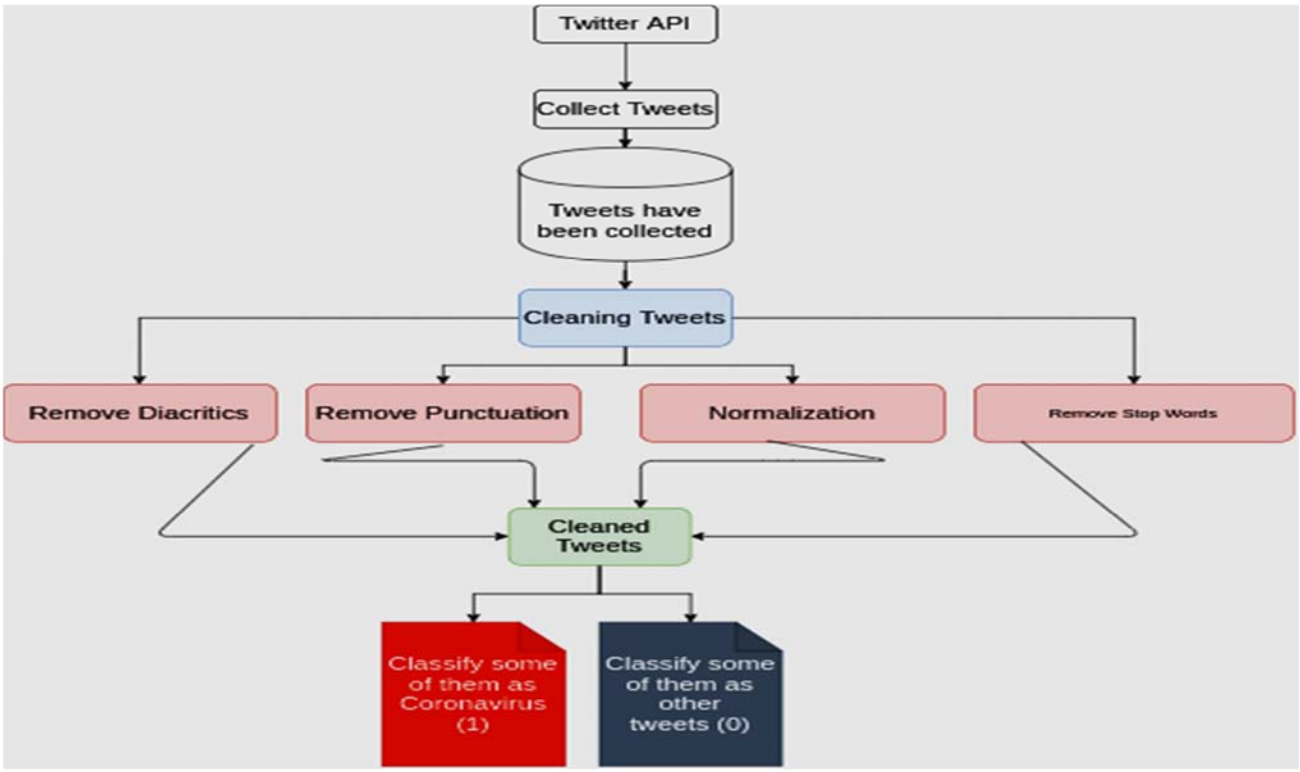

Figure 1. Pipeline of collecting, preparing, and labeling Tweets.

A. Remove Diacritics

In this pre-processing step we remove all Arabic diacritics[9] from the words Since various Arabic users of twitter write tweets in different ways, for example, some use diacritics, but most do not. Diacritics have can insignificant impact on the problem of classification since most people write tweets without them and since it will not affect our classification of tweets related to Covid or not, as they change the meaning slightly. We removed diacritics so the machine learning model doesn't get confused between tweets with diacritics and tweets without diacritics. Diacritics may confuse machine learning models, because they represent some of pattern and will take a place in our memory in the process of features engineering [10].

