CoreID JOURNAL

Vol. 2 No. 2, July 2024 (72-78)
ISSN: 2987-6990
DOI: https://doi.org/10.60005/coreid.v2i2.33

Long Short Term Memory Approach for Sentiment Analysis on COVID-19 Vaccination Policy

Fauzan Herdika Tubagus Putra[1], Wildan Budiawan Zulfikar[2], Nur Lukman[3]
[1,2,3]Department of Informatics UIN Sunan Gunung Djati Bandung, Indonesia.

Article Info

Article history:

Received June 21, 2024
Revised July 23, 2024
Accepted July 26, 2024

Keywords:

K-Fold Cross Validation
LSTM
Word2Vec
Sentiment Analysis

ABSTRACT

COVID-19 vaccination is one of the efforts to reduce the spread of COVID-19 and reduce the impact or severe symptoms of COVID-19. On social media, many Indonesians have expressed their opinions regarding the COVID-19 vaccine. With technology, we can classify Indonesian public opinion on the COVID-19 vaccine on social media, including pros or cons. Sentiment analysis using the LSTM (Long Short Term Memory) algorithm is one way. The data that has been taken will go through a cleaning and weighting process using Word2Vec before entering the LSTM algorithm. With the evaluation method of the K-Fold Cross Validation model, we can determine the performance of this LSTM algorithm. The results of the performance of this LSTM model show an average accuracy of 74.1% and have the best accuracy in the 4th Fold, which is 81%. The data that has been taken will be tested on this best model, and the results of the sentiment analysis of Indonesian public opinion on the COVID-19 vaccine are 49.4% Positive and 50.6% Negative.

Corresponding Author:

Fauzan Herdika Tubagus Putra
Department of Informatics UIN Sunan Gunung Djati Bandung,
Jl. A.H. Nasution No. 105 Cibiru, Kota Bandung, 40614, Indonesia.
Email: ojanherdika@gmail.com

1. INTRODUCTION

At the end of 2019, precisely in December, the world was shocked by the existence of a new variant of the virus that made many people anxious, known as Coronavirus Disease 2019 (COVID-19). The incident began in China, Wuhan. Initially, the virus was thought to be the result of exposure to a wholesale seafood market that sold many species of live animals. This disease quickly spread within China and also throughout the world including Indonesia. Judging from the situation of the spread since 2020, COVID- 19 has reached all provinces in Indonesia with the number of cases and or the number of deaths increasing and having an impact on the political, economic, social, cultural, defense and security aspects, as well as the welfare of the people in Indonesia, the Government of Indonesia has stipulated Presidential Decree Number 11 of 2020 concerning the Determination of Coronavirus Disease 2019 (COVID-19)

Public Health Emergency. The Presidential Decree establishes COVID-19 as a type of disease that causes Public Health Emergencies (PHE) and establishes COVID-19 PHE in Indonesia which must be carried out in accordance with the provisions of laws and regulations [1]–[5].

This virus can be stopped or reduced in various ways, one of which is by using a vaccine. The development of a safe and effective vaccine to control this pandemic is very important because it is expected to inhibit its spread and prevent its recurrence in the future. In addition, because this pandemic is spreading rapidly, a vaccine that can be produced in a fairly short time is needed, because in general, making a vaccine takes years [6]–[9].

Regarding the use of the vaccine in Indonesia itself, there are still many pros and cons, one of the Members of the House of Representatives made a number of controversial statements at a working meeting on January 12, 2021, who firmly refused to be vaccinated against COVID-19 and claimed that he chose to pay a

fine rather than get With this statement, it proves that the use of the COVID-19 vaccine is still reaping the pros and cons in Indonesian society [10].

The pandemic that occurred in Indonesia requires us as a society to carry out all activities as much as possible at home, because of that it is possible that many people express all opinions or complaints on social media. As of January 2021, 170 million Indonesians are active social media users or around 61.8% of the total Indonesian population. This active user of social media is calculated to increase its use by 10 million users from the previous year [5].

Of the many social media, in Indonesia itself, social media that is often used can be sorted based on the number of people who use social media. In the first rank is Youtube, second WhatsApp, third Instagram, fourth Facebook, fifth Twitter and many others. Although Twitter is ranked fifth compared to other social media, Twitter contains text content and the lack of advertising audiences is different from other social media. As evidenced by the data, the advertising audience on Twitter is only 6%. And with Text Mining technology, we can retrieve text data about the covid-19 vaccine on social media, especially Twitter [5].

