Sentiment Analysis on Arabic Public Opinions toward COVID-19 Vaccines Using Twitter Data.

Sara Abduljaleel[1], Ghidaa Alnefaiy[1], Najwa Alshahrani[1], Tahani Alqurashi[1], Zahyah H. Alharbi[2].

1.Department of Computer Science, Umm Al-Qura University, Makkah, Saudi Arabia

2.Management Information Systems Department, King Saud University, Riyadh, Saud Arabia

(s44286098, s44287423, s44287422, tmqurashi)@st.uqu.edu.sa, zalharbi@ksu.edu.sa

Abstract

Social media has emerged as a critical communication tool in contemporary society, bringing the world closer by making news and opinions more accessible through virtual platforms. Among these, Twitter stands out with over 350 million users worldwide. However, the highly unstructured nature of Twitter data poses significant challenges for analysis. Recently, the rich content of social media networks has spurred extensive research, particularly in light of the COVID-19 pandemic that has impacted global populations. Concurrent with the search for a vaccine, numerous studies have focused on analyzing public sentiment during the crisis, using data from various social media platforms. This study concentrates on people's opinions regarding COVID-19 vaccines, employing several machine learning models to analyze data from Twitter. It was discovered that the most accurate algorithm was the SGD Classifier, which utilized all three n-gram ranges, achieving an accuracy of 0.75.

Keywords:

Arabic Tweets, COVID-19 Vaccines, Sentiment Analysis, Social Media and Twitter Data.

1. Introduction

Social media encompasses various applications designed for the web and mobile devices, facilitating communication and content sharing. As the use of social media rapidly grows, becoming a prominent aspect of social information systems, companies and researchers are developing new methods and tools to analyze the burgeoning data [1]. A notable advancement in this field is sentiment analysis, also known as opinion mining. This subfield of Natural Language Processing (NLP) employs Artificial Intelligence (AI) techniques to identify attitudes towards specific targets or topics of interest [2-4].

Since the end of 2019 to mid-2020, COVID-19 has been identified a pandemic due to the rapid spread of the virus around the globe [5]. Its impact has been profound on all aspects of society, including the mental and physical well-being of its members. The rapid spread of COVID-19 has caused a global public health crisis. Countries and large pharmaceutical companies rushed to start the quest for a vaccine.

Consequently, there have been several vaccines developed, including Johnson & Johnson's Janssen, Astra-Zeneca, Novavax, Moderna, and Pfizer-BioNTech. The COVID-19 vaccines were granted temporary authorization for emergency use due to their complex development and long authorization process. Although scientists and authority figures emphasized the importance of vaccines in stopping the outbreak of the virus, the public still called the safety of the COVID-19 vaccines into question. These concerns and opinions were vividly ex-pressed on social media platforms, especially given the social restrictions and lockdowns imposed at that time by many governments.

According to [6], the number of Arabic-speaking Twitter users has surged. However, the application of text mining and natural language processing to Arabic text faces challenges due to the incompatibility of existing algorithms with the diverse Arabic dialects on social media. This difficulty stems from the intricate structure and grammar of the Arabic language and the linguistic diversity across the Arabic-speaking world. Consequently, most sentiment analysis models, primarily developed for languages like English, fall short in adequately addressing Arabic text[7].

This study addresses two main research questions. First: What are people's opinions regarding COVID-19 vaccines in the Arab region that have been expressed on Twitter? Second: What are the common side effects that users have stated in their tweets after receiving the vaccine?

The remaining of the paper is structured as follows: Section 2 discusses related work in the field of Arabic Sentiment Analysis. Section 3 states the implemented sentiment analysis including data collection. Section 4 contains the experiment and the evaluation. Section 5 reports the results and discussion. Finally, future work and conclusions are presented in section 6.

2. Related Work

The most common method to collect Twitter data is via Twitter API. It provides developers with programmatic access to Twitter data through a special application programming interface (API) and supports data filtering tools [8].