B. Remove Punctuation and URLs

One of the most important step in our cleaning is to remove punctuation and URLs because most of the tweets have a lot of other characters that have no meaning, such as [-,$, #] and other characters. Similarly, tweets have URLs [11] of some reference we have cleaned these tweets by removing these URLs and punctuation marks.

C. Normalize the Text

The form of the word can be written in different ways in languages like Arabic but each form can still have the same meaning, such as "العربية" , which some users write as "العربيه", [12], but actually have the same meaning, so like these words are normalized to a single form.

D. Remove Stop Words

We have designed our own stop words instead of using some of the provided lists from libraries such NLTK, because the sensitive problem we dealing with here is Classification, and like these libraries and other libraries remove Arabic words like [ليس ، لا], and other words, and like these words to be removed will change the whole contest of the tweet like from negative to positive meaning, so we have chosen our stop words [13] carefully based on our problem.

The 50 Most frequently used words

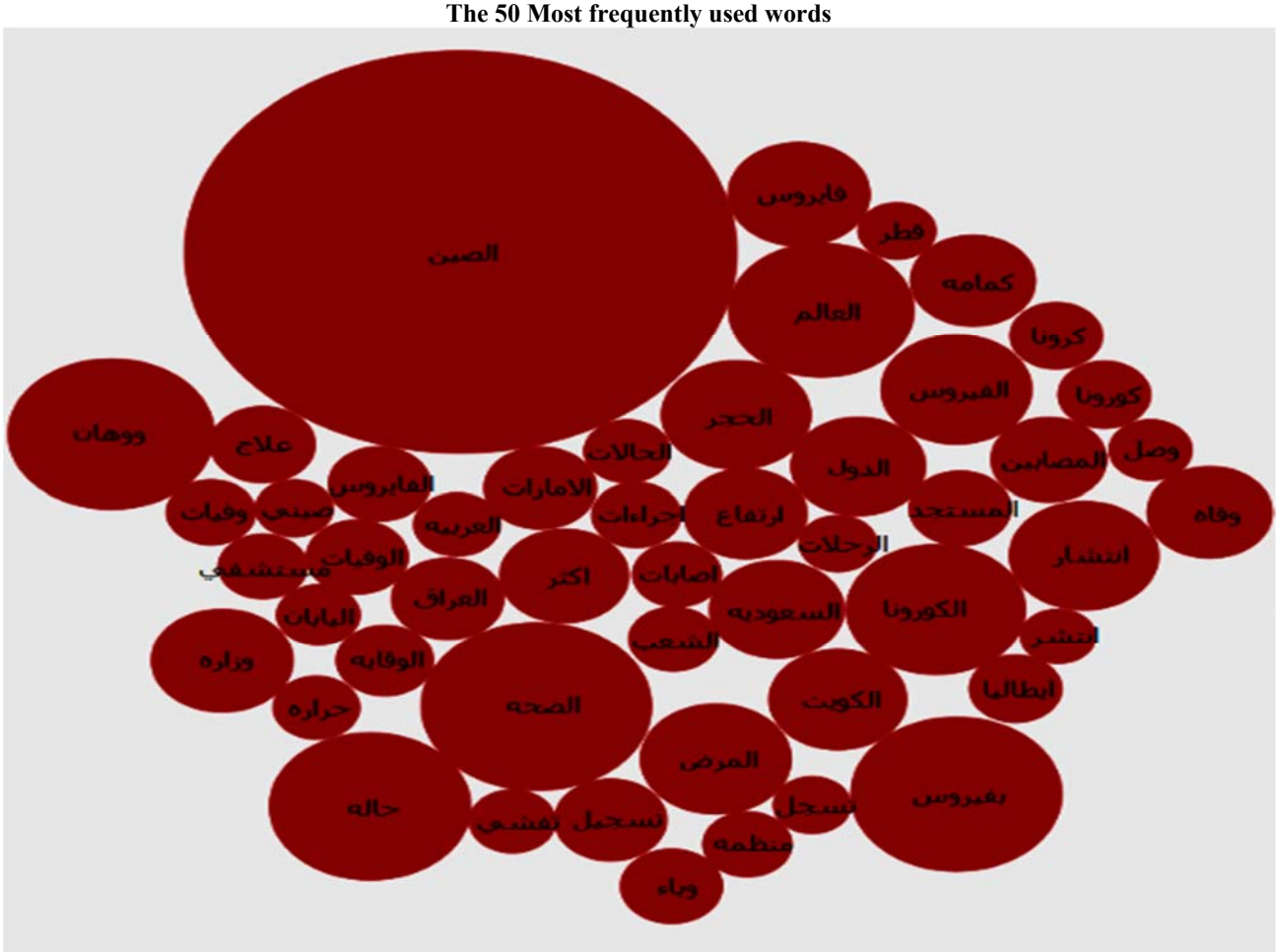

Figure 2. Frequency count of most words in the tweets.

IV. DATA ANALYSIS

In this section we perform some analysis on the data to extract some information about the Covid-19 situation in its early stages in the Arabic world from social media. Spending some time to analyze the tweets' content, such as the most repeated words and the similar words, helps us in the pre-processing phase and can help us to choose the models that can fit within the training, and get a good result in the testing phase using the weights of the models we trained [14].

A. Frequency Count

We implemented a function that counts the number of words in each tweet. It simply counts the words that are repeated frequently in the tweets; not surprisingly, we found that the most repeated world in the data set is China.

B. Tweets per day

