Classifying Various Types Of Symptoms Of COVID-19 (CTSC) In Twitter

(Text Mining)

Mahdieh Vahedipoor
Computer Engineering Student
Qom University Of Technology
Vahedipoor.M@Qut.ac.ir

Mahbobeh Shamsi
Computer Engineering Assistant
Qom University Of Technology
Shamsi@Qut.ac.ir

Saba Farhadi
Computer Engineering Student
Qom University Of Technology
Farhadi.s@Qut.ac.ir

Reza Rasouli
Computer Engineering Assistant
Qom University Of Technology
Rasouli@qut.ac.ir

***Abstract*— Data mining has many usages in the field of health, including the diagnosis of diseases, classification of patients in disease management, finding patterns for faster diagnosis of patients, and preventing complications. Research in the field of extracting public health data in social networks such as Twitter has grown exponentially. Many researchers have decided to use machine learning and deep learning algorithms for such analyzes. In this study, we present a method for classifying the types of symptoms of COVID-19 disease (CTSC) using deep learning algorithms and then analyze English Twitter data related to people who tested positive for COVID-19 for 8 days from 2021/06/26 to 2021/07/04.**

This study includes pre-processing of tweets and classification of the different symptoms of COVID-19, including Respiratory, Digestive, Muscular, Smell-Taste, and Sinusitis. In the proposed framework, Machine learning algorithms such as LR, DT, SGD, SVM, RF and deep learning algorithms such as CNN, LSTM, and GRU evaluate sentiment analysis. The results show that users diagnosed with covid19 show respiratory symptoms, including sneezing, lung problems, sore throat, ulcers, cough, fever, shortness of breath, and heart problems 18% more likely than others. We also obtained the best performance for evaluating the CTSC method using machine learning algorithms with accuracy of 87% and deep learning algorithms with accuracy of 96% .

Keywords*— *COVID-19, Respiratory, Twitter, Deep Learning, disease.

I. Introduction

The family of coronaviruses in humans, mammals, and birds has been identified genotypically and serologically with four types of alpha, beta, gamma, and delta. The types alpha and beta can cause disease in humans. A new strain of the virus, called COVID-19, first spread from Wuhan, China, in 2019, then spread to dozens of countries worldwide, causing a worldwide epidemic. Severe Acute Respiratory Syndrome or SARS and Middle East Respiratory Syndrome or MERS belong to the beta group, which has a high prevalence and has caused many risks [1]. It is not yet known how the COVID-19 virus is transmitted from animals to humans, but it has been detected in the seafood market [2]. Chen et al. [3] confirmed that the COVID-19 virus is highly associated with the bat SARS coronavirus. The prevalence of this disease is much higher than SARS in 2002 [4]. Wrap et al. [5] found that the ability of the COVID-19 to bind was 19 to 20 times more tightly than that of the SARS coronavirus. Early symptoms of COVID-19 include simple signs of the common cold, but as the disease progresses, other symptoms such as indigestion, respiratory distress, fatigue, fever, and dry cough appear [6]. On January 30, 2020 world Health Organization announced that the epidemic of COVID-19 is a public health emergency. With the spread of the COVID-19 virus, analysis of epidemiological data is essential to prepare the community for better decisions against the disease [7].

In recent years, due to the broad access to vast amounts of data available on social networks such as Twitter and the imminent need to turn data into useful information and knowledge, data mining methods have received much attention in various sciences, including medical science. Data mining means "discovering specific methods and patterns in large databases to guide decisions about future activities."
In the early stages of the epidemic, patients, nurses, and physicians used social media to exchange information about the symptoms and treatment of COVID-19[8]. During the outbreak of COVID-19 and the asylum of many people at home, the use of social networks increased widely (more than 20%) [9].

Health professionals and researchers have used traditional medical databases for years to study and understand public health. Recently, social media data, especially Twitter, has been used extensively for public health purposes. Every technological development in history affects the behavior of society. The advent of the Internet and social media is no exception. Social networks create a general flow of communication, and scientists have discovered that such information can provide a level of access to people's views and circumstances [10].

When the conventional capacity to monitor diseases is limited, public social networks and Internet data can play an essential role in uncovering the hidden dynamics of an outbreak [11]. Social networks can enable the early detection of symptoms of diseases such as COVID-19, whose treatment is not known for sure, and our knowledge about it is expanding [12].

