Sentiment Analysis of Covid-19 Tweets by using LSTM Learning Model

Yunus Emre Karaca[*1] iD , Serpil Aslan[2] iD
[1]Department of Informatics, Malatya Turgut Ozal University, Malatya, Turkey
[2] Department of Software Engineering, Malatya Turgut Ozal University, Malatya, Turkey
(yunusemre.karaca@inonu.edu.tr , serpil.aslan@ozal.edu.tr)

Abstract - Social media plays an important role in our lives due to the conditions of the age we live. Nowadays, the most popular social media platform that prioritizes meaningful content sharing is Twitter. In Twitter, which produces big data on an unprecedented scale, users have the opportunity to share their own perspectives, feelings, and experiences, as well as examine the opinions of other individuals. The Coronavirus-2019 (Covid-19) disease, transmitted through close contact and small droplets spread by people coughing, sneezing, or speaking, has created social and economic wounds worldwide. As of July 7, 2021, more than 185 million people worldwide have been diagnosed with the New Coronavirus (Covid-19), and approximately 4 million people have died from this infectious disease. This work focuses on the analysis of the sentiments that Covid-19 leaves on people, using the tweets that people share about the Covid-19 pandemic on the Twitter platform. Analyzes are based on deep learning algorithms. Sentiment analysis can provide serious benefits. In this study, we used a Long-short Term Memory (LSTM) based network model. Also, we compared the proposed model other machine learning algorithms: Support Vector Machine (SVM), Naïve Bayes and Logistic Regression. Experimental results show that our proposed method can effectively perform sentiment analysis on the Twitter dataset.

1. INTRODUCTION

The coronavirus pandemic, which is an epidemic disease, has been the focus of attention not only in the field of medicine but also in many different disciplines. The increasing technological activities and the instant use of many social media tools have created an exciting workspace in processing and finalizing different data as a concept within the IT sector. Until 20 years ago, when smartphones and pocket computers did not become more widespread, surveys were held together with field studies to keep the pulse of the society in the face of mass or political developments.

In line with the survey studies conducted later, country managers and institutions would set a roadmap against the wishes of the society, but today these surveys have been replaced by social media reviews. Thanks to the technological developments in the information age we live, it has become very easy to access the Internet with electronic devices such as smartphones and pocket computers. Also, individuals spend most of their daily activities with these devices. The general internet usages in all of the world are shown in Fig.1.

Social media networks such as Twitter, Facebook, Instagram, LinkedIn, where people express themselves, create and share their resumes, have emerged. Now, individuals express themselves, explain themselves, and be instantly informed about developments in these channels without the need for

interviewers or fieldwork. On the other hand, institutions use social media to promote products and services and communicate with customers in this increasingly popular medium [2].

Twitter is one of the most popular social media tools where the number of daily users is increasing rapidly. Twitter is a social networking service where users write and share messages called 'tweets' and interact with each other, referred to as 'the SMS of the internet, with 340 million active users as of 2020 and 98 thousand tweets per minute. [3]

It is not possible to access such serious data in a short time through a field study or surveyors. Twitter's Big data, where millions of users share their opinions, is very difficult to process. The processing of these data attracted the attention of many researchers and enabled them to produce methods. One of the most widely used techniques in the analysis of Twitter data is Sentiment Analysis Methods. Sentiment analysis is the automatic extraction of subjective information conveyed through a text. Through sentiment analysis, we can get an idea of whether a text has a positive or negative subjective orientation. With the concept of sentiment analysis, the processing of these data has become easier and reveals what society tells us, what it plans or wants.

This work focuses on the analysis of the sentiments that Covid-19 leaves on people, using the tweets shared by people about the Covid-19 pandemic on the Twitter platform.