In [9], researchers first collected two datasets, then they extracted the third one from the two datasets depending on the geolocation of the data (i.e. tweets located in Saudi Arabia). They initially collected around 6 million tweets and reached a net of 2 million remaining tweets after applying filtering techniques on the datasets. Also, in [10] Twitter API was used to collect data all related to COVID-19. In addition to the previous papers, researchers in [11] a Twitter scraper called ‘Twint’ was used to collect data from Twitter and build the dataset. Arabic was chosen from the language filtering options to retrieve tweets in Arabic only. In order to retrieve relevant tweets during the pandemic, a list of hashtags was prepared to be used as search keywords. This approach focused on capturing all tweets mentioning COVID-19 in Arab countries. Besides these hashtags, multiple queries were conducted to build join terms related to COVID-19 with the names of all Arab countries, such as #Corona_Saudia.

Furthermore, [12] presented 72 of the most famous Twitter news accounts including official news agencies, newspapers, and human and civil organizations. The choice of these accounts is very important in this context because the tweets are written in Modern Standard Arabic (MSA). The noise level in these accounts is much lower than in public accounts where the diversity of Arabic dialects has to be accounted for. The data collection resulted in 12 million tweets, and after filtering the data using Twitter API based on country names, 2.3 million modern Standard Arabic (MSA) tweets were considered. In [13], they designed models using the Twitter API to collect tweets on a particular topic (e.g., using a hashtag or search term) or from specific accounts.

In [15], the Twitter API was utilized to collect tweets from Syria. The Word2Vec tool, developed by Google, was employed to streamline the computation of vector representations of words in a lower-dimensional space, facilitating the association of common semantics among words. The dataset was then divided into training and test data using k-fold stratification [15]. In paper [16], two datasets were collected. The first, the Twitter Arabic Dialect dataset (TAD), comprises approximately 8 million tweets in the Arabic language and identifies four dialects: Gulf, Levantine, North Africa, and Egyptian. The second dataset, Twitter Arabic Dialect Emoji (TADE), was extracted from TAD and consists of tweets containing emojis. Furthermore, for feature selection, two methods were compared: the 'bag-of-words' method utilizing TF-IDF, and the more contemporary method of word embeddings [16]

Preprocessing is the crucial phase that aims to reduce the inconsistencies, nuisance, normalize and transform the data into a coherent format. Tweets scraped from Twitter often contain noise and must be cleaned before performing any of the NLP tasks.

Before performing data analysis, several steps are taken to prepare the data and remove any noise. First, researchers remove the mention symbol (@), URLs, and diacritical marks. Normalization is then applied to the data. This step involves replacing letters that have multiple forms into a unified form. For example, the Alef Letter (آ، إ، أ) is replaced with (ا); the Ya'a Letter (ي) with (ى); the Ta'a-Marbota (ة) with(ه); and the Waw-Hamzah (ؤ) with (و); to unify the shape of letters [6, 7, 8, 14]. Non-Arabic letters are also removed [7, 14]. In addition, in [16], if religious references were found, they were removed. Also, tweets containing Arabic letters but are not written in the Arabic language were removed. In [8, 13], they solved the problem of repeated letters by removing the extra letters restoring the word to its root. They also excluded duplicate Tweets during the crawling process by searching for the Tweet ID. Removing punctuation is also a common step in data preprocessing [9]. A tokenization technique was applied in [6,7]. In [13], the text of the tweets was pre-processed by removing non-word symbols like mentions, emojis, and images.

Researchers often divide the main dataset into two datasets: one for training the system and the other for testing the system (to ensure there is no overfitting). In [9], they built a corpus for Arabic Sentiment Analysis of Saudi Tweets entitled 'AraSenTi-Tweet'. Four-way sentiment classification was then applied to annotate the tweets: positive, negative, neutral, or mixed. The tweets were manually labelled. Then, the experiment was conducted on the resulting corpus by utilizing different features. Many classifications were applied on these sets. By using only positive and negative tweets, they got a two-way classification. In the three-way classification, positive, negative, and neutral classes were used. Lastly, for the four-way classification, all classes were used. F1-score was reported for evaluation. A Support Vector Machine (SVM) with linear kernel was utilized for classification. Term-presence, term-frequencies, and term frequency–inverse document frequency (TF-IDF) features were tested. The TD-IDF is an approach to convert a text document into a vector. TF-IDF proved that it was the best performance in the four-way classifications.