The most prominent symptoms of the disease, such as fever, cough, and shortness of breath, were identified early in the corona outbreak. But other symptoms such as loss of smell and taste, body aches, and diarrhea were later added to the list of COVID-19 symptoms [13].

In the CTSC method, the meaningless information is eliminated by collecting tweets from June 26, 2021, to July 4, 2021, concerning the symptoms of COVID-19. The root of the words is found, and then they are classified based on the dictionary of five categories: respiratory, digestive, muscular, smell-taste, and sinusitis. Various machine learning algorithms such as Logistic Regression (LR), Decision Tree (DT), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM) and Random Forest (RF) have been used to classify these tweets. Then, to evaluate the proposed method, we use accuracy, precision, recall and F-Measure.

We also use deep learning algorithms such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Gated Recurrent Unit (GRU) and consider Accuracy and Percentage of Error to evaluate the results.

The structure of the paper is as follows: Section 2 provides a review of the literature. Section 3 shows the Dataset of terms in text mining and classification algorithms and their details. Section 4 describes the implementation and evaluates of the proposed CTSC method. Section 5 concludes the proposed method and presents future work.

II. LITERATURE REVIEW

Using social media to gather information about public health is a growing field of research [14], and often labeled data is used to build supervised machine learning models [15]. In 2020, Tim Mackie et al. conducted research on symptoms detection, testing, and recovery associated with COVID-19 using Twitter social media. The purpose of their study is to identify and describe user-generated conversations related to the symptoms, testing, and recovery associated with COVID-19. Unsupervised machine learning algorithms are used to implement this system [16].

Antonioa et al conducted research on sentiment analysis using TF-IDF feature extraction and stochastic gradient descent algorithm. Their work can classify Indonesian text appropriately according to positive and negative sentiment. Classification Performance using TF-IDF feature extraction and stochastic gradient descent algorithm obtained an accuracy of 85.141%[17].

Okango et al used coronavirus-related tweets on Twitter, using a dictionary-based method to polarize sentiments. They also examine the co-occurrence of words to gain insights on the aspects that affect the masses. The results show that mental health issues and lack of supplies have been the direct effects of this epidemic[18].

Roshan Santosh et al. conducted research on the detection of emerging symptoms of COVID-19 using context-based twitter embedding. Their goal is to provide a repetitive graph-based method for diagnosing the symptoms of COVID-19. Their method can detect coronavirus symptoms before they are reported by the Centers for Disease Control [19].

Cristian R. Machuca et al. developed a machine-learning method for sentiment analysis of the coronavirus on Twitter. Identifying people's positive and negative sentiments about a subject is an essential challenge in machine learning. They used English tweets related to COVID-19 in 2020 and categorized them using a logistic regression algorithm with an accuracy of 78.5% [20].

Kamaran H. Manguri et al. conducted sentiment analysis of the prevalence of COVID-19 using data from Twitter. In their research, python TextBlob library was used, and after evaluating sentiments of the tweets, graphic diagrams were presented. Their study showed that people's feelings about COVID-19 are variable [21].

Boone et al. performed sentiment analysis and topic modeling on data related to COVID-19. In this research, the LDA algorithm and natural language processing approaches have been used to identify the most frequent tweets and clustering to identify templates based on keyword analysis [22].

Karisani et al. Used language tools and machine learning algorithms including simple Bayes, logistic regression, fasttext, bertbase, and Twitter data to identify COVID-19 positive test records in China[23].

In recent years, using deep learning algorithms for analyzing Twitter data has attracted the attention of many researchers. Deep learning is a subfield of machine learning based on algorithms that attempt to model high-level abstract concepts in data. This process is modeled using a deep graph with multiple processing layers consisting of several layers of linear and nonlinear transformations. In other words, it is based on learning to display knowledge and features in model layers [24].

III. DATASET

In this study, we have collected 16504 English tweets containing phrases that somehow imply COVID-19 positive cases for 8 days from June 26, 2021 to July 4, 2021 using Twitter API and python web crawlers. We have also created a dictionary of COVID-19 symptoms classified into five categories: respiratory, digestive, muscular, smell-taste, and sinusitis (Table 1).