We also draw the tweets sent per day to get insights from the Arabic data set. We found almost 100,000 tweets the day after Covid-19 was first discovered in the Middle East, showing how highly people interacted with social media in the immediate aftermath of such a huge event. The figure shows how much attraction gained the social media for tweets related to Covid19 and how much these tweets started to decrease until reach a certain level. Towards the end of the month the tweets started rising again as the society interest in the event started to gain.

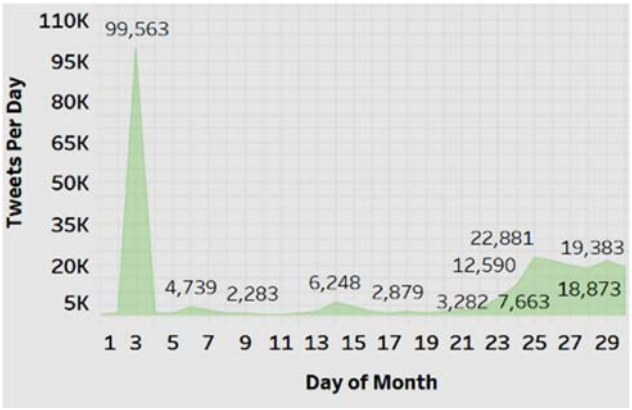

Figure 3. The number of tweets each day of January month.

C. Word2Vec analysis

The most important step in the analysis used Word2Vec to create feature vectors for the tweets. Word2vec uses a Neural Network to learn features and, unlike methods such as TFIDF xyz needs a cite methods create features simply based on count words. It builds an understanding of the words and can identify words that are related, for example, [Orange and Apple] are related to each other and can come in similar context of text [15]. Word2Vec is a two-layer neural net that processes text by "Vectorizing" words. Its input is a text corpus and its output is a set of feature vectors that represent words in that corpus.

V. DATA SET CLASSIFICATION

We use three classification techniques, namely, Random Forest classifier, SVC (support vector classification) and logistic regression to classify the data into either Class 1, in which the tweets are highly related to the Covid-19, or Class 2, in which the tweets are not related to Covid-19. We compare our results between the three models after fine-tuning the parameters in different our results show that the classification AI models prove our idea that there is more than 55% of the tweets are not related to Covid-19. The 15K tweets where divided by 25 percent training and 75 percent testing. Here we show tables for the results of the percentage between the training data set and testing data set for the three models.