In addition to that, in [10], a Non-Negative Matrix Factorization (NMF) text modelling approach was used to identify topics in documents. Before the NMF modelling technique was conducted, clean data was first transformed into TF-IDF. Then, they used the Topic Coherence–Word2Vec (TC-W2V) to catch the meaning of the word and measure the coherence between words assigned to a topic. The model was trained by utilizing the Skipgram algorithm. Then, NMF was trained again for different values of K, the number of topics. While in [14], multiple models were used. The techniques included a Convolutional Neural Network (CNN), a Long Short-Term Memory (LSTM), and an ensemble model that combines between them were used. The Arabic Sentiment Tweets Dataset (ASTD) was used and classified into four classes using the aforementioned techniques (positive, negative, neutral, and objective).

Other Machine Learning algorithms were also used in [14]: Nearest Centroid, Decision Tree, XGB, AdaBoost, Logistic Regression, SVM, Gradient Boosting, Random Forest, MultinomialNB and Convolutional Neural Network (CNN).

In [14], tweets were classified into suspicious or not suspicious using six machine learning algorithms to test the system. The algorithms used were the K-Nearest Neighbors (KNN), SVM, LSTM networks, decision trees, Artificial Neural Networks (ANN), and linear discriminant algorithm.

Researchers in [12] train a Neural Network (NN) model using a manually annotated dataset of 6000 tweets (i.e., the training dataset). The model is tested using another 1200 tweets (also manually annotated) and the Kohen's Kappa metric is calculated to measure the inter-annotator agreement (IAA).

Data were collected from Twitter API in [11]. Then Arabic emotion lexicons were used to annotate the tweets. Deep learning classifiers were utilized in the analysis to ensure the quality of the output. In order to classify symptoms, a list of COVID-19 symptom keywords has been extracted. The final step was to store the annotated dataset in a database for monitoring purposes.

Additionally, Alhumoud et al. [17] explore Arabic Sentiment Analysis for Vaccine-Related COVID-19 Tweets (ASAVACT), marking the first and largest human-annotated Arabic dataset in this domain. They analyze sentiment polarity in 32,476 annotated tweets using advanced deep learning models, such as the Stacked Gated Recurrent Unit (SGRU), the Stacked Bidirectional Gated Recurrent Unit (SBi-GRU), and an ensemble of SGRU, SBi-GRU, and AraBERT. Notably, the ensemble model has outperformed individual models, achieving at least a 7% enhancement in accuracy. This research underscores the effectiveness of ensemble deep learning approaches in Arabic sentiment analysis, particularly in the context of socially relevant topics like COVID-19 vaccination

3. Methodology

In this paper, the main interest is to identify different attitudes towards COVID-19 vaccines in the Arab re-gion. To achieve that, the methodology displayed in Fig. 1. was pursued. It consists of four steps: 1) data collec-tion, 2) data preprocessing, 3) sentiment analysis 4) result evaluation.

First, Arabic tweets that contain individual opinions were collected. Then, data preprocessing was performed to ensure the suitability of the collected data for the sentiment analysis step. In the sentiment analysis step, sev-eral classification algorithms were applied to classify the data into classes and then evaluated to determine the best classifier.

The following subsections explain the methodology in more details:

3.1 Data Collection

To collect data, Twitter's streaming API was used. Twitter API allows for the scraping of a limited number of tweets within a 30-day period only. To cover the time interval of interest, from September 2020 to April 2021 - the time period that involved the release of vaccines and their distribution and administration around the world - the collected dataset was merged with another dataset that also covered the same time interval. Therefore, after merging the two datasets, the final dataset consisted of a total number of 34,953 Arabic tweets.