Every data taken in the form of text about the COVID-19 vaccine on social media can be known to be good or bad through the sentiment analysis process. Sentiment analysis is a field of study that analyzes opinions, sentiments, assessments, evaluations, attitudes, and emotions in text data using a person's text analysis techniques related to a particular topic, service, product, individual, organization, or activity. Sentiment analysis is carried out to identify the sentiment of a person's opinion or comment on a problem has a positive or negative sentiment and can be used as a reference in improving the quality of products, services, individuals, organizations, or certain activities [6], [11]–[13].

With the Deep Learning method, especially the LSTM (Long Short Term Memory) algorithm where this algorithm is suitable for text data [14]–[16]. Because the LSTM (Long Short Term Memory) algorithm has several advantages, namely that it can store history or trace data if the data is not used in a process, but the data can be reused if needed at any time for the next process [17]–[20]. So with these advantages, this method can analyze the sentiment of public opinion in the form of text or words that are pro (positive) or contra (negative) more accurately, effectively and efficiently.

2. METHOD

2.1 Data Understanding

This stage focuses on what problem to solve or what question to answer. In a data analysis, understanding the business is a very important stage. In Indonesia, there are about 170 million people who use social media (as of January 2021). With the PSBB, the use of the internet and social media has increased from the previous year. And many people often express opinions or comments regarding certain cases or topics. For example on twitter. Because twitter data is in the form of text tweets, this twitter is used in sentiment analysis processing. Therefore, to find out and overcome the problems of complaints or public comments that often occur related to the COVID-19 vaccine, a system is proposed that can analyze the opinions and comments of social media users including into positive or negative social media.

With a system that can analyze the sentiment of public opinion on the COVID-19 vaccine on twitter social media, whether the results are negative or positive, these results can be used as a reference in improving the quality of services, organizations, or certain activities. In addition, the results can later be used to help select decisions regarding what steps to take next.

2.2 Data Understanding

The data that will be taken in this sentiment analysis research is taken by utilizing the API (Application Programming Interface) facility provided free of charge by twitter by registering as a twitter developer beforehand. With this facility, the process of scrapping data or retrieving tweets data is easier than with other social media.

The data that will be scrapped from twitter is data about the COVID-19 vaccine on May 30 to 31, 2021. Because at

that time the vaccination rate in Indonesia was still far from the target national vaccination rate (181,554,465 people) while the number of people who had vaccinated was calculated at 27,045,507 people on May 31, 2021, and the number of people who vaccinated decreased from the previous days.

As the national vaccination target was not met by the number of people who had vaccinated at that time and the lack of people who vaccinated at that time, the data to be scrapped on that date is suitable for this study to be used as input to the LSTM model to be built. Collected 999 documents from scrapping this data, which will then be labeled manually positive with negative.

Meanwhile, the data that will be searched for the results of sentiment analysis is the combined data between data after the COVID-19 vaccine is spread and before the COVID-19 vaccine is spread. In December 16-19 2020 before the vaccine was spread, only about 37% of Indonesians stated that they would vaccinate

against COVID-19 if the vaccine was available, the remaining 17% would not, and 40% thought first. As many as 2000 data were collected from this combined data, which will then be used as input for the LSTM model that has been trained with data for May 30-31, 2022. After being inputted, this model will issue predictions on the results of analyzing the sentiment of the Indonesian people towards the COVID-19 vaccine. In addition to the combined data, the model that has been trained will also be tested with data on various types of vaccine brands such as AstraZeneca, Sinovac and Pfizer. Then the model will issue predictions of the results of sentiment analysis on these various vaccine brands.