A. Random Forest Classifier

A Random Forest classifier is an ensemble of a large number of decision trees that work together. We use Random Forest Classifier with Tf-IdF and word2vec feature extraction on our 15000 annotated tweets. We obtained 99.7% accuracy with our training data set and 92.7% with the testing data set. The full results for Random Forest Classifier are shown in Table I.

TABLE I. RANDOM FOREST CLASSIFIER ACCURACY

Features Engineer	Training	Testing
word2vec	99.9	85.2
TF-Idf	99.7	92.7

B. SVC

Support vector classifiers are another type of supervised machine learning classifiers that are widely used in classifying text such as tweets. It usually used to classify between two classes which is our case where we are trying to classify between Covid-19 related tweets and non-Covid-19 related tweets. Support vector machine classifier are fast and can work efficiently with limited data to analyze. The idea of support vector classifier is simple. It is finding the hyper plane (decision boundary) between two classes in the features plan. We got the results in Table II from the support vector classifier. Again in this case, using Tf-IDF gave the best test performance, also scoring 92.7%.

TABLE II. SUPPORT VECTOR CLASSIFIER ACCURACY

Features Engineer	Training	Testing
word2vec	89	84.9
TF-Idf	98.6	92.7

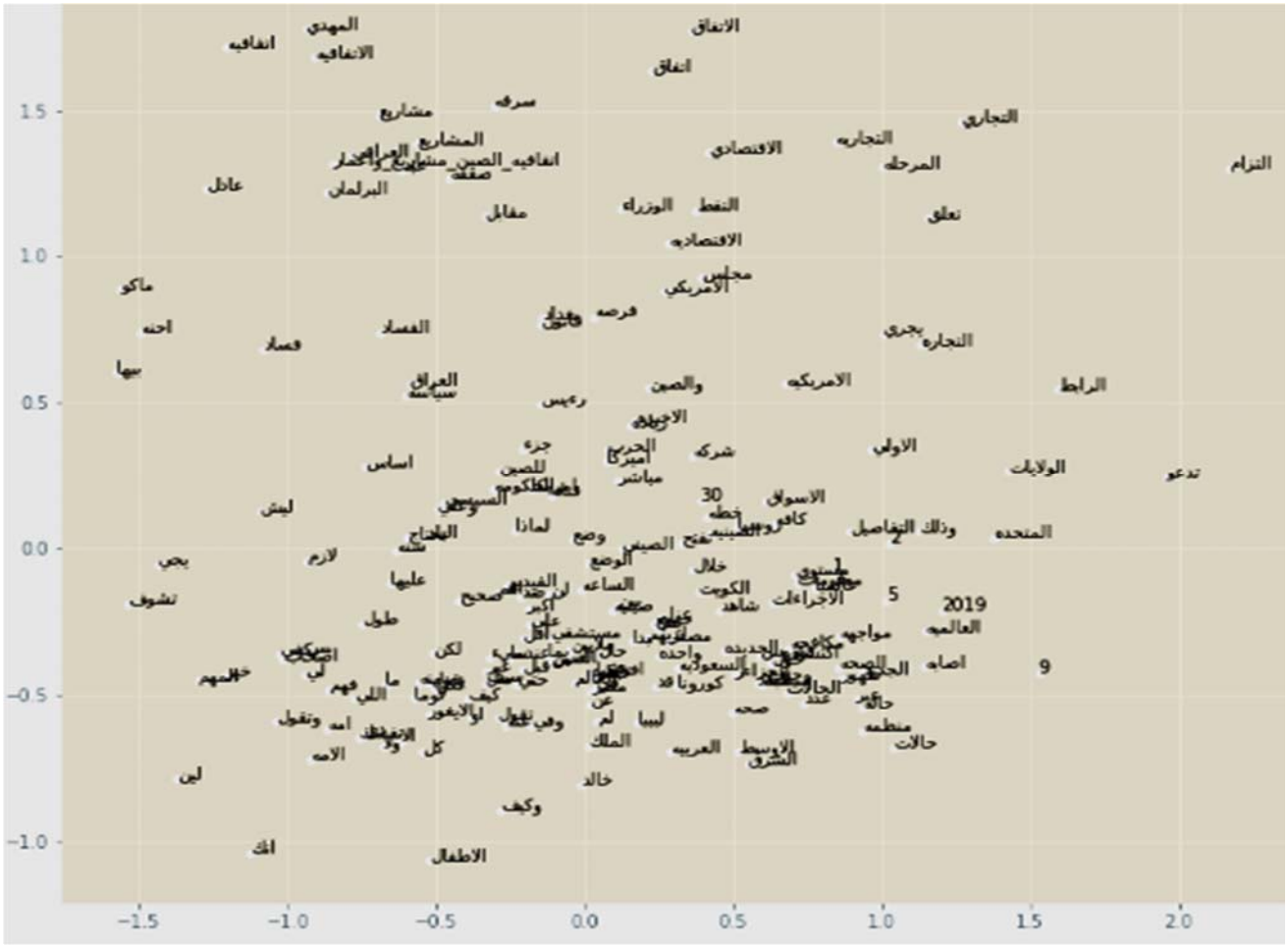

Figure 4. Word2Vec Words.

C. Logistic Regression Classifier

Logistic regression Classifier is a statistical learning technique mainly used in classification tasks. Logistic regression is mainly used in binary classification problems such as our classification task. Our results shown in Table III.

TABLE III LOGISTIC REGRESSION CLASSIFIER ACCURACY

Features Engineer	Training	Testing
word2vec	90.1	86.2
TF-Idf	93.6	91.8

We applied the Logistic Regression classifier on the entire data set and the results as follows the number of coronavirus tweets are 128,971 and the numbers of non-coronavirus tweets are 175,447. Figure 5 shows the results of classification of the entire data set.

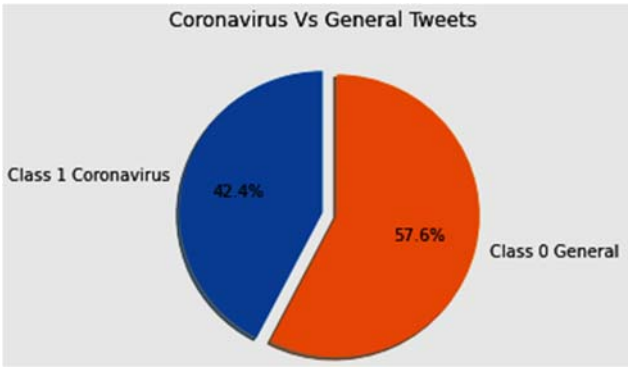

