International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Design Text Mining Classifier for Covid-19 by using the Machine Learning Techniques

Suvarna Lakshmi C[1], Dr. Sameer Saxena[2], B. Suresh Kumar[3]

Submitted: 24/10/2022 **Revised:** 27/12/2022 **Accepted**: 28/01/2023

Abstract: In the starting of year 2020, WHO identified COVID-19 as a new pandemic and issued a statement to that effect. This fatal virus was able to disperse and propagate over several nations all over the globe. During the course of the epidemic, social media platforms like Twitter generated significant and substantial volumes of data that helped improve the quality of decisions pertaining to health care. As a result, we suggest that the opinion expressed by users might be analysed via the use of efficient Supervised Machine Learning (SML) algorithms to forecast the occurrence of illness and offer early warnings. In this paper we proposed a text mining classifier for generate the summarized text using machine learning techniques. After collecting the tweets, we got them ready for pre-processing and generate the class label for all instances such as correct, incorrect and neutral etc. In the second phase, numerous features are extracted from text by using a number of frequently used approaches, such as TF-IDF, co-relational, NLP and relational dependency features are extracted to generate the feature vector. As classification module we used one binary classification algorithm and five machine learning algorithms for evaluation of proposed model. NLP-SVM and TFIDF-SVM produces higher accuracy 95.10% and 93.50% classification accuracy respectively. This demonstrates the proposed model is effective for classification of large text for COVID-19 on tweet data.

Keywords*: COVID-19, machine learning classifier, supervised learning, feature extraction, feature selection, NLP, SVM.*

1. Introduction

Basically machine learning are divided into three different kind of learning methods such as supervised, semi supervised and reinforcement learning. The method known as supervised machine learning involves training the computer with the assistance of labelled datasets. In this method, the instances are accurately classified in accordance with the category to which they belong. The data will be analysed by the machine, and ultimately it will be able to make predictions about future occurrences based on information that it has learnt from previous examples. Unsupervised machine learning, on the other hand, is different from supervised machine learning in that it can learn on its own without the existence of data that has been properly categorised. During the unsupervised learning phase of machine learning, the machines will be provided with training samples to analyse, and it will be the computer's responsibility to uncover any concealed patterns within the dataset. For the purpose of reinforcement learning, the computer plays the role of an agent that seeks to identify the most suitable behaviours by using a technique based on experimentation and monitoring of the surrounding environment. In the event that the machine fails to successfully complete a task, it will be compensated by having its state increased; in the event that it does successfully complete the task, it will be forced its state decreased, and this strategy will be utilised multiple times until the computer system learns how to effectively complete a particular task. In order to teach robots how to conduct human-like jobs and provide special assistance, reinforcement learning is a technique that is used.

Therefore, the purpose of this research is to offer a deep learning method for the categorization of the feelings expressed in tweets relating to COVID-19. An examination of the comments made by people may assist determine if a person is feeling happy or negative emotions. In fact, there is an emphasis on automatically discovering texts in the huge text that include negative emotions. This method enhances the feature weights by recurrent neural network with semantically ordering by SVM, which, to the best of our knowledge, has not been applied similarly by earlier methods, particularly in comprehending sentiments from tweets connected to COVID-19. The primary objective is to maximise effective weight via the semantic connections between individual words using context learning and SVM. The

1Research Scholar
Amity University, Jaipur
suvarnalakshmic@adc.edu.in
2Associate Professor
AmityUniversity,Jaipur
ssaxena1@jpr.amity.edu
3Associate Professor,
SanjayGhodawat University,Kolhapur
sureshkumarbillakurthi@gmail.com

organization of the paper literature survey is described in section II by numerous existing researchers. The proposed system architecture and research methodology are demonstrated in section III. The algorithm details are described in section IV, which describes supervised classifier training and testing. Section V focus on extensive experimental analysis of proposed system and various existing systems, while section VI demonstrates a conclusion and future work proposed model.

2. Literature Survey

A technique to classifying the feelings expressed in microblog evaluations that incorporated emojis was published by Li et al. [1]. In the form of an emoji-text-incorporating bi-LSTM model that was given the name ET-BiLSTM. Emojis, as vector representations, were provided to the model that was being suggested. Based on the findings, it is clear that ET-BiLSTM improves the overall performance of sentiment categorization. They classified people's feelings using a BERT model that they implemented. After that, they performed an investigation into the manner in which attitudes towards the vaccination are dispersed around the globe by analysing the hotspot locations and making use of kernel density estimation. Another research was suggested by Balli et al. [2] at COVID-19 to classify public datasets and Sentiment data that had been manually labelled for positive and negative Turkish tweets. During the pre-processing of the dataset, the Zemberek [3] library and the SnowBall library were both used as separate libraries. In addition, the data were tokenized using TF-IDF before being provided to ML algorithms like CNN and RNN, however for LSTM, the data were presented by the tokenizer class. This was done so that the data could be processed by the ML algorithms. It was found that the models that were applied to Sentiment had superior performances, and the accuracy of the data when it was negatively weighted was greater than when it was positively weighted.

