UDC 004.6.032.24:615.371 **doi: 10.32620/reks.2022.4.02**

Hassan BADI[1], Imad BADI[2], Karim El MOUTAOUAKIL[1], Aziz KHAMJANE[2], Abdelkhalek BAHRI[2]

[1]***Sidi Mohamed Ben Abdellah University, Fes, Morocco***
[2]***Abdelmalek Esaadi University, Tétouan, Morocco***

SENTIMENT ANALYSIS AND PREDICTION OF POLARITY VACCINES BASED ON TWITTER DATA USING DEEP NLP TECHNIQUES

The global impact of COVID-19 has been significant and several vaccines have been developed to combat this virus. However, these vaccines have varying levels of efficacy and effectiveness in preventing illness and providing immunity. As the world continues to grapple with the ongoing pandemic, the development and distribution of effective vaccines remains a top priority, making monitoring prevention strategies mandatory and necessary to mitigate the spread of the disease. These vaccines have raised a huge debate on social networks and in the media about their effectiveness and secondary effects. This has generated big data, requiring intelligent tools capable of analyzing these data in depth and extracting the underlying knowledge and feelings. There is a scarcity of works that analyze feelings and the prediction of these feelings based on their estimated polarities at the same time. In this work, first, we use big data and Natural Language Processing (NLP) tools to extract the entities expressed in tweets about AstraZeneca and Pfizer and estimate their polarities; second, we use a Long Short-Term Memory (LSTM) neural network to predict the polarities of these two vaccines in the future. To ensure parallel data treatment for large-scale processing via clustered systems, we use the Apache Spark Framework (ASF) which enables the treatment of massive amounts of data in a distributed way. Results showed that the Pfizer vaccine is more popular and trustworthy than AstraZeneca. Additionally, according to the predictions generated by Long Short-Term Memory (LSTM) model, it is likely that Pfizer will continue to maintain its strong market position in the foreseeable future. These predictive analytics, which uses advanced machine learning techniques, have proven to be accurate in forecasting trends and identifying patterns in data. As such, we have confidence in the LSTM's prediction of Pfizer's ongoing dominance in the industry.

Keywords: *Natural Language Processing (NLP); Machine learning; Big Data; COVID-19; Sentiment analysis; Prediction; Vaccines; Long short-term memory (LSTM); Apache Spark Framework (ASF).*

1. Introduction

Today, the massive amounts of information circulating in social networks and the enormous sharing of opinions and feelings among users have led specialists to think about creating tools for exploiting this data to identify decision-support strategies or rather carry out anthropological, political, or social-economic analyses [1, 2].

Social media are experiencing a drastic and daily change in content and the sharing of views permanently about news, especially when it comes to a phenomenon that has a global aspect like pandemics, which invites specialists to process this flow of data carefully and attentively to release all useful information that can be a subject for strategic monitoring. Twitter is one of the social networks where the daily population shares their surveys, opinions, or points of view [3 – 5].

1.1 Problematic

Since March 11, 2020, the World Health Organization (WHO) has been declaring the danger relating to the epidemic new virus (COVID-19), and it has insisted on the need for a global commitment to resist and fight the pandemic.

The WHO has created a complete guideline to slow down the dizzying spread of the pandemic while putting mechanisms and strategies in place for the search for effective and safe vaccines. But in emergency conditions and the need to react immediately, the guideline was based on public health and physical and social measures, restricting mass gatherings, closing schools, universities, shopping malls, and restaurants, and limiting travel to places where COVID-19 is spreading. Despite these precautions, humanity was developing a safe solution for resuming a normal and secure life. Everyone must protect themselves and each other against COVID-19.

This situation has forced most countries facing challenges to monitor and slow down the pandemic with all available means. Despite these precautions and even with the appearance of certain vaccines, the pandemic has not stopped spreading. Two years later, COVID-19 continues to have a substantial impact on people's daily lives and has spread across more than 210 countries and territories [6], affecting more than 607 million people in the world and resulting in 6.15 million deaths [7].

This work consists of comparing two vaccines in terms of effectiveness and popularity based on tweets using NLP algorithms.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

1.2 Big Data Analysis on Social Media

Today's world is based on digitization, which presents us with a problem we have never faced before [8]. Every device is currently connected to the Internet of Things (IoT), which means there is instant and always-on access to data collection. The flow of data circulating on social networks has the potential for companies and organizations to better understand their behavior, but big data goes even further. It is capable of helping specialists to face global problems because it provides the information needed for appropriate decision-making [4, 9, 10].

Social media analytics is one of the best examples of how big data is shaping our lives today. Scientists face a challenge in identifying decision support strategies based on the proper data use, and artificial intelligence tools [11].

1.3 Objective and approach

Twitter is used by hundreds of millions of people around the world, which gives credibility to studies carried out on the platform. Given the huge number of users, the current estimate is 330 million monthly active users and 145 million daily active users.

The main objective of this work is to carry out an exploratory and visual analysis of the tweets collected and cleaned in our dataset for the different vaccines (AstraZeneca, etc.). To highlight the number of tweets that vaccines have generated, we give the map statistic based on the number of patients that have been administered vaccines so far; see fig. 1. As of 4 September 2022, a total of 125 400 61501 vaccine doses have been administered! To determine the polarity of tweets, we first use the English word list developed by Finn rup Nielsen (AFINN) lexicon [10], which is a list of English terms containing over 3,300 words. In the second step, classify these vaccines based on the calculated polarities. The sentiment of the tweets can be positive, neutral, or negative. In the last step, calculate the future prediction behavior for polarity vaccines using the LSTM time series.

2. Related work

Different studies have been carried out for risk assessment or decision-making regarding COVID-19 in which advice to prioritize populations with social conditions is necessary for more effective control of the epidemic in its next phase and should become the norm in the planning, prevention, and mitigation of all health problems [12, 13]. In this context, Twitter has been widely used for studies relating to the COVID-19 pandemic vaccine [14, 15]. In these studies, an analysis of the sentiments and emotions of the tweets was carried out to understand the attitude of the citizens towards the waves of COVID-19 [16 – 18].