Figure 5. Classification of the entire data set

VI. CLUSTERING

We performed a clustering exercise using K-means clustering on the data to determine the quality of our predictions.

This is a second mechanism to prove our results we are getting from the machine learning models are valid we used k-means to cluster the data set we will able to reach 70% accuracy which prove that the data set is divided to two parts tweets related to Covid-19 and non-related Covid-19 the low percentage however comes from that the distinction is not apparent to the clustering model. Table IV shows our results on clustering the data set.

TABLE IV. CLUSTERING ACCURACY

Features Engineer	Training	Testing
Tf-idf	68.8	67.3

VII. CONCLUSION AND FUTURE WORK

In this work we introduce our cleaning, analysis, Classification and clustering on a large Arabic data set. We noticed that large percentage of the data set is not related to Covid-19. To prove our conclusion using machine learning models we made pre-processing to the data set. We did some analysis to the tweets like frequency count where we count the frequency of repetition of each word in the tweets. Using this analysis we were able to get some insights from the data set. We have manually annotated 15000 tweets randomly from the data set and annotated them highly related to Covid-19 tweet and non-related to Covid-19 tweets to test our hypo-sis that many tweets are not related to Covid-19 we used three machine learning models random forest, SVC and logistic regression. We got working models with good results this shows more than 55% of the data set is not related to Covid19 but only uses similar worlds that is related to Covid-19. We also applied a clustering technique to verify our hypo-sis using k-means classification we reached 70% accuracy which is affected by similar words but prove our hypo-sis that their is a large portion of the data set is not highly related to Covid19 and additional pre-processing step is required to work with such data sets that relay on related words to collect data set. Future work will include sentiment analysis for the highly related tweets.