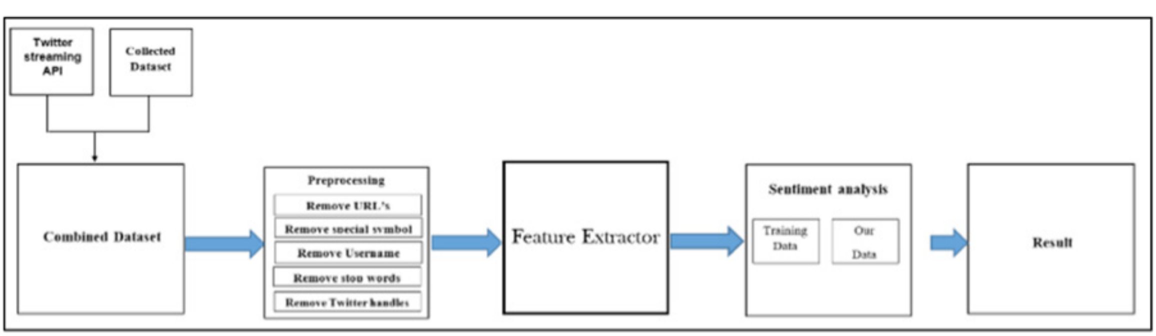

Fig. 1. Research Methodology

3.2 Data Preprocessing

Pre-processing consists of reducing trivial noise and cleaning the data by applying several steps to the dataset.

The following text pre-processing methods were utilized: First, the cleaning step was applied to remove noisy data. This included mentions, URLs, emojis, and punctuation. Hashtags were also removed maintaining the content of the tweets by using multiple Python's libraries. Regular expression (re) Python module was used to delete the @ and username string (i.e., @user). The NLTK library, which contains Arabic stop words, was also used to perform the following pre-processing tasks:

- Removal of URLs. Twitter data include various types of information. If a user posted a link that is of no interest to the sentiment analysis, it was removed from the tweet.
- Removal of special symbols. There are different types of symbols utilized in tweets such as punctuation marks (e.g. exclamation marks, periods, etc.….) which do not entail any sentiment. These special symbols were removed from the tweet.
- Converting emoticons and emojis. Table 1. shows a sample of various emoticons that were contained in a collected tweets. Nowadays, users use emoticons to express their feelings, emotions, and reactions. Sentiment analysis is heavily influenced by emotions. Therefore, it is important to include emoticons in the sentiment analysis by converting the emoticon into a word corresponding to the emotion it entails.
- Removal of Arabic punctuation, numbers, and special characters.
- Removal of diacritics.
- Removal of Arabic stop words.
- Removal of Twitter handles and #hashtags.
- Removal of duplicate tweets.

Table 1 shows an example of tweets before and after the preprocessing step.

After cleaning the data, the normalization step ensued:

- Letters such as (أ،إ،آ) were converted into (ا), (ؤ) into (و) , and (ي) into (ى).
- Arabic ligatures were removed and words were returned to its root. For example: (اللقـاح) was returned to (لقـاح) and (الخميـس) to (خميـس).
- Word connectors such as 'و' were also removed.
- After the conclusion of the data preprocessing process, sentiment analysis can be performed on the clean, processed data.

Table 1. An example of Tweets before and after the Data Pre-processing step.

	Tweets
Before	اخذت اللقاح يوم الخميس وتوني امس اصحصح 😌 الحمدلله باقي الجرعه الثانيه ،💞 https://t.co/EsmsIxLTVC
Translation	I took the vaccine on Thursday, and just recovered yesterday. Thanks God only one dose left.
After	اخذت لقاح خميس وتوني اصحصح حمدللهه باقي جرعه ثانيه

	Tweets
Before	لحماية نفسي ووطني أخذت لقاح كورونا وسأبقى ملتزمًا بالإجراءات الاحترازية. أدعوك للتسجيل عبر تطبيق صحتي لنحمي مجتمعنا، #خذ_الخطوة خذ اللقاح
Translation	To save my country and I, I got the vaccine and I will be following the precautions. I encourage you to register in Sehhaty Application to save our community.
After	لحماية نفسي وطني اخذت لقاح كورونا وسأبقى ملتزما اجراءات احترازية. ادعوك تسجيل تطبيق صحتي لنحمي مجتمعنا، خذ خطوة لقاح

3.3 Feature Extraction

In this step, the text data must be converted into numerical data that can be easily handle by classification algorithms. Thus, the most popular approach in extracting feature from text was used, which is the Term Frequency-inverse Document Frequency (TF-IDF).