In their paper [19], the authors present the NLP (bag of words) implementation technique for message alliance, relying on the Twitter omicron infocatalog. On the Twitter omicron tweet, infocatalog uses the sentiment preparation information from the Twitter statistics set to give a plan to further extend the categorization. In this sense, the authors consider the exploration of sentiments in the Twitter omicron tweet to arrange the inquiry of all consumers, regardless of whether they are favourable, unfavourable, or dispassionate.

The authors of [13] enable sentiment analysis by utilizing Twitter data for research. First, the authors collect currently accessible tweets and hashtags about different types of COVID vaccinations posted on Twitter using the Twitter API. Then, the loaded tweets are parameterized to produce a combination of non-trained rules and random variables. To build their model, they used Tweepy, which is a wrapper for the Twitter API.

Fig. 1. The map statistic is based on the number of patients who have received a treatment

Then, as part of the sentiment analysis of new posts, the software generates donut-shaped charts. In [1], the authors investigated tweets related to coronavirus vaccines to comprehend people's feelings regarding various gender-level brands of vaccines. This approach concentrated on the impact of COVID-19 vaccines on gender by taking into account descriptive, diagnostic, predictive, and prescriptive analyses on the Twitter dataset.

These researches are much diversified and each one of them has treated the problems, related to the COVID-19 pandemic, from a different point of view using NLP and Big data techniques; the most used data source, among the social networks, is Twitter. In this work, the data are also extracted from Twitter but we will consider several aspects: (a) comparison of the two vaccines Pfizer and AstraZeneca based on the polarity index and (b) prediction of these polarities for the next period, which will allow a deep predictive comparison thanks to recursive neural networks with deep learning.

3. Methodology and implementation

Natural Language Processing (NLP) is a discipline that focuses on the understanding, manipulation, and generation of natural language by machines.

NLP is the interface between computer science and linguistics. It, therefore, relates to the capacity of the machine to interact directly with humans. The first applications of NLP were for automatic translation. A little later, the fields of application widened to include the sentimental analysis of the text in which this work is located.

Sentiment analysis involves identifying subjective information in a text to extract the author's opinion [19].

In general, the analysis of feelings makes it possible to measure the level of satisfaction of users concerning the themes or services provided by an organization or company. It has proven to be much more effective than conventional methods.

3.1 System architecture

To collect data efficiently and in a reasonable time frame, we used Apache Spark Framework (ASF), to get the maximum number of tweets in the minimum amount of time, thanks to the low latency, and their massively parallel processing framework as detailed in fig. 2.

Data collection is the process of assembling information on a particular topic in a methodical way. Collected tweets will be stored as structured and unstructured data in a data lake, which is a centralized repository, see fig. 3. In this architecture, we present steps from the collection, transformation, and visualization as a graphic.

3.2 Collection of tweets

When working with social media and especially textual data, the user community creates and manages the content. This implies that NO RULES exist! This implies that one must perform additional steps for the preparation of the collected data to ensure correct analysis. Next, we explore the text associated with a set of tweets accessed using Tweepy and the Twitter API.

This is done using standard natural language processing (also called text mining) approaches.

The first step is the collection of Tweets. To start, we must perform the following operations:

- create a standard or academic Twitter account;
- using this Twitter account, one can request developer access and then create an application that will generate the API credentials that one uses to access Twitter from Python;
- import the Tweepy package. Once the Twitter application is configured, we start by importing the Python libraries needed to access the tweets. To access the Twitter API, you need 4 Twitter application keys. These keys can be found in your Twitter app settings under the Keys and Access Tokens tab.

When we have authentificated ourselves, we launch the query to collect the data set.