After pre-processing and counting the symptoms in each category and appropriate labeling of each type of symptoms of COVID-19, the results show that respiratory type contains 3049 records (18% of tweets), muscle type contains 1809 values (11% of tweets), gastrointestinal and smell-taste types contain 1096 and 1121 records respectively (7% of tweets) and finally sinusitis type contain 594 records (4% of tweets) and the rest are ineffective and did not belong to any category.

IV. IMPLEMENTING AND EVALUATING THE PROPOSED CTSC METHOD

In this section, we first mention the details of the implementation of the article and then evaluate the accuracy of performance the proposed model.

A. Implementing The CTSC method

In this research, we collected tweets from June 26, 2021, to July 4, 2021, using the Twitter API and web crawlers by extracting the features related to the COVID-19 disease.
To classify the types of symptoms of COVID-19, we created a dictionary of English tweets for COVID-19 that has five distinct categories such as respiratory, gastrointestinal, muscular, olfactory-taste, and sinusitis, each of which includes several specific symptoms according to the table 1.
For example, the respiratory category includes symptoms such as sneezing, pneumonia, sore throat, cough, fever, shortness of breath, and heart disease. The gastrointestinal category has symptoms such as chills, muscle aches, diarrhea and vomiting, fatigue, nausea and stomach pain. Muscular category has symptoms such as chills, muscle aches, pain in the body, bone pain, fatigue and weakness. The Smell-Taste category has symptoms such as problems with the nose and sense of smell, problems with the sense of taste and problems with the five senses. Finally, Sinusitis includes symptoms such as headache, hypotension, impaired vision and runny nose.

Respiratory	**Digestive**	**Muscular**	**Smell-Taste**	**Sinusitis**
sneezing	chill	muscle	nose	headache
pneumonia	muscle	discomfort	smell	pressure
Throat sore	flux	pain	taste	vision
cough	puke	bone	sense	runny
fever	fatigue/tired	fatigue/tired		
breath/gasping	nausea	weak		
heart	stomach	chill		

Table 1 - Dictionary created by CTSC method for various symptoms of Covid disease 19

These tweets go through a process of eliminating meaningless words such as non-English letters and duplicated marks and letters. Then after normalizing and stemming, symptoms of each type are determined based on the symptoms dictionary. Then each tweet gets a category label of respiratory, digestive, muscular, smell-taste, or sinusitis based on the frequency of each category symptoms presented on that tweet.

Then various deep learning algorithms such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) are used to classify the tweets. After preparing the data, we used 70% of them for training and 30% for testing. The above algorithms are executed with three layers and batch size of 64 and epochs of 10. Then, to evaluate the proposed CTSC method, we have considered the accuracy and loss. Figure 1 shows the steps of the proposed method.

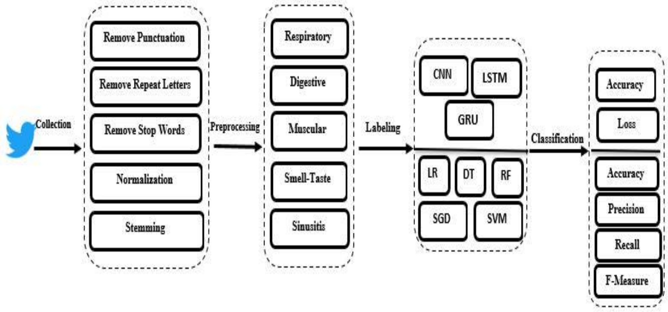

Fig. 1 - Review of the proposed CTSC Method

B. Evaluate The CTSC method

In this research, we have created a dictionary of various types of symptoms of COVID-19 disease, including Respiratory, Digestive, Muscular, Smell-Taste, and Sinusitis. Each class contains related symptoms of that type. According to this dictionary, we can count the symptoms declared in each tweet and identify the type of symptoms of the COVID-19 disease. Figure 2 shows the number of each type of symptoms of COVID-19 in 8 days. As a result, regardless of the neutral class, the respiratory type of COVID-19 with a value of 3049(18% of the total tweets) contains the most records, the Digestive type contains 1809 records (11% of the total tweets), and Muscular and Smell-Taste type contain 1121 and 1096 records Digestive (7% of the total tweets), and sinusitis contains 594 records (4% of the total tweets).