2.3 Data Preparation

Before heading to the next stage, the data must be labeled. Where for each negative comment is labeled "0" while for positive comments is labeled "1". The amount of data that has been labeled negative and positive, as in fig 1 below:

```
1    539
0    460
Name: label, dtype: int64
```

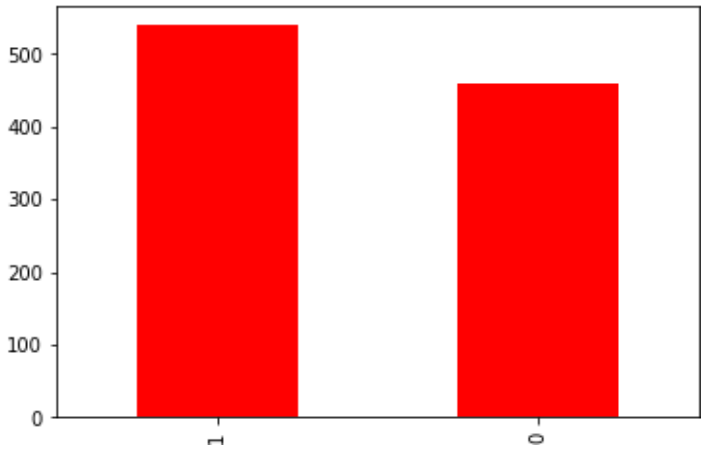

Figure 1. Number of Positive and Negative Label of Dataset

The amount of data that has been labeled negatively and positively is not much difference between negative and positive, only about 79 documents or about 7.9% of the total data. So that when this data is processed the possibility of overfitting a label will be smaller when compared to the amount of data that is not balanced between positive and negative.

The preprocessing stages used include tokenizing, casefolding, regex removing, stopwords, and stemmming. The following is an example of implementation at the preprocessing stage:

Table 1. Sample of Preprocessing

Proses	Result
Tokenizing	'RT' '@JPenerangan:' 'Vaksin' 'COVID-19' 'sangat' 'dianjurkan' 'untuk' 'mengurangi' 'gejala' 'COVID-19'
Casefolding	'rt' '@jpenerangan:' 'vaksin' 'covid-19' 'sangat' 'dianjurkan' 'untuk' 'mengurangi' 'gejala' 'covid-19'
Regex Removing	vaksin covid sangat dianjurkan untuk mengurangi gejala covid
Stopword	vaksin covid sangat dianjurkan mengurangi gejala covid
Stemming	vaksin covid sangat dianjurkan mengurangi gejala covid

Word weighting is done after the data has gone through the cleaning process. This data will be represented in the form of vectors and create vocabularies with training data models. Word2vec has several different dimensions, some are 300, 250, 200, 100, 50. This study uses a dimension of 50 because the number of documents in the dataset is not so much only about 900 train data and 100 test data. After producing vectors and matrices, the data is ready to be processed at the next stage, namely as input to the model of the algorithm that this research uses, namely Long Short Term Memory.

Table 2. Word2vec Result

Words	Result
'vaksin'	array([0.24500501, 0.0902098 , -0.20089197, -0.01869769 , 0.17395455, 0.11627325, 0.07738613, 0.13911141, -0.01105479, 0.145 37545, 0.23377107, 0.3468209 , 0.05481519, -0.05975091, -0.032 90469, 0.22630256, 0.02026901, 0.1273412 , -0.08736784, 0.251 87492, -0.03066862, -0.2675387 , 0.08699781, 0.11180223, -0.00 380542, 0.05280958, -0.18406153, -0.16166554, -0.09523913, 0.03 162546, 0.02122332, -0.11426074, 0.138523 , 0.08397825, 0.0381 3726, -0.08696777, -0.11997165, -0.14710099, -0.23615319, -0.1 6981688, -0.03978463, 0.03761901, -0.15567556, -0.12094911, 0.01 363597, -0.04471155, 0.07599332, 0.2758858 , -0.12551658, 0.05 177052], dtype=float32)
'aman'	array([-0.11729756, 0.27817184, 0.1058146 , 0.10125296, 0.05426468, -0.13822293, 0.08791998, 0.26641732, -0.13501757, 0.00 531872, 0.24590042, 0.11126411, -0.05601184, -0.16899724, 0.03 294804, 0.04638429, -0.19432683, 0.08012074, 0.00966968, 0.072 33725, -0.18399319, 0.05424116, -0.15434904, 0.12575826, 0.14 421642, 0.10227595, -0.06958544, 0.04340191, 0.07185905, 0.198 29726, 0.08549559, 0.18439698, 0.3243559 , 0.20644447, -0.141 5537 ,