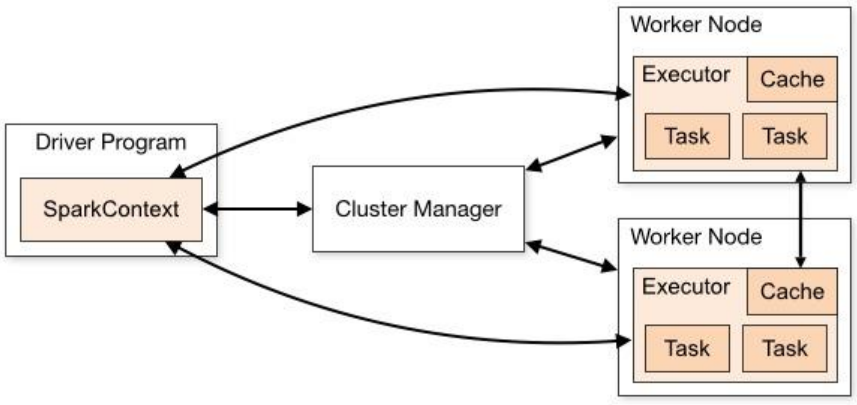

Fig. 2. Parallel processing illustration of Spark

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

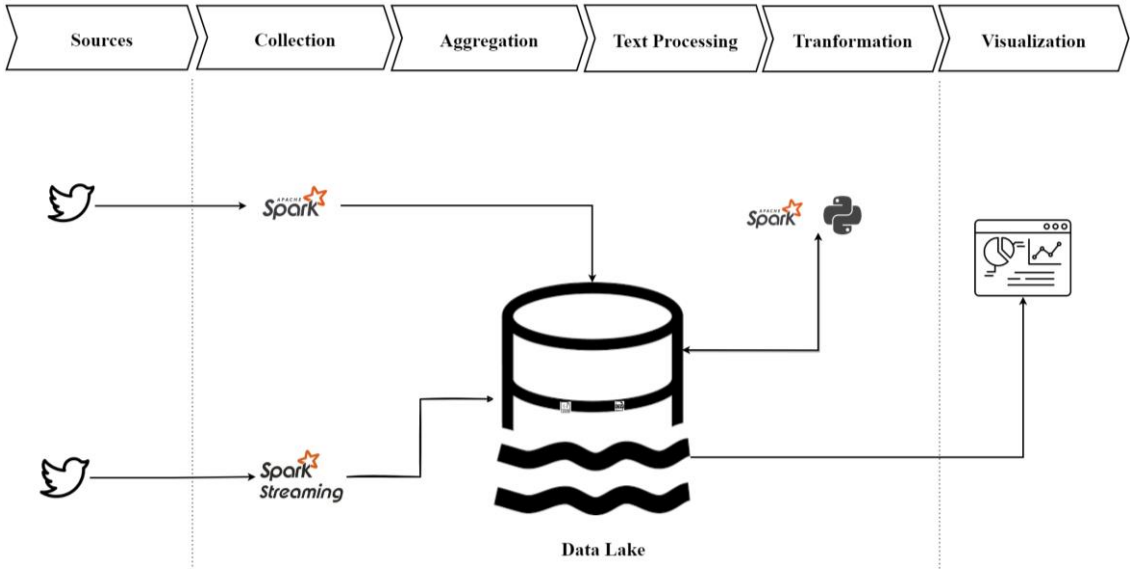

Fig. 3. Data Architecture

3.3 Text processing and transformation

We proceed with the cleaning of the tweets to analyze the frequency of the words found. Most text mining algorithms respect the following steps:

- remove Tweet URLs;
- clean up tweet text, including case differences (e.g. upper, lower) that will affect unique word counts, and remove words that are not useful for analysis;
- summarize and count words found in tweets;
- replicate the results in the form of dashboards.

The NLP method is a procedure that requires a vital stage of text preprocessing. It reconstructs text into a clear and more usable format for machine learning algorithms. Text preprocessing steps include removing retweets, URLs, and punctuation, converting emojis to words, tokenization, removing stop words, radicalizing, and removing collection words.

Retweets

A retweet is when a user shares another user's tweet. The removal of retweets is very important because duplicate tweets can distort the results of the word frequency. That is why the authors have started the treatment by removing duplicate tweets.

URLs and punctuations

In this step, we remove URLs and punctuation from tweets. Cleaning URLs is essential because it doesn't make sense and won't affect their sentimental value, but they too can be word frequency because each tweet has a link [20]. Using Python packages, you can remove unnecessary signs and punctuation from tweets, such as «?, /,! ».

Emoji

After cleaning the tweets of retweets, URLs, and punctuation, the next step is to convert the emojis to words using emoji.demojize() from the Python library. Some users use emojis to express their feelings. Therefore, suddenly, converting these emojis into sentences improves the study of sentiment analysis in tweets.

Stop Words

After separating the words from the tweets, the next step is to remove the stop words. To do this, we use the Sklearn package "stop words." Removing stopwords will be useful for calculating the sentiment of the tweet, as these stopwords are not useful for analysis. These stop words will skew a bucket of phrases from the tweet.

4. Long Short-Term Memory Layer to forecast the vaccines polarities

Among the well-known recurrent neural network in the deep learning domain, we find Long short-term memory (LSTM) [21]. Using the feedback connections, LSTM can learn static and sequential data, such as time series [22]. A common LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM is useful for prediction-making based on time series data [23].

Let i, f, g, and o be the input gate, the forget gate, the input layer, and the output gate, respectively. The diagram in Figure 3, shows the flow of data at time step t based on the network knowledge in the previous time steps.

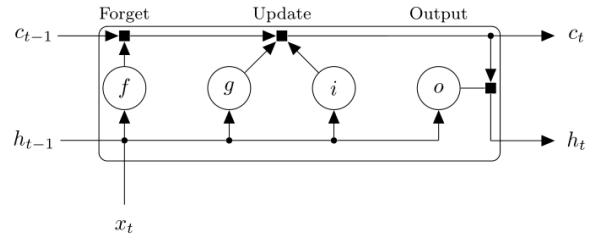

Fig. 4. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell

In figure 4, the vectors x_t, h_t, and c_t are defined as follows:

x_t: input vector of ti the LSTM unit;

h_t: hidden state vector is also known as output vector of LSTM unit;

ct : cell state vector.

The weights and the bias of the LSTM are randomly initialized. Then we can use the stochastic gradient descent algorithm, root mean square propagation algorithm or derived from the adaptive moment estimation algorithm (Adam) to set the weights and bias. In our case, we use the Adam algorithm, which uses a parameter update with momentum[24]:

$$m_l = \beta_1 m_{l-1} + (1 - \beta_1)\nabla E(\theta_l),$$

$$v_l = \beta_2 v_{l-1} + (1 - \beta_2)[\nabla E(\theta_l)]^2,$$

β_1 and β_2 decay rates and α are chosen by the user. To update the network parameters, Adam uses the equation:

$$\theta_{l+1} = \theta_l - \frac{\alpha m_l}{\sqrt{v_t} + \varepsilon},$$

where ε is a small non-negative real number.

In the experimental session, we will use the LSTM to forecast the vaccines polarities for future weeks. This choice can be justified by the fact that the recurrent network can give a good prediction based on a few data points.

5. Experimental Result and Discussion

This study was made with the aim of making a comparison between two types of vaccines widely used by most countries, based on tweets from social networks (Twitter). A sentimental study was carried out using the PNL algorithm.

In the first instance, we assigned sentimental analyses to tweets about two vaccines: AstraZeneca and Pfizer. In our case, we treated the polarity of the words related to each separate vaccine (negative, neutral, or positive).

5.1 System classification

The architecture of the system is illustrated in fig. 5, which includes the different steps mentioned above.

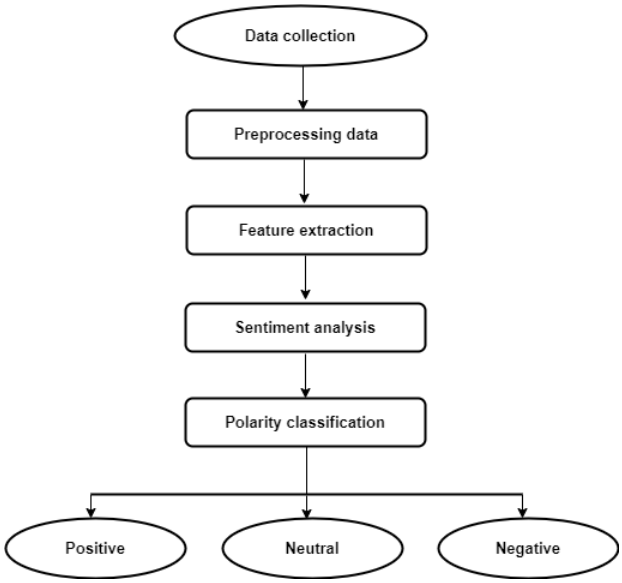

Fig. 5. System Architecture

We start with the step of collecting data on Twitter to build our own dataset, then we proceed to the data preprocessing phase before extracting the useful characteristics for the evaluation, and finally, we move on to a thorough examination of the feelings and characterization of the polarity [1].