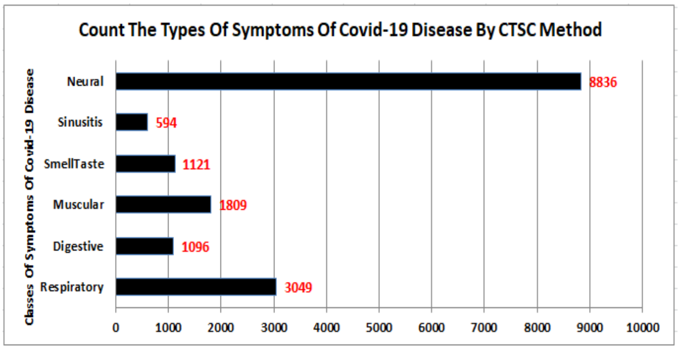

Fig. 2 – Counting Types of Symptoms of COVID-19 Disease by CTSC Method in Twitter

The accuracy obtained in the machine learning methods after applying the proposed CTSC method with 10Fold test is shown in Figure 3. Therefore, machine learning algorithms such as logistic regression (LR) with 83% accuracy and decision tree (DT) with 86% accuracy and random gradient reduction (SGD) and support vector machine (SVM) and random forest (RF) with 87% accuracy are results with the most accuracy. According to the results obtained from the machine learning algorithms, it can be said that the CTSC method had a good performance. However, this method is more efficient in deep learning algorithms (97% accuracy) than machine learning algorithms.

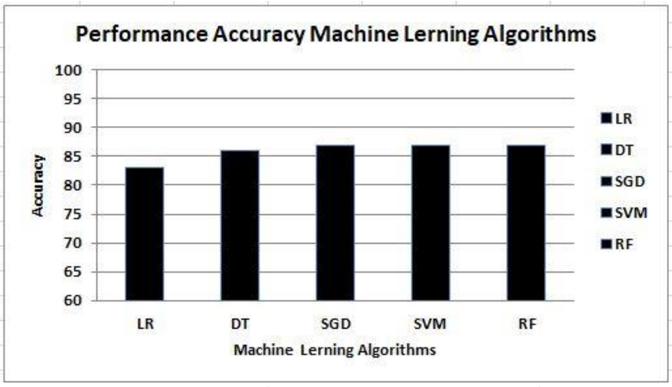

Fig. 3 - Accuracy Results of machine learning algorithms after applying CTSC method

The precision obtained in machine learning methods after applying the proposed CTSC method with 10Fold test is shown in Figure 4. Therefore, machine algorithms such as logistic regression (LR) with 72% precision and decision tree (DT) with 88% accuracy and random gradient reduction (SGD) with 79% precision and support vector machine (SVM) and random forest (RF) with 77% precision are results with the most precision. According to the precision results of the machine learning algorithms, it can be said that the CTSC method had a good performance in terms of correct diagnosis. The decision tree method with 88% had the highest and the logistic regression method with 72% had the lowest Precision.

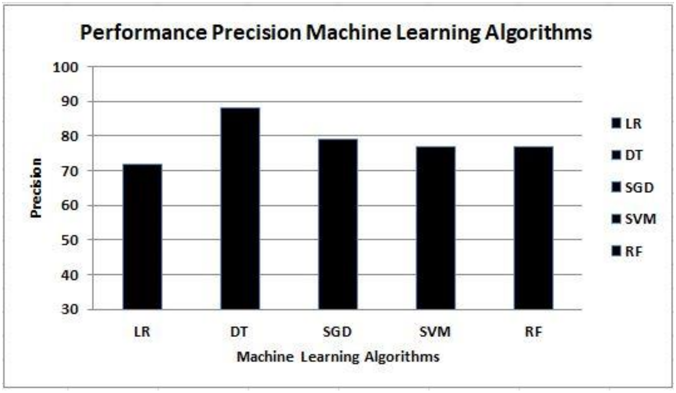

Fig. 4 - Precision Results of machine learning algorithms after applying CTSC method

The recall obtained in machine learning methods after applying the proposed CTSC method with 10Fold test is shown in Figure 5. Machine algorithm methods such as logistic regression (LR) had a 28% recall and decision tree (DT) showed 85% recall and random gradient reduction (SGD) had 79% recall and support vector machine (SVM) and random forest (RF) had a recall of 74%. According to the recall results obtained from the machine learning algorithms, it can be said that the CTSC method had a good performance in terms of correct recognition compared to the whole data and the decision tree method with 85% had the highest and the logistic regression method with 28% had the lowest recall score.