In the second chapter, academic studies related to the subject are mentioned. In chapter three, the concept of sentiment analysis and the LTSM network are mentioned. In the fourth chapter, the proposed method is explained. In the fifth chapter, the experimental results are obtained. In the sixth chapter, a general conclusion is drawn to the work.
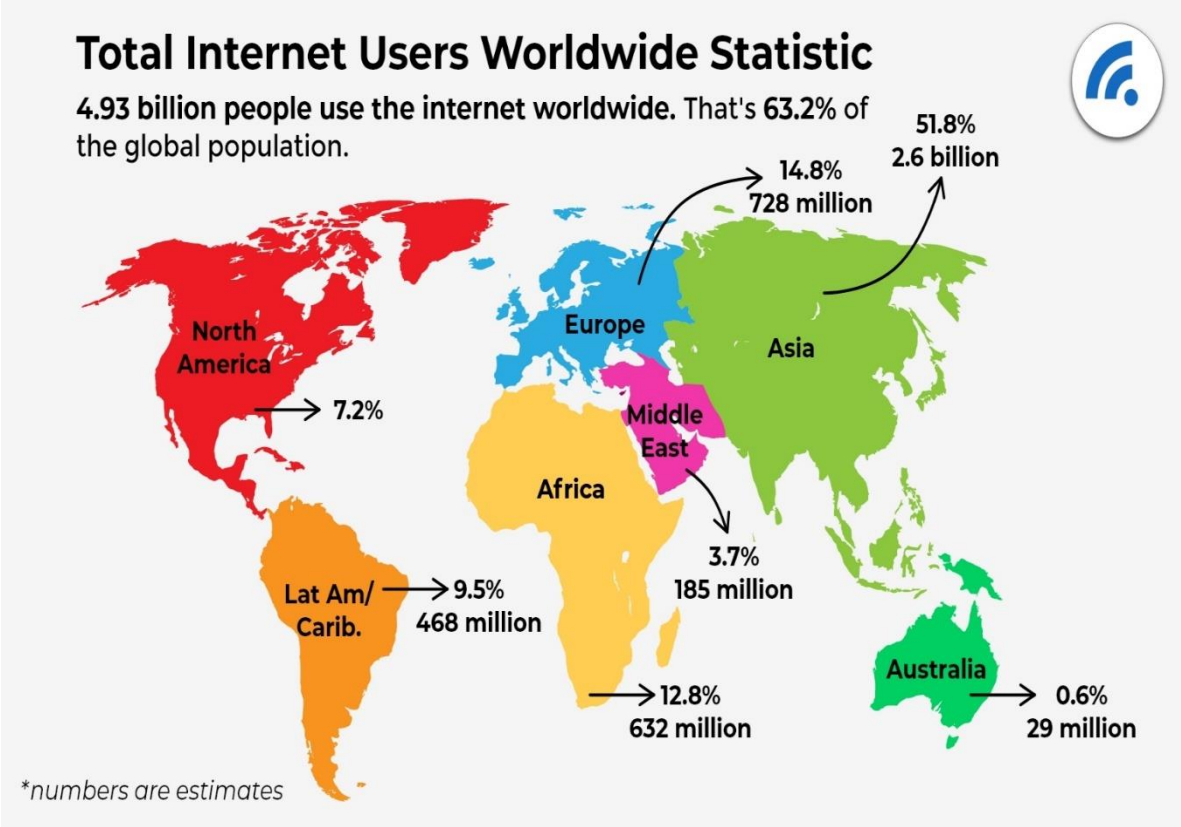

Figure 1. General Internet Usage Statistics 2021 [1]

2. BACKGROUND AND RELATED WORKS

The rapid increase in information sharing or access to information in social media networks gives rise to the idea of transforming information in this field into informatics. In this context, the concept of opinion mining, in other words, sentiment analysis, is on its way to becoming the shining star of the information sector, which grows in parallel with the developing technology. If we examine the reason a little bit, the internet is a very skillful place in terms of emotional information [4]. While emotion is an attitude, thought is a judgment driven by emotion. Sentiment analysis, also known as opinion mining, examines people's feelings towards certain entities and aims to determine people's feelings about a topic by analyzing their posts and different actions on social media. [5]. So, the user can post their point of view, their content via various social media such as forums, microblogs, or online social networking sites [4]. Also, people rely more on posts shared on social networking sites like Twitter, Facebook, and Instagram [6].

While technology was advancing rapidly, humankind experienced an event that could turn everything upside down at a time when it was just a click away from everything they wanted via the phone in their pocket. While the name of this event was initially defined as a virus-related disease in the form of the flu, as of March 11, 2020, the World Health Organization named the coronavirus a pandemic. As of July 7, 2021, the coronavirus, which shook the whole world with its effects, has been diagnosed with more than 185 million people all over the world. About 4 million people died from this contagious disease. The coronavirus has created a process that causes social traumas that deeply affect the work and family life, habits, holidays, in short, all their lives.

In this process full of differences, extensive studies have been carried out using both coronavirus and artificial intelligence technologies. During these studies, studies focusing on sentiment analysis of COVID-19 using social media data stand out **[6-10]**. That's because individuals, companies, and governments are using social media to communicate with each other on a range of issues related to the COVID-19 pandemic.