In the present study, we performed a sentiment analysis for the data we collected during the months of September and November 2021. From figure 5, we notice that the word that has more negative feelings is that of AstraZeneca (2974 negative words).

5.2 Observation and analysis of results

In this study, we built a dataset from the collected and cleaned tweets to perform our analysis. One way to analyze social media data, especially related to events such as pandemics, is to calculate the frequency and the polarity of words related. Figure 6 gives the most common negative words in tweets.

The most important thing to remember when carrying out a sentiment analysis is to detect whether a statement is subjective or objective in nature. If the given line is considered subjective, it is necessary to determine its polarity (i.e., whether positive, negative, or neutral) [25]; see figures 7 and 8.

This section of the study presents the results of two keywords, "Pfizer" and "AstraZeneca" in the same period, as well as an emotive analysis of Twitter comments. In this case, subjectivity determines whether a word is

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

subjective or objective. On the other hand, polarity tells us about a person's positive and negative responses to a keyword or phrase. The zero point separates negative and positive feedback. We notice that positive feedback for the Pfizer vaccine is higher than negative feedback in the case of the AstraZeneca vaccine, and the percentage of positive tweets is higher than the percentage of negative tweets for the Pfizer vaccine, unlike that of AstraZeneca. The number of tweets processed after cleaning for the Pfizer vaccine is very large compared to that of AstraZeneca.

A diagnostic analysis explores ideas using information from different attributes during a sentimental analysis of tweets, it can be seen that there are high-frequency and widely used words related to vaccines that are collected to explore the general understanding of human attitudes toward different vaccines.

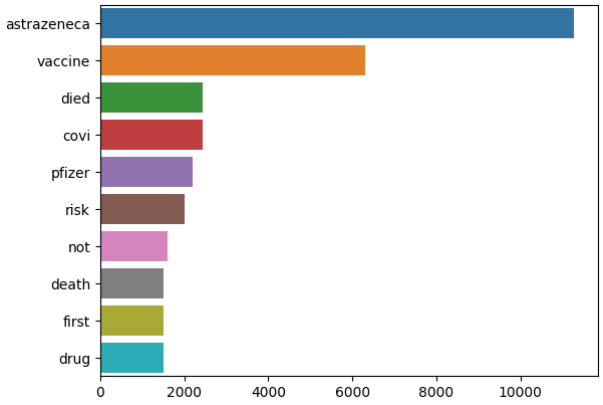

Fig. 6. Most common negative words in tweets (AstraZeneca)

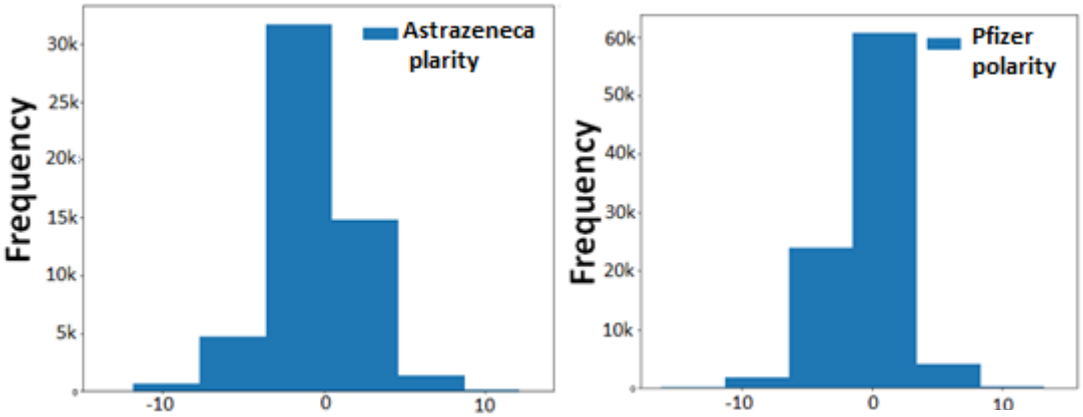

Fig. 7. Distributions of Tweets by Polarity for AstraZeneca and Pfizer vaccines

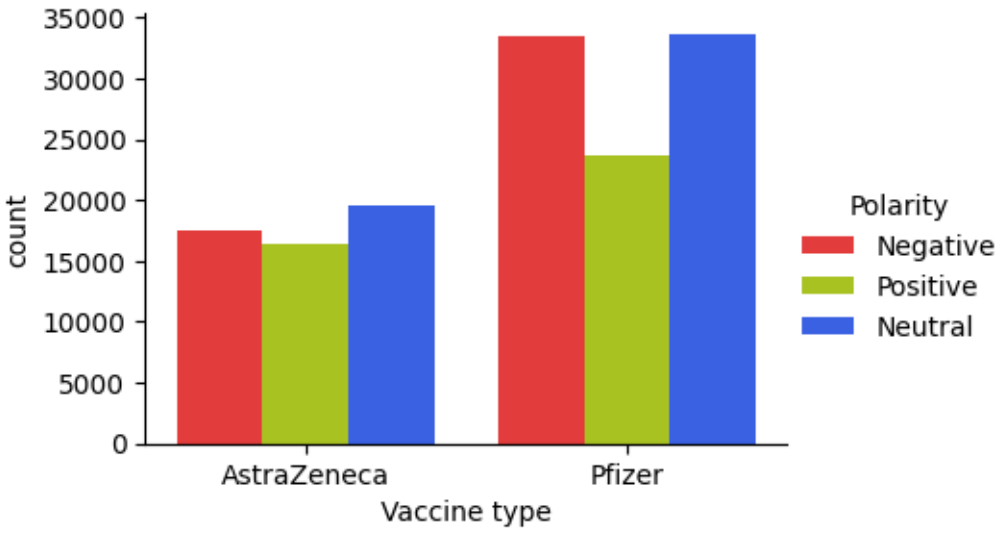

Fig. 8. Polarity of Tweets for AstraZeneca and Pfizer vaccines

Examples of diagnostic analysis are shown in fig. 8. Tweets with positive sentiments frequently contain words with a score of "> 0", while tweets with negative sentiments frequently contain words with a score of "0."