In fact, TF-IDF combines two scores, the term frequency (TF) and the inverse document frequency (IDF). TF calculates the frequency of word in each tweet, while calculating he IDF increase the weights of words that are recurring very rarely and reduce the weights of words that are recurring frequently. The. TF-IDF is defined as shown in Eq. 1.

$$TF\text{-}IDF(w, t, D) = TF(w, t) * IDF(w, D) \qquad (1)$$

Where, w represents a word in a tweet t. IDF defined as shown in Eq. 2, where D is the total number of tweets in the dataset and the df(w) is the number of tweets in which word w appears in D.

$$IDF(w, D) = log(\frac{|D|}{\mathrm{df(t)}}) \qquad (2)$$

3.4 Sentiment Analysis

The aim of the analysis was to classify each tweet into one of two classes, either positive or negative based on the sentiment it entails. As the data is big and unlabeled, the transfer learning technique was used. Transfer learning is a machine learning approach that aims to transfer knowledge learned from a previous problem to solve another similar problem [12].

Therefore, a previously labeled Arabic corpus was downloaded and utilized for this study. Various classifiers were then trained to categorize the sentiment of each tweet as either positive or negative. Subsequently, these trained models underwent evaluation against a test dataset to assess their performance. The results of this evaluation are detailed in the following section.

4. Results and Discussion

Results of the trained models are reported in Table 2. For each algorithm, three different N-grams were modeled and tested (i.e., unigrams (one word), bigrams (two adjacent words), trigrams (three adjacent words)). Results were evaluated using accuracy, precision, and recall measures.

The three most accurate algorithms were the SGDClassifier, BernoulliNB, and MultinomialNB as shown in Table 2. The best was the SGDClassifier in all grams with a score of 0.75 in terms of accuracy. After applying the SGDClassifier to the test dataset, frequencies of positive and negative tweets were tabulated as reported in Fig. 2. The frequency of positive tweets is shown to be 11620, and the frequency of negative tweets is shown to be: 23337.

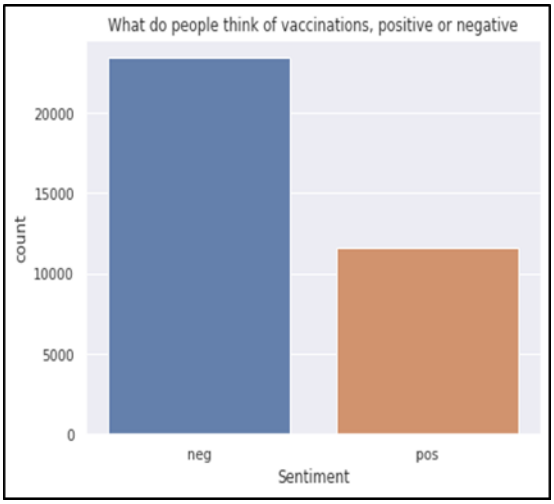

Fig. 2. Sentiment analysis on people's opinion on Covid-19 vaccination

To be more specific, the research focused on the opinions on two specific vaccines: Pfizer and AstraZeneca. Most people were found to feel positively towards both vaccines, as presented in Figure 3, although the majority of individuals feel negatively towards vaccines in general.