2.4 Modeling

There are several layers in creating a sentiment analysis model in this research, namely Embedding Layer, Dropout, LSTM, Dense Sigmoid. After going through the word weighting process, words have their respective value vectors and words that have the same or similar meaning will have similar values so as to facilitate the next input process. Then the next layer is Dropout, this is to reduce the potential for overftting in the model created. This dropout function discards networks that are not needed in existing layers. The value vector obtained from the embedding layer becomes an input to the LSTM model, which will be spread over 3 gates. As shown in fig 2 below:

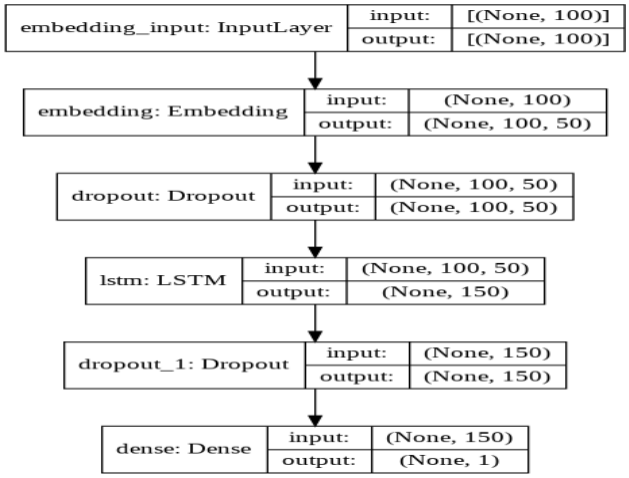

Figure 2. Rancangan Arsitektur Model LSTM

3. RESULT AND DISCUSSION

Evaluation is done to adjust whether the results of the modeling that has been made previously meet business understanding or not. Before the evaluation, a testing process is carried out first. In this research, testing uses the K-Fold Cross Validation method. Which is divided into 10 folds, 1 fold consists of 10% test data or about 100 data from the overall dataset and the other 9 folds consist of 90% train data or about 900 data from the overall dataset.

This process uses the k-flod cross validation method and confusion matrix. Where all data will be divided into 10 partitions, testing will be carried out 10 times where each test of one partition will become testing data and the rest will become training data. The results of testing and evaluation that have been carried out can be seen in table 3:

Table 3. Summary Of Test Case

K	*Precission* (%)		*Recall* (%)		F1 – *Score* (%)		Accuration (%)
	1	0	1	0	1	0	
1	82	67	70	79	75	72	74
2	73	74	55	87	63	80	74
3	73	84	85	72	78	78	78
4	80	82	78	83	79	83	81
5	67	81	76	72	71	76	74
6	74	68	51	85	61	76	70
7	66	87	82	74	73	80	77
8	62	71	64	70	63	70	67
9	73	63	65	72	69	67	68
10	74	83	83	73	78	78	78
Avg	72,4	76	70,9	76,7	71	76	74,1

Based on table 3, the average accuracy result of the modeling is 74.1%. The accuracy shows that overall the model can classify sentiment quite well. Factors that affect accuracy involve all true negatives, true positives, false negatives, and false positives. Then the factor that affects precision positive is how many true positives and false positives while precision negative is how many true negatives and false negatives. The factor that affects recall positive is how many true positives and false negatives while recall negative is how many true negatives and false positives.

Based on the results of the 10 tests above, the best model was obtained, with the highest accuracy of 81% at the 4th K- Fold. In training the model produces accuracy that goes up and loss that goes down. However, validation accuracy and validation loss are not as good as accuracy and loss in training as in figure 3 below:

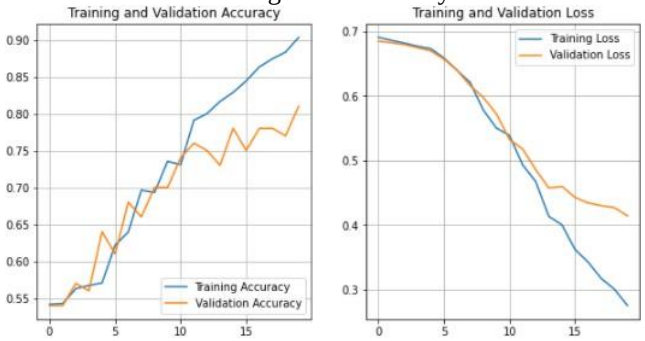

Figure 3. Training and Validation Graph

The contributing factor is the lack of data, both test and train, then the data used in the training process does not match the testing data, which can result in decreased accuracy and high loss. As well as language confusion, because there is some data that uses mixed language between English and Indonesian, causing confusion.