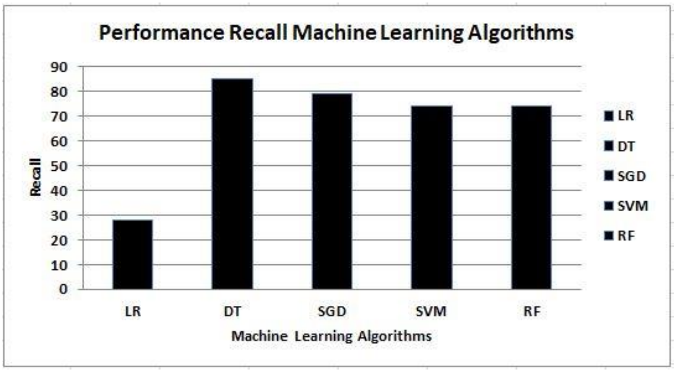

Fig. 5 - Recall Results of machine learning algorithms after applying CTSC method

The value of F-Measure obtained in machine learning methods after applying the proposed CTSC method with 10Fold test is shown in Figure 6. Therefore, machine learning algorithms such as logistic regression (LR) with F-Measure of 31%, decision tree (DT) with F-Measure of 84%, random gradient reduction (SGD) with F-Measure of 78%, support vector (SVM) and a random forest (RF) with F-Measure of 73% are results with the most F-Measure. According to the results of F-Measure obtained from machine learning algorithms, it can be said that the CTSC method had good performance in terms of balanced combination of precision and recall. The decision tree with 84% had the highest, and the logistic regression method with 31% had the lowest F-Measure.

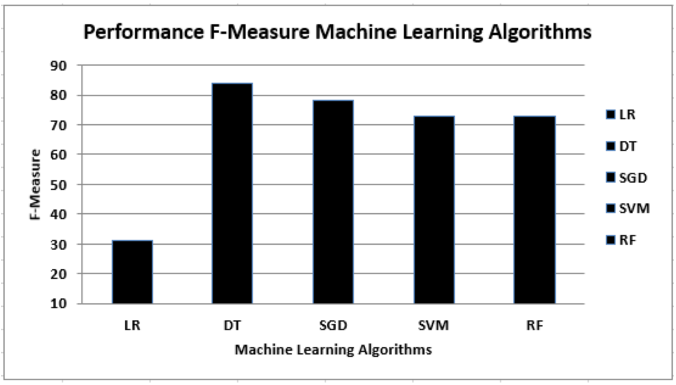

Fig. 6 – F-Measure Results of machine learning algorithms after applying CTSC method

The accuracy obtained from the convolutional neural network algorithm after applying the proposed CTSC method in 10 different rounds is shown in Table 2 and Figure 7.

The amount of accuracy obtained in training and testing in the third round has reached 97%. Training in the last two rounds has increased by one percent. Also, the Loss of convolution neural network algorithm after applying the proposed CTSC method in 10 different rounds is shown in Table 2 and Figure 8. The result obtained in the training data in the sixth round has reached 0.04%, but for the testing data in the fourth round, Loss has a fixed number of 0.05%. In the last two rounds, Loss reaches 0.06%, which is more than the training data. In general, the CNN method has the best performance on the data, and after applying the proposed CTSC method, we reach an accuracy of 97% and 0.06% Loss.

Table 2. Results of accuracy and Loss of training and testing of CNN after applying CTSC method

Round	Accuracy Train	Accuracy Test	Loss Train	Loss Test
Round 1	0.87	0.91	0.32	0.25
Round 2	0.94	0.96	0.17	0.1
Round 3	0.97	0.97	0.07	0.06
Round 4	0.97	0.97	0.05	0.05
Round 5	0.97	0.97	0.05	0.05
Round 6	0.97	0.97	0.04	0.05
Round 7	0.97	0.97	0.04	0.05
Round 8	0.97	0.97	0.04	0.05
Round 9	0.98	0.97	0.04	0.06
Round 10	0.98	0.97	0.04	0.06