As not much is known about the issues shared on social media platforms related to COVID-19, analyzing such information can help politicians and healthcare organizations assess their stakeholders' needs and address them appropriately [7]. Analyzing public mood is an essential prerequisite for appropriate crisis management during extraordinary events such as pandemics [8]. In a study by Jim Samuel and colleagues [9], he outlines a methodological approach to analyzing Twitter data to identify sentiment, keywords, connotations, and trends specific to this crisis to determine public sentiment, tracking the progression of fear associated with the rapid spread of COVID-19 infections. Another study analyzes how people are responding to this pandemic, how aware they are of the disease and its symptoms, what precautions they take, and whether people follow the guidelines of political rulers [10]. Although we have talked about the positive side of the shares on social media until now, if the shares in this formation, which we call social media, are malicious, that is, if false information is shared, while there is already a crisis environment with the pandemic, sharing inaccurate information and fearing may drag the societies into this psychological depression more.

That's why all governments should use social media fact-checkers to prevent further sharing of unnecessary information for cases of such serious concern [6].

Figure 2. Word cloud highlighting some hashtags related to the topic.

3. SENTIMENT ANALYSIS

While developing technology and increasing internet use cause an increase in the information shared in this field, questions such as what to do and how to process this information have created a field for researchers in this field. To give an example of the areas created by this increase in data, research areas such as clustering, automatic text summarization, data classification, sentiment analysis can be given. When we look at it in this context, when we browse different web pages or through social media posts, we can easily see people's comments, what they like and what they don't like, what they want, their complaints, their health, and economic conditions and many other criteria. What we saw formed the content of Idea Mining, or in other words, Sentiment Analysis [11]. Sentiment analysis is the automatic extraction of subjective information conveyed through a text. Through sentiment analysis, we can get an idea of whether a text has a positive or negative subjective orientation.

The concept of sentiment analysis in social media posts is one of the useful tools [12]. However, with the significant growth of user-generated content on the Web, especially in recent years, it has been transformed into a world in itself. This environment has even changed the way people express their views and opinions. Now, individuals can post their opinions about a product they bought on their site, and they can express their opinions about almost anything on internet forums, discussion groups, and blogs, social media.

If a person wants to buy a product, they will not only be able to ask their friends or family members, but they will be able to read opinions made by current buyers of the product they are going to buy on the Web [12]. Companies need to capture the positive and negative feedback of customer service at the right time.

Implementing an effective sentiment analysis tool can increase customer satisfaction and reduce costs. As mentioned above, the amount of accessible information on the Web is increasing with the contribution of forums, columns, blogs, and social media. [13].

Companies may no longer have to conduct surveys, organize focus groups, or use external consultants to gain consumer insights on their products and competitors' products because user-generated content on the Web can give them all the information they want.

However, finding the basis for opinions and tracing them on the Web is a challenging area, as there are so many different sources and each source can also have a huge amount of opinionated text.

Ideas can often be found in long forum posts and blogs. For this reason, it is difficult for the reader to find relevant sources, extract sentences about ideas, read, summarize and organize them into usable formats [12]. Therefore, automatic opinion finding and summarization systems are needed. While

processing this information, extracting subjectivity, and classifying emotion, creates the basic outlines of emotion analysis, it also contains difficulties due to its structure.