Tweets with a positive score about a vaccine indicate that it has shown its effectiveness and that it can be counted on for good protection against the pandemic, while those with a negative score give a signal to the concerned for improvement or review, knowing what people like or dislike can lead to more confident and clear decision-making.

Figure 9 gives the word count for positive tweets for AstraZeneca and Pfizer vaccines. Figure 10 gives the time series of different vaccines' polarities. From this figure, we observe Pfizer vaccine has more credibility and is more popular than AstraZeneca in the studied period.

We used LSTM to forecast the polarities of different vaccines in the coming weeks. In recent weeks, we considered the polarities of these vaccines and divided them into learning (80 %) and testing (20 %) data set.

The LSTM algorithm was configured as follows: number of Hidden neurons (200); learning algorithm (Adam); MaxEpochs (250); Gradient Threshold (1); Initial Learn Rate (0.005); Learn Rate Schedule (piecewise); Learn Rate Drop (125); Learn Rate Drop Factor (0.2). We use Root-mean-square error (RMSE) to measure the performance of the proposed system that compares the predicted polarities to the right ones.

Figures 11 and 12 give, respectively, the evolution of the loss and RMSE with time. We notice that the LSTMs associated with Pfizer and AstraZeneca start to converge after 150 epochs.

Concerning the predictions of the polarities of the two vaccines for the coming weeks, figures 13 and 14 give, respectively, the obtained predictions using the trained LSTMs. In this sense, we note that the polarities of the two vaccines will have different behaviors in the future. Indeed, Pfizer will maintain its dominance over AstraZeneca except that, toward the end of the prediction, the popularity of the second vaccine will gain several points at the expense of Pfizer.