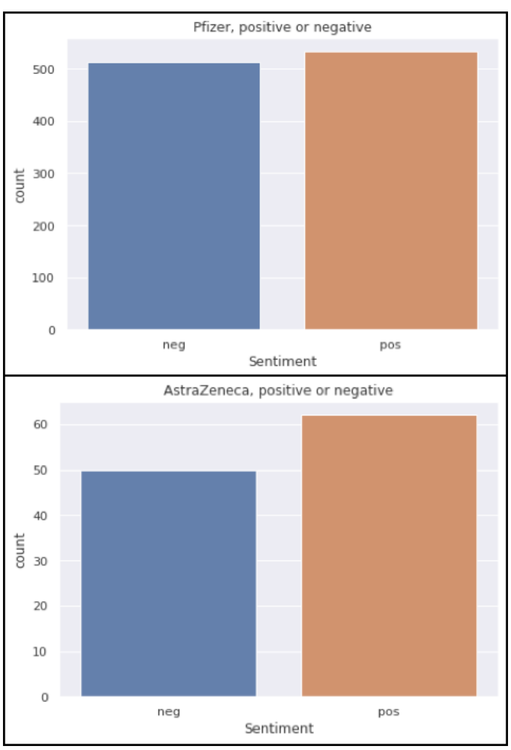

Fig. 3. Sentiment analysis on people's opinion on Pfizer and AstraZeneca.

Regarding the side effects that people reported, after receiving the vaccine, in their tweets: (حرارة) and (صداع) were the most frequent. Figure. 4 illustrates the side effects and their corresponding frequencies.

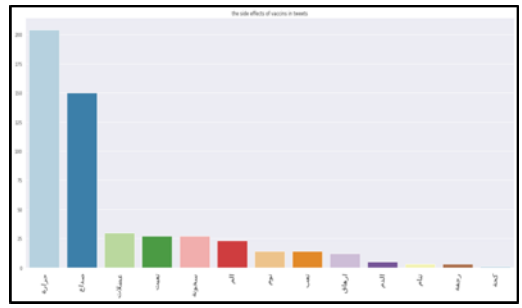

Fig. 4. The frequency of the symptoms that people mentioned in their tweets.

Table 2. The results of tested classifiers algorithms.

Algorithm	Ngram	Accuracy	Precision	Recall
AdaBoostClassifier	1	0.624	0.663	0.624
GradientBoosting-Classifier	1	0.631	0.666	0.631
XGBClassifier	1	0.629	0.673	0.629
MultinomialNB	1	0.738	0.739	0.738
BernoulliNB	1	0.749	0.757	0.749
SGD Classifier	1	0.753	0.760	0.753
DecisionTree Classifier	1	0.560	0.676	0.560
RandomForestClas-sifier	1	0.512	0.623	0.512
KNeighbors Classifier	1	0.694	0.739	0.694
AdaBoost Classifier	2	0.620	0.658	0.620
GradientBoosting-Classifier	2	0.632	0.665	0.632
XGBClassifier	2	0.623	0.674	0.623
MultinomialNB	2	0.746	0.747	0.746
BernoulliNB	2	0.749	0.765	0.749
SGDClassifier	2	0.755	0.761	0.755
DecisionTreeClas-sifier	2	0.561	0.680	0.561
RandomForestClas-sifier	2	0.512	0.682	0.512
KNeighborsClas-sifier	2	0.686	0.734	0.686
AdaBoostClassifier	3	0.621	0.659	0.621
GradientBoosting-Classifier	3	0.636	0.670	0.636

Algorithm	Ngram	Accuracy	Precision	Recall
XGBClassifier	3	0.626	0.665	0.626
MultinomialNB	3	0.748	0.748	0.748
BernoulliNB	3	0.749	0.769	0.749
SGDClassifier	3	0.756	0.760	0.756
DecisionTreeClassifier	3	0.561	0.681	0.561
RandomForestClassifier	3	0.509	0.672	0.509
KNeighborsClassifier	3	0.680	0.729	0.680

5. Conclusion and Future Work

Social media offers numerous advantages, primarily as a platform for global communication where individuals can share their opinions and feelings about various aspects of their lives. Given the vast and rich data it harbors, social media has recently become a focal point of extensive research. During the COVID-19 pandemic, particular attention was paid to analyzing public opinions on available vaccines. This research specifically aimed to investigate opinions regarding the Pfizer and AstraZeneca COVID-19 vaccines within the Arabic Twitter community.