With the best model in this 4th K-Fold, this model will be saved and will be tested with combined data, and data on various types of vaccine brands. Furthermore, the results of the trial with this combined data will be taken as the results of the analysis of public sentiment towards the COVID-19 vaccine. The following are the predictions of the sentiment analysis results of the best model that has been made:

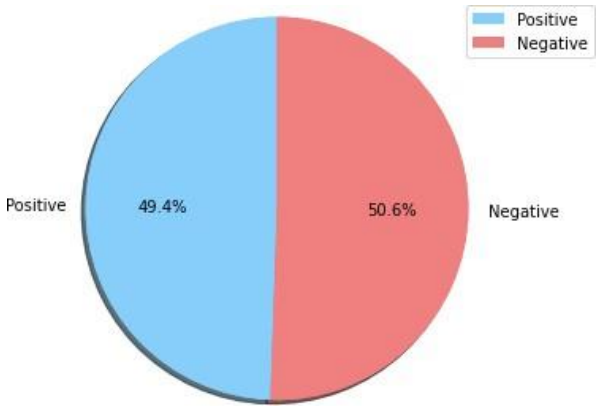

Figure 4. Results of Analysis of Public Sentiment towards the COVID-19 Vaccine

Fig. 5 explains that the results of the analysis of public sentiment towards the COVID-19 vaccine are 49.4% Positive and 50.6% Negative.

Figure 5. WordCloud Keseluruhan Data

In wordcloud, the more the word is used, the larger the font on the word will be. Figure 15 shows the words 'vaksin, 'vaksinasi covid', 'vaksin vovid', 'sinovac' as words that appear frequently.

4. CONCLUSION

Based on the test results using the K-fold Cross Validation method, the performance of the Long Short Term Memory algorithm produces a fairly good average accuracy of 74%, with the highest accuracy value in the 4th fold of 81%, the factors that affect this accuracy are the amount of data used, data quality, and data confusion. Based on the results of the 4th fold which has the highest accuracy value, the results of sentiment analysis using the Long Short Term Memory algorithm on public opinion regarding the COVID-19 vaccine on Twitter social media get 49.4% positive and 50.6% negative results.

REFERENCES

[1] A. T. Budiman, Y. Ruldeviyani, and A. N. Hidayanto, "Predictive Analytics Comparison of Indonesia Government Budget With or Without Covid-19 Pandemic," *SIST 2022 - 2022 International Conference on Smart Information Systems and Technologies, Proceedings*, 2022, doi: 10.1109/SIST54437.2022.9945772.

[2] R. R. Kartika Winahyu, M. Somantri, and O. D. Nurhayati, "Predicting Creditworthiness of Smartphone Users in Indonesia during the COVID-19 pandemic using Machine Learning," *2021 International Seminar on Machine Learning, Optimization, and Data Science, ISMODE 2021*, pp. 223–227, 2022, doi: 10.1109/ISMODE53584.2022.9742831.

[3] D. F. Murad, R. Hassan, Y. Heryadi, B. D. Wijanarko, and Titan, "The Impact of the COVID-19 Pandemic in Indonesia (Face to face versus Online Learning)," *Proceeding - 2020 3rd International Conference on Vocational Education and Electrical Engineering: Strengthening the framework of Society 5.0 through Innovations in Education, Electrical, Engineering and Informatics Engineering, ICVEE 2020*, Oct. 2020, doi: 10.1109/ICVEE50212.2020.9243202.

[4] S. F. Pane, Heriyanto, A. G. Putrada, N. Alamsyah, and M. N. Fauzan, "The Influence of The COVID-19 Pandemics in Indonesia On Predicting Economic Sectors," *2022 7th International Conference on Informatics and Computing, ICIC 2022*, 2022, doi: 10.1109/ICIC56845.2022.10006897.