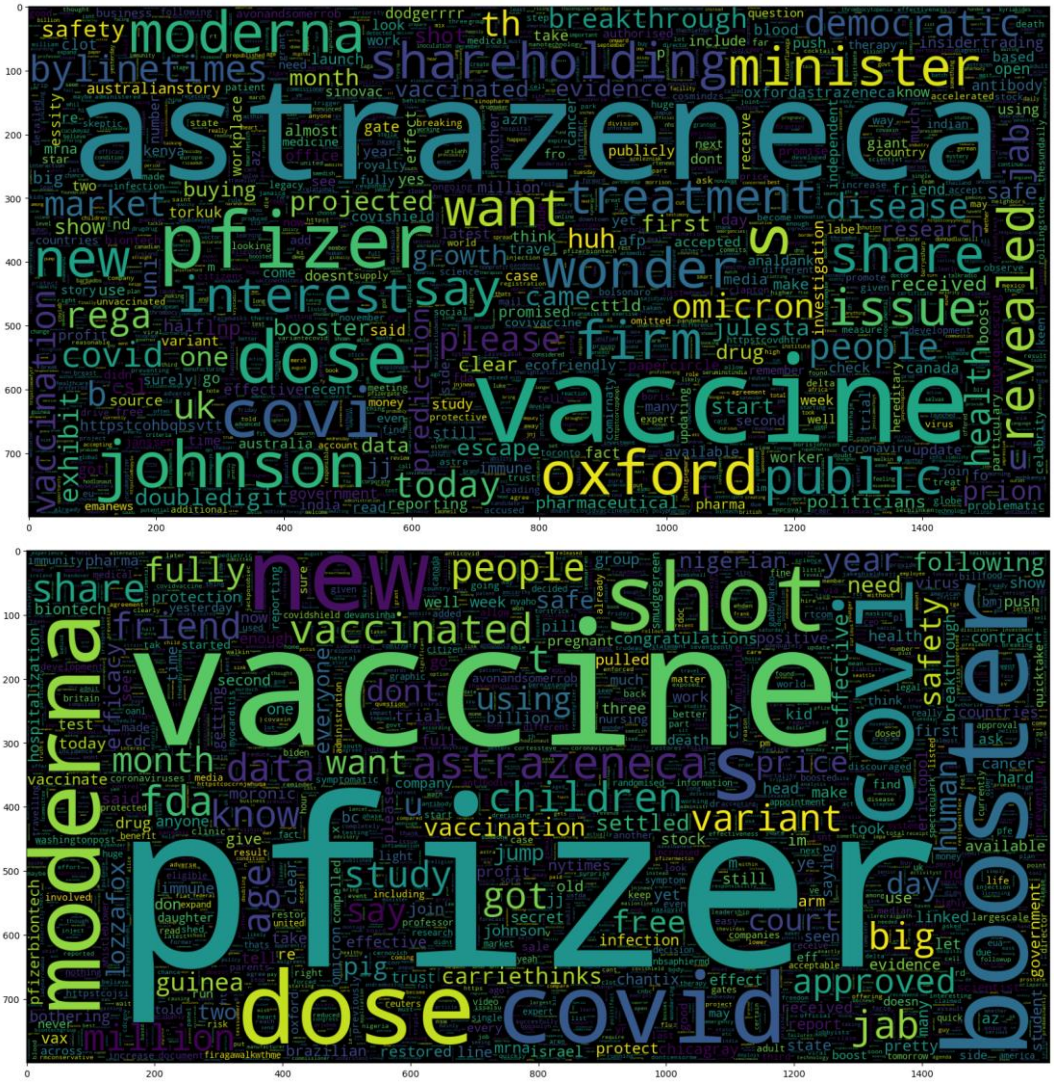

Fig. 9. Word Count for Positive Tweets for AstraZeneca and Pfizer vaccines

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

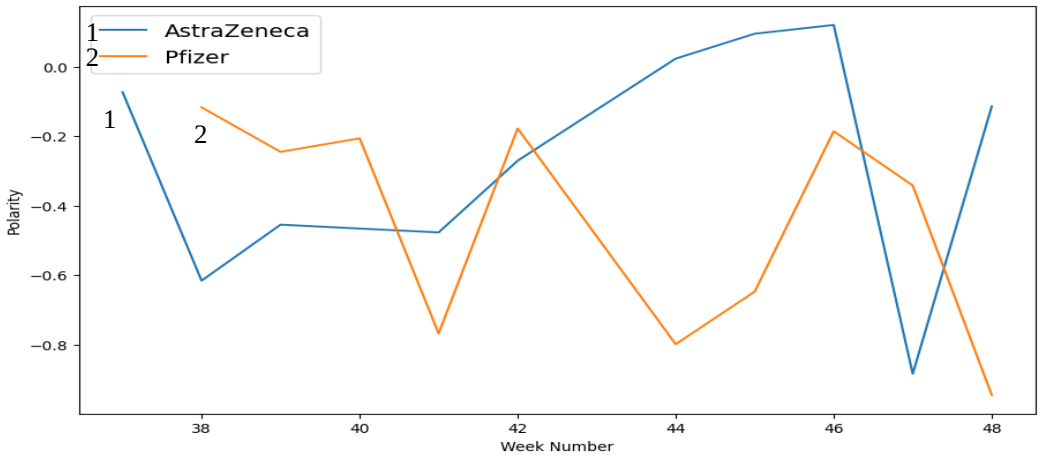

Fig. 10. Time series of vaccines polarity

Fig. 11. LSTM performance on Pfizer polarity

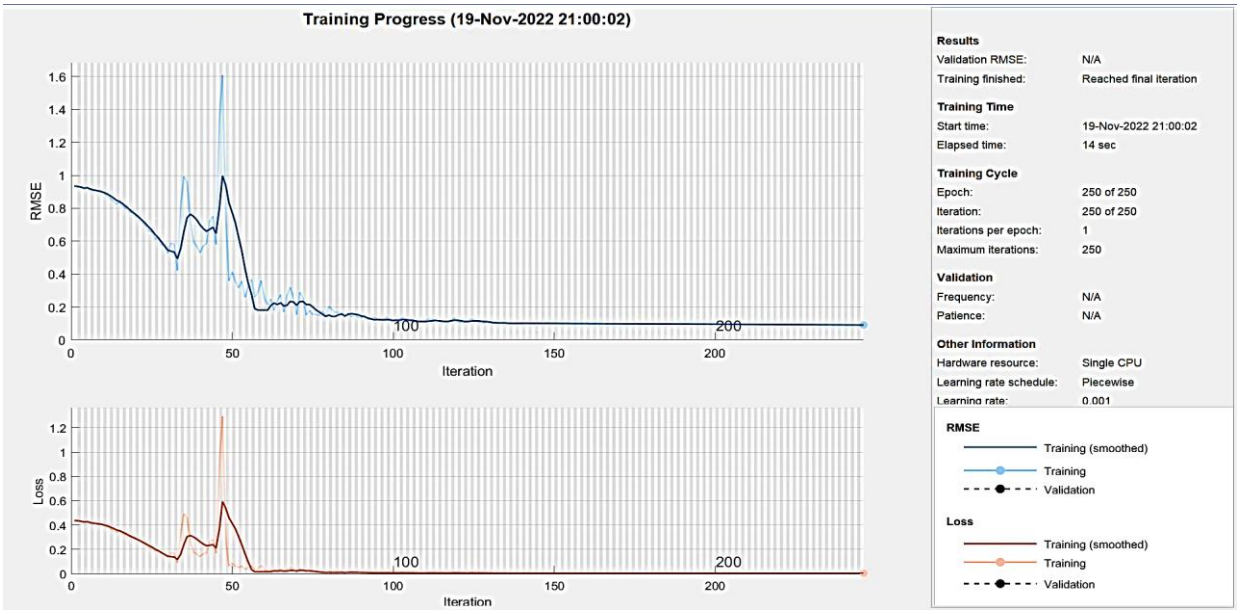

Fig. 12. LSTM performance on AstraZeneca polarity

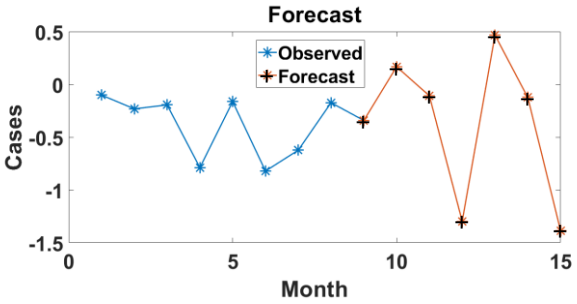

Fig. 13. Prediction on Pfizer polarity

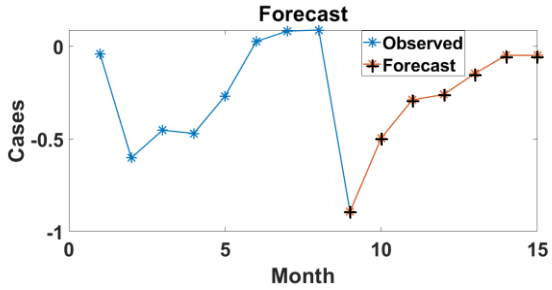

Fig. 14. Prediction on AstraZeneca

6. Conclusion

The development of vaccines was an urgent and necessary necessity at a time when the world was suffering from an exceptional health crisis, so it was the main objective of scientific research in the medical field, from which a lot of material and immaterial resources have been reversed. Along with preventive measures, the vaccine helps to control and halt the dizzying spread of COVID-19 on a large scale. Meanwhile, vaccine hesitancy is another problem to be addressed, especially with the distrust and lack of trust on the part of citizens.

In the present study, we did a sentimental analysis of tweets from a properly integrated dataset collected from Twitter. This analysis was based on natural language processing (NLP) using the AFINN lexicon, which is a list of English terms containing more than 3,300 words. Each word is associated with a partition of sentiment. This study allows us to visualize the attitudes of people toward the Pfizer and AstraZeneca vaccines over a very specific period. We have also applied the LSTM algorithm to predict the polarities of these vaccines for several future weeks based on current polarities. We note that the polarities of the two vaccines will behave differently in the future. In addition, Pfizer will continue to outperform AstraZeneca, but near the end of the forecast, the popularity of the second vaccine will gain several points over Pfizer.