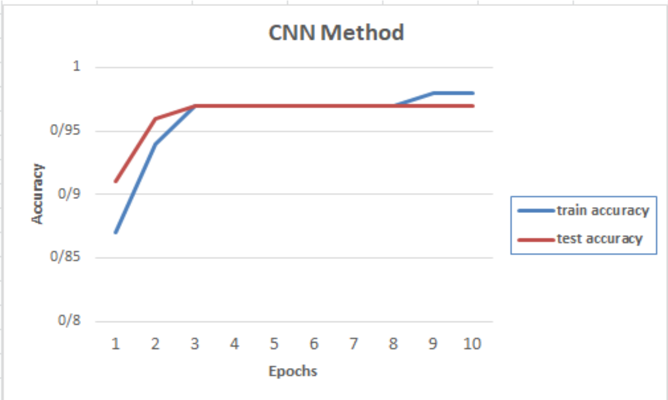

Fig. 7 - Accuracy of CNN after applying CTSC method

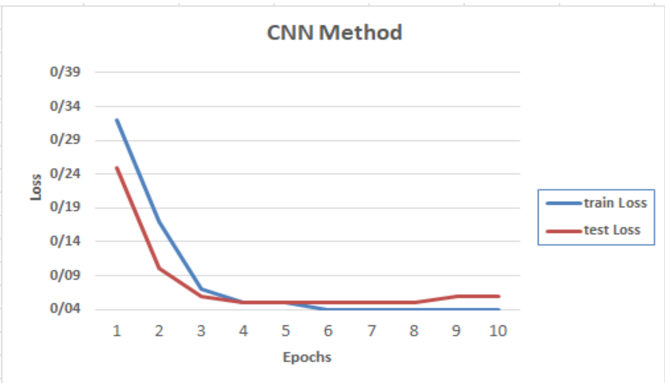

Fig. 8 - Loss of CNN after applying the CTSC method

The accuracy obtained from the Long Short-Term Memory algorithm after applying the proposed CTSC method in 10 different rounds is shown in Table 3 and Figure 9.

The accuracy obtained in training and testing in the third round has reached 97%, and in the last two rounds of training, we have a 1 percent increase in accuracy. Also, the Loss of the Long Short-Term Memory algorithm after applying the proposed CTSC method in 10 different rounds is shown in Table 3 and Figure 10. It can be analyzed that the percent of Loss in the training data in the eighth round has reached 0.04%, but for the testing data in the fifth round we have a Loss of 0.07%. In the last two rounds, we have a Loss of 0.06%, which is more than the training data. In general, the LSTM method has the best performance on the data and after applying the proposed CTSC method with 97% accuracy and 0.06% Loss.

Table 3. Accuracy and Loss rate in training and testing of LSTM method after applying CTSC method

Round	Accuracy Train	Accuracy Test	Loss Train	Loss Test
Round 1	0.89	0.92	0.28	0.17
Round 2	0.95	0.97	0.11	0.08
Round 3	0.97	0.97	0.07	0.07
Round 4	0.97	0.97	0.06	0.06
Round 5	0.97	0.97	0.06	0.07
Round 6	0.97	0.97	0.05	0.07
Round 7	0.97	0.96	0.05	0.07
Round 8	0.97	0.97	0.04	0.07
Round 9	0.98	0.97	0.04	0.06
Round 10	0.98	0.97	0.04	0.06