In a separate piece of research, Sitaula and Shahi [4] developed a hybrid feature in order to represent tweets from Nepal. They used two-word representations, such as the bag-of-words (BOW), with FastText-based approaches as well as methods that were specialised to the domain. Convolutional neural networks with several channels were then fed the concatenated representations as input (MCNN). According to the findings, the combination of features performed better than the individual features with an accuracy of 69.7 percent, while the MCNN model attained 71.3 percent accuracy in comparison to traditional methods. Using increased feature weighting and the attention mechanisms of LSTM-RNN, Singh et al. [5] performed research to categorise sentiment Twitter data linked to COVID19. TF-IDF was used to extract tweets features. The findings of the experiments demonstrated that the suggested technique surpassed the rest of the traditional machine learning algorithms. In a study that was conducted by Parimala et al. [6,] the researchers classified tweets that were relevant to catastrophic occurrences by using LSTM in conjunction with feature extraction approach. They suggested use an algorithm called risk assessment sentiment analysis (RASA). When compared with XGBoost and binary classifiers, the findings demonstrate that RASA attained a higher level of accuracy [6].

On a regular basis, people contributed their ideas, news, and experiences in confronting this virus via social media, which is regarded to be a large data centre throughout the epidemic. The use of social networking sites (SNSs) like Twitter as important resource for the identification and monitoring of various events, such as the spread of illness, has become more common. This online platform encouraged academics to analyse, in real time, the tweets that include people's sentiments and responses about a variety of topics, including election voting, the financial market, criminal activity, and hate speech [7].

In addition, the purpose of artificial intelligence (AI) in this present crisis has unquestionably helped to the research of the shift in human behaviours and worries in association with COVID-19 patients and fatalities during and after the pandemic. As a result, several COVID-19 surveillance models are looking for an efficient method of text processing and extracting information from COVID-19-related postings. This would result in early reports being generated, which may be critical for the prevention of outbreaks. The method in question is known as sentiment analysis (SA) [8], sometimes known as emotions mining [9], and it involves assigning positive, negative, or neutral connotations to a variety of texts based on their views. These phrases are given a preliminary processing using natural language processing (NLP) and then given a classification by machine learning (ML) [10]. These perspectives are highly helpful for constructing disease monitoring systems that are more efficient. There have been many research that have contributed to the analysis of tweets written in English; however, these studies have not taken into account the connection that exists between syntactic and semantic information and ML approaches that are based on feature types.

The researchers Samuel et al. [11] suggested two machine learning methods to classify tweets based on their sentiment: naive Bayes and logistic regression. These models separated tweets into two categories: positive and negative. In their study, the authors examine the efficacy of these models on two categories of data

with varying lengths of characters. The first group has data with less than 77 characters, while the second category contains data with 120 characters per tweet. In all categories, Naive Bayes fared better than logistic regression. For shorter tweets, NB reached an accuracy of 91.43 percent, whereas this was only 74.29 percent for LR. For lengthier tweets, NB achieved an accuracy of 57.14 percent, while LR evaluated an accuracy of 52 percent. An strategy to analysing tweets concerning the COVID-19 epidemic, based on the frequency of terms and sentiment analysis, has been utilised by researchers such as Rajput et al. [17]. In their method, word-level frequencies, bi-gram frequencies, and tri-gram frequencies are used to express word rates according to a power law distribution. As a result of this, we were able to identify three distinct kinds of tweets: negative, positive, and neutral.

An investigation that examines and depicts the global impact of COVID-19 has been presented by Muthausami et al. [18]. This investigation is based on analysis and visualisation. They divided the tweets into three categories using the approach of machine learning. There is a positive class, a neutral class, and a negative class. They used a variety of classifiers, such as support vector machines (SVM), naive Bayes, random forests, decision trees, LogitBoost, and MaxEntropy. The results obtained by the LogitBoost ensemble classifier were shown to be superior to those obtained by the other algorithms by the approach that was suggested.

Deep learning models, such as LSTM recurrent neural networks, were used in a study that was carried out by Jelodar et al. [19], and the researchers developed an algorithm to identify feelings based on these models (LSTM RNN). The subject modeller for COVID-19 was stated on social media, and natural language processing was used to create the classifier.