Contribution of authors

Creation of dataset from social media network, implementation of the sentiment analysis algorithm, analysis of the literature, redaction of the paper – **Hassan Badi**; definition of the problematic, redaction of the paper – **Imad Badi**; setting and substantiation of the purpose and objectives of the study, Implementation of LSTM to predict the polarities of different vaccines – **Karim El Moutaouakil**; research methodology and presentation of results. Revision of the document –**Aziz Khamjane** and **Abdelkhalek Bahri**.

All the authors have read and agreed to the published version of the manuscript.

References

1. Shahriar, K. T., Islam, M. N., Anwar, M. M. and Sarker, I. H. COVID-19 analytics: Towards the effect of vaccine brands through analyzing public sentiment of tweets. *Informatics in Medicine Unlocked*, 2022, vol. 31, article no. 100969. DOI: 10.1016/j.imu.2022.100969.

2. Bibi, M., Abbasi, W. A., Aziz, W., Khalil, S., Uddin, M., Iwendi, C., Gadekallu, T. R. A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. *Pattern Recognition Letters*, 2022, vol. 158, pp. 80-86. DOI: 10.1016/j.patrec.2022.04.004.

3. Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. Sentiment Analysis of Twitter Data. *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 2011, pp. 30-38. Available at: https://aclanthology.org/W11-0705. (accessed March 20, 2022)

4. Sunitha, D., Patra, R. K., Babu, N. V., Suresh, A. and Gupta, S. C. Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries. *Pattern Recognition Letters*, 2022, vol. 158, pp. 164-170. DOI: 10.1016/j.patrec.2022.04.027.

5. Joshi, M., Prajapati, P., Shaikh, A. and Vala, V. A Survey on Sentiment Analysis. *International Journal of Computer Applications*, 2017, vol. 163, no. 6, pp. 34-38. DOI: 10.5120/ijca2017913552.

6. Chumachenko, D., Pyrohov, P., Meniailov, I. and Chumachenko, T. Impact of war on COVID-19 pandemic in Ukraine: the simulation study. *Radioelectronic and Computer Systems*, 2022, no. 2, pp. 6-23. DOI: 10.32620/reks.2022.2.01.

7. *About Worldometer*. Available at: https://www.worldometers.info/about/ (accessed March 28, 2022).

8. Zhang, H., Zang, Z., Zhu, H., Uddin, M. I. and Amin, M. A. Big data-assisted social media analytics for business model for business decision making system competitive analysis. *Information Processing & Management*, 2022, vol. 59, iss. 1, article no. 102762. DOI: 10.1016/j.ipm.2021.102762.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

9. Chen, Y.-J. and Chen, Y.-M. Forecasting corporate credit ratings using big data from social media. *Expert Systems with Applications*, 2022, vol. 207, article no. 118042. DOI: 10.1016/j.eswa.2022.118042.

10. Chumachenko, D., Chumachenko, T., Kirinovych, N., Meniailov, I., Muradyan, O. and Salun, O. Barriers of COVID-19 vaccination in Ukraine during the war: the simulation study using ARIMA model. *Radioelectronic and Computer Systems*, 2022, no. 3, pp. 20-35. DOI: 10.32620/reks.2022.3.02.

11. Wongkoblap, A., Vadillo, M. A. and Curcin, V. 6 - Social media big data analysis for mental health research. *Mental Health in a Digital World*, Academic Press Publ., 2022, pp. 109-143. DOI: 10.1016/B978-0-12-822201-0.00018-6.

12. Afifi, R. A. et al. 'Most at risk' for COVID19? The imperative to expand the definition from biological to social factors for equity. *Preventive Medicine*, 2020, vol. 139, article no. 106229. DOI: 10.1016/j.ypmed.2020.106229.

13. Chinnasamy, P., Suresh, V. et al. COVID-19 vaccine sentiment analysis using public opinions on Twitter. *Materials Today: Proceedings*, 2022, vol. 64, Part 1, pp. 448-451. DOI: 10.1016/j.matpr.2022.04.809.

14. Nezhad, Z. B. and Deihimi, M. A. Twitter sentiment analysis from Iran about COVID 19 vaccine. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 2022, vol. 16, no. 1, article no. 102367. DOI: 10.1016/j.dsx.2021.102367.

15. Paul, S. Analyzing the attitude of Indian citizens during the second wave of COVID-19: A text analytics study. *International Journal of Disaster Risk Reduction*, 2022, vol. 79, article no. 103161. DOI: 10.1016/j.ijdrr.2022.103161.

16. Anastasiou, D., Ballis, A. and Drakos, K. Constructing a positive sentiment index for COVID-19: Evidence from G20 stock markets. *Social Science Research Network*, 2021, 38 p. DOI: 10.2139/ssrn.3895548.

17. Huynh, T. L. D., Foglia, M., Nasir, M. A. and Angelini, E. Feverish sentiment and global equity markets during the COVID-19 pandemic. *Journal of Economic Behavior & Organization*, 2021, vol. 188, pp. 1088-1108. DOI: 10.1016/j.jebo.2021.06.016.

18. Sv, P., Tandon, J., Vikas, and Hinduja, H. Indian citizen's perspective about side effects of COVID-19 vaccine – A machine learning study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 2021, vol. 15, iss. 4, article no. 102172. DOI: 10.1016/j.dsx.2021.06.009.

19. Hosgurmath, S., Petli, V. and Jalihal, V. K. An Omicron Variant Tweeter Sentiment Analysis Using NLP Technique. *Global Transitions Proceedings*, 2022, vol. 3, iss. 1, pp. 215-219. DOI: 10.1016/j.gltp.2022.03.025.

20. Zulfiker, M. S., Kabir, N., Biswas, A. A., Zulfiker, S. and Uddin, M. S. Analyzing the public sentiment on COVID-19 vaccination in social media: Bangladesh context. *Array*, 2022, vol. 15, article no. 100204. DOI: 10.1016/j.array.2022.100204.