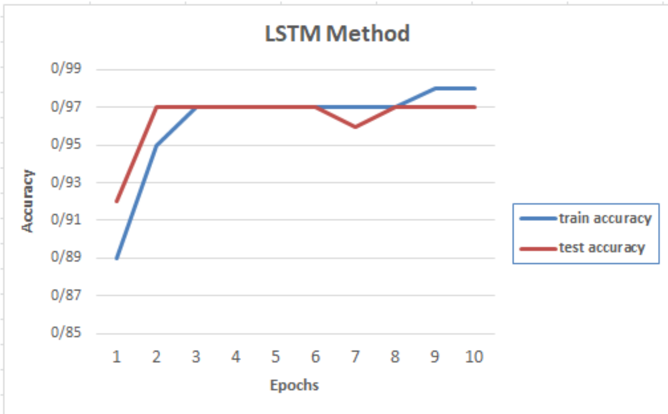

Fig. 9 - Accuracy of LSTM algorithm after applying the CTSC method

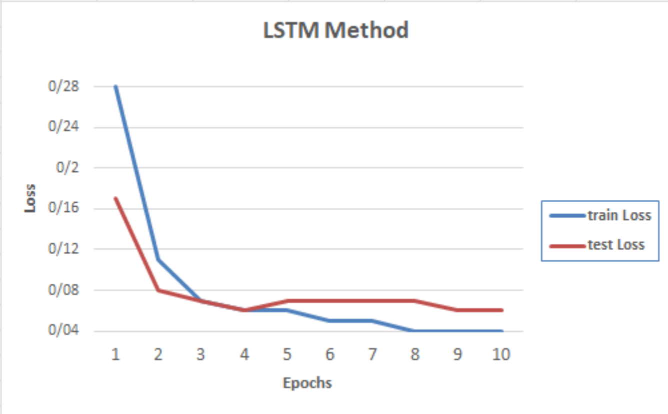

Fig. 10 - Loss rate of LSTM algorithm after applying the CTSC Method

The accuracy obtained from the Gated Recurrent Unit algorithm after applying the proposed CTSC method in 10 different rounds is shown in Table 4 and Figure 11.

The amount of accuracy obtained in training in the sixth round has a fixed number of 97%, and in the seventh round has a fixed number of 96%, and in the last round, it has reached 97%. Also, the Loss of the Gated Recurrent Unit algorithm after applying the proposed CTSC method in 10 different rounds is shown in Table 4 and Figure 12. It can be analyzed that the results obtained in the training and test data are different in 10 rounds. The Loss of rate training data has reached 0.05, and the Loss of testing has reached 0.06%, which is one percent more than the Loss training. In general, the GRU algorithm has the best performance on the data and after applying the proposed CTSC method with 97% accuracy and 0.06% Loss.

Table 4. Accuracy and Loss of training and testing of GRU algorithm after applying CTSC method

Round	Accuracy Train	Accuracy Test	Loss Train	Loss Test
Round 1	0.86	0.90	0.30	0.28
Round 2	0.91	0.92	0.24	0.21
Round 3	0.93	0.94	0.19	0.16
Round 4	0.95	0.95	0.13	0.12
Round 5	0.96	0.96	0.10	0.10
Round 6	0.97	0.97	0.08	0.08
Round 7	0.97	0.96	0.07	0.09
Round 8	0.97	0.96	0.07	0.08
Round 9	0.97	0.96	0.06	0.07
Round 10	0.97	0.97	0.05	0.06

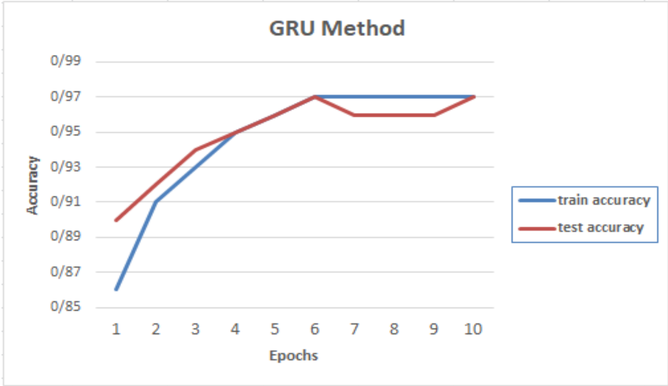

Fig. 11 - Accuracy of GRU algorithm after applying the CTSC method

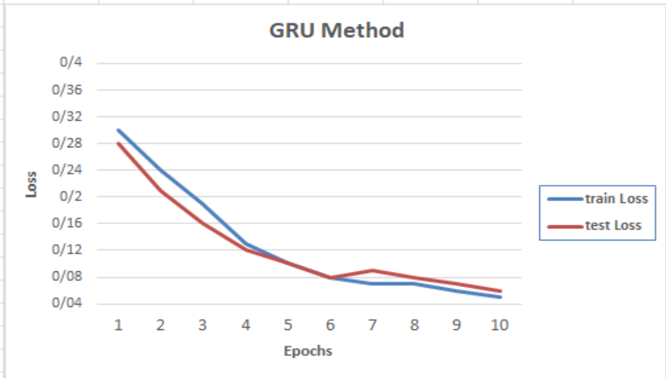

Fig. 12 - Loss rate of GRU algorithm after applying the CTSC Method

V. Conclusion

In this research, we provide a method based on different types of symptoms of COVID-19 using five machine learning algorithms: LR ،DT ،SGD ،SVM و RF and three deep learning algorithms: CNN, LSTM, and GRU.