[5] "Digital in Indonesia: All the Statistics You Need in 2021 — DataReportal – Global Digital Insights." https://datareportal.com/reports/digital-2021-indonesia (accessed Jun. 22, 2023).

[6] P. Gupta and M. Gupta, "Managing Congregations of People by Predicting Likelihood of a Person being Infected by a Contagious Disease like the COVID Virus," *Proceedings - 2020 IEEE International Conference on Cloud Computing in Emerging Markets, CCEM 2020*, pp. 32–36, Nov. 2020, doi: 10.1109/CCEM50674.2020.00017.

[7] G. Dobrita, A. Bara, S. V. Oprea, C. Baroiu, and D. C. Barbu, "Mobility, COVID-19 cases and virus reproduction rate data analysis for Romania using Machine Learning Algorithms," *2022 26th International Conference on System Theory, Control and Computing, ICSTCC 2022 - Proceedings*, pp. 244–251, 2022, doi: 10.1109/ICSTCC55426.2022.9931806.

[8] S. Zhang, M. J. Ventura, and H. Yang, "Network Modeling and Analysis of COVID-19 Testing Strategies," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 2003–2006, 2021, doi: 10.1109/EMBC46164.2021.9629754.

[9] S. Zhang, S. Yang, and H. Yang, "Statistical Analysis of Spatial Network Characteristics in Relation to COVID-19

Transmission Risks in US Counties," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 2278–2281, 2021, doi: 10.1109/EMBC46164.2021.9629892.

[10] "Ribka Tjiptaning, Orang Pertama Menolak Vaksin di Indonesia." https://www.cnnindonesia.com/nasional/20210113074635 -32-592938/ribka-tjiptaning-orang-pertama-menolak- vaksin-di-indonesia (accessed Jun. 22, 2023).

[11] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón- Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Syst Appl*, vol. 223, p. 119862, Aug. 2023, doi: 10.1016/J.ESWA.2023.119862.

[12] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Natural Language Processing Journal*, vol. 2, p. 100003, Mar. 2023, doi: 10.1016/J.NLP.2022.100003.

[13] H. Wang and M. Hou, "Quantum-like implicit sentiment analysis with sememes knowledge," *Expert Syst Appl*, p. 120720, Jun. 2023, doi: 10.1016/J.ESWA.2023.120720.

[14] F. Iacono, L. Magni, and C. Toffanin, "Personalized LSTM-based alarm systems for hypoglycemia and hyperglycemia prevention," *Biomed Signal Process Control*, vol. 86, p. 105167, Sep. 2023, doi: 10.1016/J.BSPC.2023.105167.

[15] J. Li, L. Lu, C. Liu, and Y. Gong, "Improving Layer Trajectory LSTM with Future Context Frames," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May, pp. 6550–6554, May 2019, doi: 10.1109/ICASSP.2019.8683783.

[16] D. Haputhanthri and A. Wijayasiri, "Short-term traffic forecasting using LSTM-based deep learning models," *MERCon 2021 - 7th International Multidisciplinary Moratuwa Engineering Research Conference, Proceedings*, pp. 602–607, Jul. 2021, doi: 10.1109/MERCON52712.2021.9525670.

[17] S. D. Kumar and D. P. Subha, "Prediction of depression from EEG signal using long short term memory(LSTM)," *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, vol. 2019-April, pp. 1248–1253, Apr. 2019, doi: 10.1109/ICOEI.2019.8862560.

[18] S. Chakraborty, J. Banik, S. Addhya, and D. Chatterjee, "Study of Dependency on number of LSTM units for Character based Text Generation models," *2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020*, Mar. 2020, doi: 10.1109/ICCSEA49143.2020.9132839.

[19] S. R. Patra, H.-J. Chu, and Tatas, "Regional groundwater sequential forecasting using global and local LSTM models," *J Hydrol Reg Stud*, vol. 47, p. 101442, Jun. 2023, doi: 10.1016/J.EJRH.2023.101442.

[20] K. Duan, R. Wang, S. Chen, and L. Ge, "Exploring the predictability of attention mechanism with LSTM: Evidence from EU carbon futures prices," *Res Int Bus Finance*, p. 102020, Jun. 2023, doi: 10.1016/J.RIBAF.2023.102020.