21. Kudo, M., Toyama, J. and Shimbo, M. Multidimensional curve classification using passing-through regions. *Pattern Recognition Letters*, 1999, vol. 20, no. 11-13, pp. 1103-1111. DOI: 10.1016/S0167-8655(99)00077-X.

22. Baytas, I. M., Xiao, C., Zhang, X. S., Wang, F., Jain, A. K. and Zhou, J. Patient Subtyping via Time-Aware LSTM Networks. *KDD '17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 65-74. DOI: 10.1145/3097983.3097997.

23. Voelker, A. R., Kajić, I. and Eliasmith, C. Legendre Memory Units: Continuous-Time Representation in Recurrent Neural Networks. *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, Canad, 2019, article no. 1395, pp. 15570-15579.

24. Kingma, D. P.and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv*, 2014. 15 p. DOI: 10.48550/ARXIV.1412.6980.

25. Hamzah, F. A. et al. *CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction. 2020*. Available at: https://www.researchgate.net/publication/340032869_CoronaTracker_World-wide_COVID-19_Outbreak_Data_Analysis_and_Prediction. (accessed March 28, 2022).

Надійшла до редакції 17.09.2022, розглянута на редколегії 20.11.2022

АНАЛІЗ СЕНТИМЕНТУ ТА ПРОГНОЗ ПОЛЯРНОСТІ ВАКЦИН НА ОСНОВІ ДАНИХ ТВІТТЕРА З ВИКОРИСТАННЯМ ТЕХНІК ГЛИБОКОГО NPL

Хасан Баді, Імад Баді, Карім Ель Мутуакіл, Азіз Хамджане, Абдельхалек Бахрі

Глобальний вплив COVID-19 був значним, і для боротьби з цим вірусом було розроблено кілька вакцин. Однак ці вакцини мають різний рівень ефективності та результативності в профілактиці захворювань і

забезпеченні імунітету. Оскільки світ продовжує боротися з пандемією, яка триває, розробка та розповсюдження ефективних вакцин залишаються головним пріоритетом, що робить стратегії моніторингу профілактики обов'язковими та необхідними для пом'якшення поширення цієї хвороби. Ці вакцини викликали бурхливу дискусію в соціальних мережах і ЗМІ щодо їх ефективності та вторинних ефектів. Це призвело до створення великих даних, що потребує інтелектуальних інструментів, здатних глибоко аналізувати ці дані та витягувати основні знання та відчуття. Мало робіт, які одночасно здійснюють аналіз почуттів і прогнозування цих почуттів на основі їх оціночної полярності. У цій роботі, по-перше, ми використовуємо великі дані та інструменти обробки природної мови (NLP), щоб виділити сутності, виражені в твітах про AstraZeneca та Pfizer, і оцінити їх полярності; по-друге, ми використовуємо нейронну мережу довготривалої короткочасної пам'яті (LSTM), щоб передбачити полярності цих двох вакцин у майбутньому. Щоб забезпечити паралельну обробку даних для великомасштабної обробки через кластерні системи, ми використовуємо Apache Spark Framework (ASF), яка дає змогу розподілено обробляти великі обсяги даних. Результати показали, що вакцина Pfizer є більш популярною та надійнішою, ніж AstraZeneca. Крім того, згідно з прогнозами, створеними моделлю довгострокової короткочасної пам'яті (LSTM), цілком імовірно, що Pfizer продовжить зберігати свої сильні позиції на ринку в осяжному майбутньому. Ця прогностична аналітика, яка використовує передові методи машинного навчання, довела свою точність у прогнозуванні тенденцій і виявленні закономірностей у даних. Таким чином, ми впевнені в прогнозі LSTM про постійне домінування Pfizer у галузі.

Ключові слова: обробка природної мови (NLP); машинне навчання; великі дані; COVID-19; аналіз настроїв; прогнозування; вакцини; довга короткочасна пам'ять (LSTM); Apache Spark Framework (ASF).

Хасан Баді – PhD, Лабораторія інженерних наук, Мультидисциплінарний факультет Таза, Університет Сіді Мохамеда Бен Абделла, Фес, Марокко.

Імад Баді – проф., Департамент математики та інформатики Університету Абдельмалека Есааді, Лабораторія прикладних наук Марокко.

Карім Ель Мутауакіл – проф., Лабораторія інженерних наук, Мультидисциплінарний факультет Тази, Університет Сіді Мохамеда Бен Абделлаха, Фес, Марокко.

Азіз Хамджане – проф., Департамент математики та інформатики Університету Абдельмалека Есааді, Лабораторія прикладних наук Марокко.

Абдельхалек Бахрі – проф., Департамент математики та комп'ютерних наук Університету Абдельмалека Есааді, Лабораторія прикладних наук Марокко.

Hassan Badi – PhD, Laboratory of Engineering Sciences, Multidisciplinary faculty of Taza, Sidi Mohamed Ben Abdellah University, Fes, Morocco,
e-mail: hassan.badi@usmba.ac.ma, ORCID: 0000-0002-1568-9790.

Imad Badi – Prof., National School of Applied Sciences Al-Hoceima, Laboratory of Applied Sciences Al-Hoceima, Abdelmalek Esaadi University, Tétouan, Morocco,
e-mail: ibadi@uae.ac.ma, ORCID: 0000-0002-2844-4629.

Karim El Moutaouakil – Prof., Laboratory of Engineering Sciences, Multidisciplinary faculty of Taza, Sidi Mohamed Ben Abdellah University, Fes, Morocco,
e-mail: karim.elmoutaouakil@usmba.ac.ma, ORCID: 0000-0003-3922-5592.

Aziz Khamjane – Prof., National School of Applied Sciences Al-Hoceima, Laboratory of Applied Sciences Al-Hoceima, Abdelmalek Esaadi University, Tétouan, Morocco,
e-mail: akhamjane@uae.ac.ma, ORCID: 0000-0002-3508-8968.

Abdelkhalek Bahri – Prof., National School of Applied Sciences Al-Hoceima, Laboratory of Applied Sciences Al-Hoceima, Abdelmalek Esaadi University, Tétouan, Morocco,
e-mail: abahri@uae.ac.ma, ORCID: 0000-0002-8527-7281.