Additionally, it focused on the most frequently reported side effects post-vaccination, as mentioned in the tweets. To this end, data was collected using Twitter's streaming API, cleaned, and analyzed using the Python programming language. The results aligned somewhat with expectations, revealing that the most positive tweets predominantly concerned the Pfizer vaccine across various countries. Commonly mentioned side effects included fever (حرارة) and headache (صداع).

The primary objective of this research was to analyze Arabic tweets using Machine Learning techniques and to classify sentiments using several classification algorithms. The findings indicated that the SGD Classifier was the most accurate algorithm. For future work, the plan includes collecting a more diverse dataset for analysis and conducting a deeper investigation into the reasons behind the observed results

Conflict of Interest

None declared

References

[1] Appel G., Grewal L., Hadi R., Stephen A.T. The future of social media in marketing. Journal of the Academy of Marketing Science. 2020;48:79–95.

[2] Liu, B., 2020. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge university press.

[3] Singh, R., Singh, R. and Bhatia, A., 2018. Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. Int. J. Adv. Sci. Res, 3(2), pp.19-24.

[4] Yanchang, Z. R and Data Mining: Examples and Case Studies. ISBN 978-0-12-396963-7, December 2012. Academic Press, Elsevier. 256 pages. URL: http://www.rdatamining.com/docs/RDataMining-book.pdf.

[5] World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19 - 23 October 2020. Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-

at-the-media-briefing-on-covid-19---11-march-2020

[6] Beech. Mark, 2020 , "COVID-19 Pushes Up Internet Use 70% And Streaming More Than 12%, First Figures Reveal". Available from: https://www.forbes.com/sites/markbeech/2020/03/25/covid-19-pushes-up-internet-use-70-streaming-more-than-12-first-figures-reveal/?sh=28a09afe3104

[7] M. A. B and M. Hadzikadic , 2017, "Social Computing and Social Media. Applications and Analytics," vol. 10283, pp. 191–202.

[8] Twitter, About Twitter's APIs, 2019, https://help.twitter.com/en/rules-andpolicies/twitter-api. (Accessed 5 March 2023).

[9] Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., & Al-Ohali, Y. (2017). Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. Procedia Computer Science, 117, 63-72.

[10] Hamoui, B., Alashaikh, A., & Alanazi, E. (2020, December). COVID-19: What Are Arabic Tweeters Talking About?. In International Conference on Computational Data and Social Networks (pp. 425-436). Springer, Cham.

[11] Al-Laith, A. and Alenezi, M., 2021. Monitoring People's Emotions and Symptoms from Arabic Tweets during the COVID-19 Pandemic. Information, 12(2), p.86.

[12] Al-Laith, A. and Shahbaz, M., 2021. Tracking sentiment towards news entities from Arabic news on social media. Future Generation Computer Systems, 118, pp.467-484.

[13] Alharthi, R., Alhothali, A. and Moria, K., 2021. A real-time deep-learning approach for filtering Arabic low-quality content and accounts on Twitter. Information Systems, 99, p.101740.

[14] AlGhamdi, M. A., & Khan, M. A. (2020). Intelligent Analysis of Arabic Tweets for Detection of Suspicious Messages. Arabian Journal for Science and Engineering, 1-12.

[15] El-Alfy, E. S. M., & Al-Azani, S. (2020). Empirical study on imbalanced learning of Arabic sentiment polarity with neural word embedding. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-12.

[16] Khalifa, Y., & Elnagar, A. (2020, December). Colloquial Arabic Tweets: Collection, Automatic Annotation, and Classification. In 2020 International Conference on Asian Language Processing (IALP) (pp. 163-168). IEEE.

[17] Alhumoud, S., Al Wazrah, A., Alhussain, L., Alrushud, L., Aldosari, A., Altammami, R.N., Almukirsh, N., Alharbi, H. and Alshahrani, W., 2023. ASAVACT: Arabic sentiment analysis for vaccine-related COVID-19 tweets using deep learning. PeerJ Computer Science, 9, p.e1507.