3.1. LSTM Network

Long short-term memory (LSTM) is a type of repetitive neural network (RNN) supervised deep learning algorithm. The idea behind RNN is based on remembering what information was in previous neurons. So these neurons can relay information to themselves in the future for further analysis. LSTMs usually contain three gates. The "gateway" controls the entry of new information into memory. The "forget gate" controls how long certain values are kept in memory. Finally, the "output gate" controls how much the value stored in memory affects the output activation of the block [14]. LSTM architecture is as shown in Figure-3.
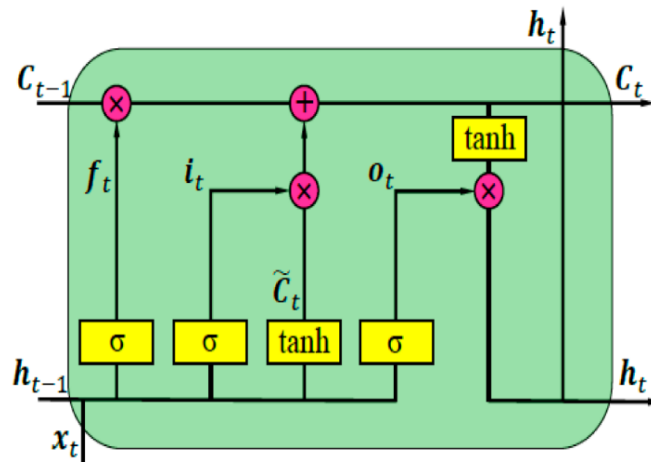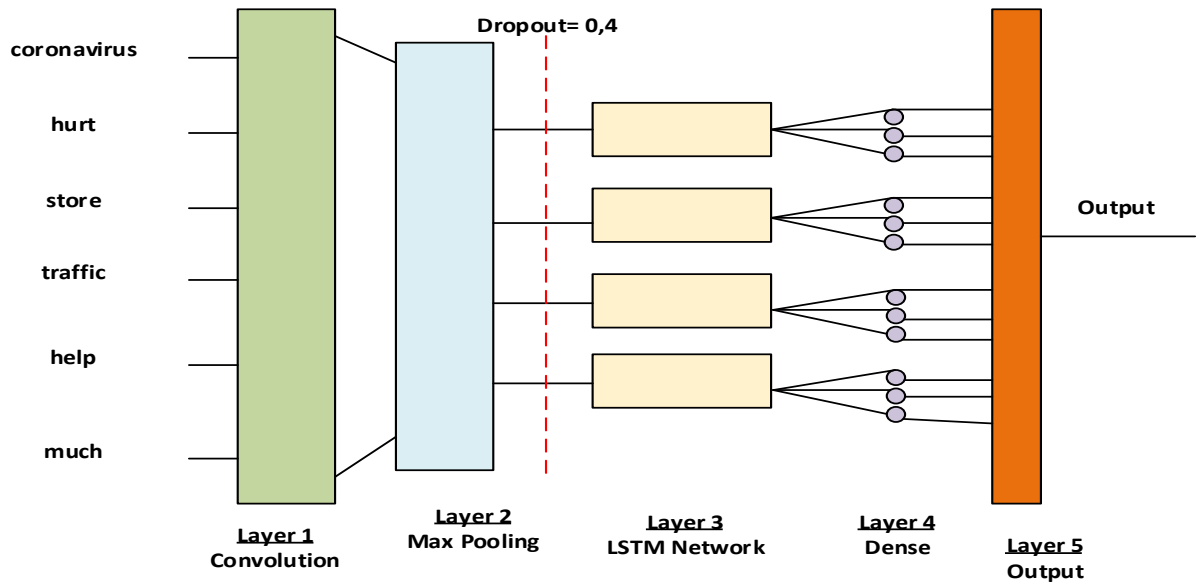

Figure 3. LSTM architecture [15]

4. METHOD

The open-source dataset from Kaggle was used to test the proposed method. The dataset contains Corona_NLP_test.csv and Corona_NLP_train.csv files. We used Corona_NLP_train.csv file to train the Natural Language Process (NLP) model. We used Corona_NLP_test.csv to evaluate the proposed model. The training dataset consists of 41,157 tweet record and the test dataset consists of 3,798 tweet records. The used dataset is a raw dataset. So, it must be cleaned. For these reasons, we performed a preprocessing on the dataset. The preprocessing steps shown as Fig. 1. Firstly, we dropped Retweet from dataset. Then, all URLs, punctuations and stop words removed because they do not contain any sentimental information. After then, we tokenized and stemmed the rested tweets with NLTK which is a python library. Finally, we applied lemmatization on the preprocessed tweet records.

The next step of the preprocessing is extracted sentiments from tweets. We used TextBlob to label the tweets as positive, neutral or negative. TextBlob is a Python library used in many natural language processing (NLP) tasks such as part of speech tagging, sentiment analysis, stemming, translation, classification, and many more [16,17]. The TextBlob library returns two properties, "Polarity" and "Subjectivity" [18]. For this purpose, subjectivity and polarity scores were calculated using the NLTK and TextBlob libraries. These scores are used for labeled the tweets sentiment.

Figure 4. The Preprocessing Steps

The next step of labeling tweets' sentiments, we built our model and trained it. For this purpose, we used LSTM network to building the model. The details of LSTM network model is explained Section III. The architecture of the proposed model as shown Fig. 6.

Figure 5: WordCloud of Preprocessed Dataset

Figure 6. The architecture of LSTM Network Model

5. EXPERIMENTAL RESULTS

We used three machine learnings algorithm to evaluate the success of LSTM network model. The compared algorithms are SVM (Support Vector Machine), Logistic Regression and Naive Bayes Model. We tested the experiments on Google Collaborate with Python programming language. All experiments were performed on an Intel Core i7 computer with 16 GB RAM.