In this study, the Twitter data has been pulled out from Twitter social media, and tweets are extracted based on positive COVID-19 hashtags. Tweets are then cleaned and matched against a symptoms lexicon and then labeled based on various types of symptoms of COVID-19 using mentioned algorithms. The evaluation results show that the presented method using SGD ،SVM و LR had an accuracy of 87%. Also, DT algorithm resulted in a precision of 88%, recall of 85% and F-Measure of 87%. Deep learning methods with an accuracy of 97 percent and a Loss of 0.06 had the best performance. In future work, we can apply this method to Persian tweets and include more symptoms in each category to obtain more accurate results.

References

[1] Y. Chen, Q. Liu, and D. Guo, “Emerging coronaviruses: genome structure, replication, and pathogenesis. Medical Virology,” 2020.

[2] L. Pub, "National Health Commission’s briefing on the pneumonia epidemic situation"., 2020.

[3] J.F. Chan, K.H. Kok, Z. Zhu, H. Chu, K.K To, S. Yuan and et al., “Genomic characterization of the 2019 novel human pathogenic. coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerging Microbes & Infections”, 2020.

[4] A.E. Gorbalenya, “Severe acute respiratory syndrome related coronavirus,” 2020.

[5] D. Wrapp, N. Wang, K.S. Corbett, J.A. Goldsmith, C.L. Hsieh, O. Abiona and et al., “Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation,” 2020.

[6] WHO coronaviruses (COVID-19). Retrieved March 30, 2020 from https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[7] N.S. Punn, K.S. Sanjay and A. Sonali,” COVID-19 epidemic analysis using machine learning and deep learning algorithms.”2020.

[8] G. Iacobucci, "COVID-19: diabetes clinicians set up social media account to help alleviate patients’ fears." 2020.

[9] Venkatraman., “Weekly time spent in apps grows 20% year over year as people hunker down at home”., 2020

[10] O. Edo-Osagie, I.B. De La, I. Lake and O. Edeghere, “A scoping review of the use of Twitter for public health research”., 2020.

[11] J. Zhang and D. Centola, “Social Networks and Health: New Developments in Diffusion, Online and Offline”., 2019.

[12] R. Del, Carlos, and N.M. Preeti, "COVID-19—new insights on a rapidly changing epidemic." , 2020.

[13] M.C. Grant and et al. "The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): a systematic review and meta-analysis of 148 studies from 9 countries." , 2020.

[14] K. O’Connor and et al. "Pharmacovigilance on twitter? Mining tweets for adverse drug reactions.", 2014.

[15] Y. Luo and et al. "Natural language processing for EHR-based pharmacovigilance: a structured review.", 2017.

[16] Mackey, Tim and et al. "Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: retrospective big data infoveillance study.", 2020.

[17] V.D. Antonioa, S. Efendia, H. Mawengkanga, “sentiment analysis for covid-19 in Indonesia on Twitter with TF-IDF featured extraction and stochastic gradient descent”,2022.

[18] E. Okango, H. Mwambi, “Dictionary Based Global Twitter Sentiment Analysis of Coronavirus (COVID-19) Effects and Response”, 2022.

[19] R. Santosh and et al. "Detecting emerging symptoms of COVID-19 using context-based twitter embeddings.", 2020.

[20] R. Machuca, G. Cristian, and M. Renato. "Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach.", 2021.

[21] K.H. Manguri, N. Rebaz, Ramadhan, and R. Pshko, A. Mohammed "Twitter sentiment analysis on worldwide COVID-19 outbreaks.", 2020.

[22] S. Boon-Itt, and S. Yukolpat, "Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study.", 2020.

[23] N. Karisani, and P. Karisani, "Mining coronavirus (COVID-19) posts in social media.", 2020.

[24] Y.Bengio, “Learning Deep Architectures for AI” and “Archived from the original”, 2014.