ACKNOWLEDGMENT

The authors want to thank the reviewers for their time and helpful suggestions. This work was supported with the financial support of the Science Foundation Ireland grants16/IA/4605 and 13/RC/2094.

REFERENCES

[1] A Alajmi, E. S. a. R. D., 2012. Toward an ARABIC Stop-Words List Generation.
[2] Ademola, A. O. a. E., 2014. A Review of Big Data Management, Benefits and Challenges.
[3] Almajed., A. M. A. a. R. S., 2013. A survey of automatic Arabic diacritization techniques.
[4] Eisa Alanazi, A. A. S. A. a. A. A., 2020. Identifying and Ranking Common COVID-19 Symptoms From Tweets in Arabic: Content Analysis.
[5] Hamdy Mubarak, S. H., 2021. ArCorona: Analyzing Arabic Tweets in the Early Days of Coronavirus (COVID-19) Pandemic.
[6] Hyeju Jang, E. R. G. C. N. J., 2020. Exploratory Analysis of COVID-19 Related Tweets in North America to Inform Public Health Institutes.
[7] Irene Li, Y. L. T. L. S. A.-N. D. G.-G. T. S., 2020. What are We Depressed about When We Talk about COVID19: Mental Health Analysis on Tweets Using Natural Language Processing.
[8] Jim Samuel, G. A. M. R. E. E. Y., 2020. COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification.
[9] Khemchandani., T. R. a. V., 2019. Feature Engineering (FE) Tools and Techniques for Better Classification Performance.
[10] Lama Alsudias, P. R., 2020. COVID-19 and Arabic Twitter: How can Arab World Governments and Public Health Organizations Learn from Social Media?.
[11] Nora Al-Twairesh, H. A.-K. A. A. a. Y. A.-O., 2017. AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets.
[12] Pedram Hosseini, P. H. a. D. A., 2020. Content analysis of Persian/Farsi Tweets during COVID-19 pandemic in Iran using NLP.
[13] Rouhia M Sallam, H. M. M. a. M. H., 2016. Improving Arabic Text Categorization using Normalization and Stemming Techniques.
[14] Sarah Alqurashi, A. A. E. A., 2020. Large Arabic Twitter Dataset on COVID-19.
[15] Tomas Mikolov, K. C. G. C. a. J. D., 2013. Efficient Estimation of Word Representations in Vector Space.

BIOGRAPHY

Prof. Conor Ryan is a Professor of Machine Learning in the Computer Science and Information Systems (CSIS) department, a Funded Investigator within Lero (the Irish Software Research Centre) and a Science Foundation of Ireland Principal Investigator. He was a Fulbright Scholar at the Massachusetts Institute of Technology in 2013/14 and was CTO of NVMdurance, a company that optimized Flash Memory products, until it had a successful exit in early 2018. He is the inventor of Grammatical Evolution, one of the most widely used Automatic Programming systems.

Ayman is currently a postdoctoral researcher in limerick university BDS group. Ayman also a researcher/Assistant professor in electronics research institute. he got his Bsc form electronics and communication department Cairo university. He got his master degree from electronics and engineering department. He got his PHD in electronics and communication from department AIN shams university. His research interest includes ANN, fuzzy logic, FPGA design, and photovoltaic systems.

Abdelrahman is a Teaching Assistant at Arab Open University Egypt-Branch and research in NLP, Machine & Deep Learning. Also, he was a Software Developer for two years besides of solved more than 500 problems on different online judges, I got two internships, and one of these interns was with an international research center in Qatar Computing Research Institute (QCRI). Also, he got Second place in Africa with the Hand Speak team in BeChangeMaker program Sponsered by HP foundation, wordlskillsafrica, and Africa union, with EUR cash reward, paid mentorship, HP laptop & BCM trophy.