To evaluate our experimental results, we use F1 score and Precision evaluation metric to compare the different deep learning algorithms. Precision and F1 Score are calculated as sequentially Equation 1 and Equation 2.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \qquad (1)$$

$$F1\ Score = 2x\frac{Precision\ x\ Recall}{Precision\ x\ Recall} \qquad (2)$$

$$Recall = \frac{TruePositivies}{TruePositives + FalseNegatives} \qquad (3)$$

Table 1. Precision results of deep learning algorithms

DL Algorithm	Precision Score
LR (Logistic Regression)	0,5472
Naïve Bayes	0,6829
SVM (Support Vector Machine)	0,5314
LSTM Model	**0,8865**

As shown Table 1, compared with other baseline algorithms, the proposed LSTM network model is the best model all of them. Because, the LSTM is recurrent training network model. Also, as shown Table 2, it outperforms all other models for F1 score evaluation metric. It can be observed that the LSTM model achieved the best prediction results.

Table 2. F1 results of deep learning algorithms

DL Algorithm	Precision Score
LR (Logistic Regression)	0,4472
Naïve Bayes	0,5201
SVM (Support Vector Machine)	0,6003
LSTM Model	**0,7550**

6. CONCLUSION

Social media plays an important role in our lives due to the conditions of the age we live in. Twitter which has millions of users is the most popular social media platform. It offers huge data source for data mining to the researchers. On Twitter, users share tweets expressing their feelings with 140 characters reserved for them.

In this study, we aim to analyze sentiments of tweets about Covid 19 pandemics. The proposed model was established using an LSTM network. We used three machine learnings algorithm to compare the success of LSTM network model. LSTM is a special RNN model. LSTM has the ability to capture sequential data taking into account previous data. It also has feedback links unlike standard RNNs.

The compared algorithms are SVM (Support Vector Machine), Logistic Regression and Naive Bayes Model. Experimental results show that our proposed method can effectively perform sentiment analysis on the Twitter dataset.

REFERENCES

[1] Internet Users Worldwide Statistic, Available at: https://www. broadbandsearch.net/blog/internet-statistics, Anonymous, retrieved 28th July, 2021.

[2] He, W., Wu, H., Yan, G., Akula, V., & Shen, J. *"A novel social media competitive analytics framework with sentiment."* Elsevier, 1-12, 2015.

[3] *Twitter*. (2021, 07 13). wikipedia:https://tr.wikipedia.org/wiki/Twitter

[4] Fang, X., & Justin, Z., *"Sentiment analysis using product review data."* Journal of Big Data, 1-14, 2015.

[5] Christi, J., & Jain, G., *"Sentiment Categorization through Natural Language Processing :A Survey.",* 104-107, (2019, 11 15).

[6] Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E., *"Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media."* Elsevier, 2020.

[7] Alrazaq, A. a., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z., *"Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study."* JOURNAL OF MEDICAL INTERNET RESEARCH, 1-10, 2020.

[8] Gencoglu, O., *"Large-Scale, Language-Agnostic Discourse Classification of Tweets During COVID-19."* Machine Learning and Knowledge Extraction, 603–616, 2020.

[9] Samuel, J., Ali, G. G., Rahman, M., Esawi, E., & Samuel, Y., *"Covid-19 public sentiment insights and machine learning for tweets classification."* Information, 11(6), 314, 2020.

[10] Iyer, P., & Kumaresh, S., *"Twitter Sentiment Analysis On Coronavirus Outbreak Using Machine Learning Algorithms."* European Journal of Molecular & Clinical Medicine, 2663-2676, 2020.

[11] BİLEN, B., & HORASAN, F., *"LSTM network based sentiment analysis for customer reviews."* JOURNAL of POLYTECHNIC, 2021.

[12] Türkmenoğlu, C., & Tantuğ, A. C., *"Sentiment analysis in Turkish media."* Workshop on Issues of Sentiment Discovery and, (1-11), 2014.

[13] Liu, B., *"Sentiment analysis and subjectivity."* Handbook of natural language processing, 2, 627-666, 2010.

[14] Xiao, Y., & Yin, Y., *"Hybrid LSTM neural network for short-term traffic flow prediction."* Information, 10(3), 105, 2019.

[15] Loria S. textblob Documentation. Release 015. 2018; 2.

[16] Sohangir S, Petty N, Wang D., *"Financial sentiment lexicon analysis."* In: 2018 IEEE 12th International Conference on Semantic Computing (ICSC), p. 286–289, 2018.

[17] Ankit and Saleena, N*., "An ensemble classification system for twitter sentiment analysis."* Procedia Computer Science, 132(2):937–946, 2018.