Aljameel et al. [20] examined a large dataset of tweets written in Arabic that were connected to COVID-19. The authors developed a model using machine learning in order to forecast and categorise the responses of Saudi Arabian residents to actions taken by the government and efforts to manage pandemics. In order to improve accuracy, they used uni-gram and bi-gram TF-IDF in conjunction with SVM, naive Bayes, and KNN classifiers. With an accuracy of 85 percent, the output findings demonstrated that SVM performed better than both KNN and naive Bayes. SVM and naive Bayes are two examples of machine learning classifiers that were used in Al-sukkar et al[21] .'s introduction of a sentiment analysis technique for analysing Arabic tweets as either negative or positive. They employed N-gram TF-IDF with 10-fold crossvalidation so that the accuracy of the classifiers may be improved. Based on the findings of the experiments, the support vector machine (SVM) demonstrated the best accuracy of 83.16 percent when using uni-gram, but the naive Bayes algorithm reached 81.93 percent accuracy when utilising bi-gram and tri-gram.

Imran et al. [22] have performed categorization of sentiments linked to COVID-19 tweets using a deep learning system called LSTM. Through the use of pre-trained Glove Twitter embedding, the application of LSTM on the sentiment 140 dataset was made much more effective. The primary purpose of this strategy was to determine, from users' tweets, the polarity of their sentiments as well as the users' feelings. As a consequence, the authors demonstrated that there is a substantial connection in sentiment polarity across nations that are geographically close to one another. Alam et al. [23] analysed tweets written in Arabic and English using SVM, FastText, and BERT respectively on a total of 504 tweets. The FastText model produced the most accurate results when applied to Arabic text. The methodology that was presented by Alqurashi et al. [24] included the use of a variety of machine learning classifiers that were applied to Arabic Tweets in order to identify false information relating to COVID-19. Additionally, the authors utilised TF-IDF, Word2Vec, and FastText feature embedding techniques in order to improve the accuracy of the classifiers. According to the findings, FastText achieved a high level of accuracy with its standard classifier, XGBoost, reaching 86.8 percent. Word2Vec, on the other hand, achieved a superior level of accuracy with its deep learning classifiers, reaching 85.7 percent using CNN. In a related manner, Naseem et al. [25] advocated the use of a variety of pre-trained embedding representations, such as TF-IDF, Word2Vec, and BERT, to extract features from a Twitter dataset. In addition, for the classification, they used deep learning approaches such as naive Bayes and SVM in addition to numerous traditional machine learning classifiers. Bi-LSTM was one of the deep learning methods. Other feature extraction approaches employing standard classifiers, such as SVM and RF, were outperformed by the TF-IDF model and FastText.

In addition, Basiri et al. [26] provided a model that combines five different models, including naive Bayes support vector machines (NBSVM), CNN, and bidirectional gated recurrent unit, using COVID-19 tweets in eight nations that are severely impacted. Their system, which is enhanced by a mechanism known as meta learning, obtained a high level of accuracy in the classification of attitudes, reaching 85.80 percent. The authors of the paper [27] offered a method for classifying COVID-19 tweets that was based on a variety of conventional machine learning methods. These algorithms included decision trees, XGBoost, extra tree

classifiers (ETC), random forests, and LSTM. They used bag-of-words (BOW) and TF-IDF algorithms in order to more accurately capture the text. According to the findings of the experiments, ETC was able to attain a greater level of accuracy with 93%.

In addition, Nemes and Kiss [28] created a model that uses an RNN model and the TextBlob technique [29] to identify tweets as either positive or negative. TextBlob was surpassed by the work that they suggested. A hybrid heterogeneous SVM technique was presented by Kau et al. [30] for the purpose of sentiment tweet categorization in connection with COVID-19 (HH-SVM). According to the findings, the suggested method performed much better than RNN. Researchers are looking for a strategy that will improve accuracy since they have found that sentiment analysis is a highly useful source of information in all of their prior investigations. In the current investigation, we examine several machines learning techniques, including TF-IDF, Word2vec, FastText, and Glove word embedding models, and offer our findings in the form of a comparison research. As a result, in order to get a high level of precision, we propose a hybrid approach that combines TF-IDF with two powerful word representations, namely embedding in Glove and FastText.

3. Proposed System Design

Initially data has extracted through the Twitter API which contains profile information as well as Twitter users' tweets and number of extracted Twitter records is 5000. Therefore, the data set consists of almost 3500 Twitter accounts. The extracted data is stored in both the MySQL database and the .txt file. The data collected from the Twitter account using the search query for data cleaning and filtration of null values from records also removes those records associated with misclassified instances or values. Data cleaning can be done interactively by transaction processing or systematic sampling technique. The systematic sampling techniques have been used for data filtering; once filtration has done, it becomes balanced data that eliminates the normalized dataset's misclassified instances

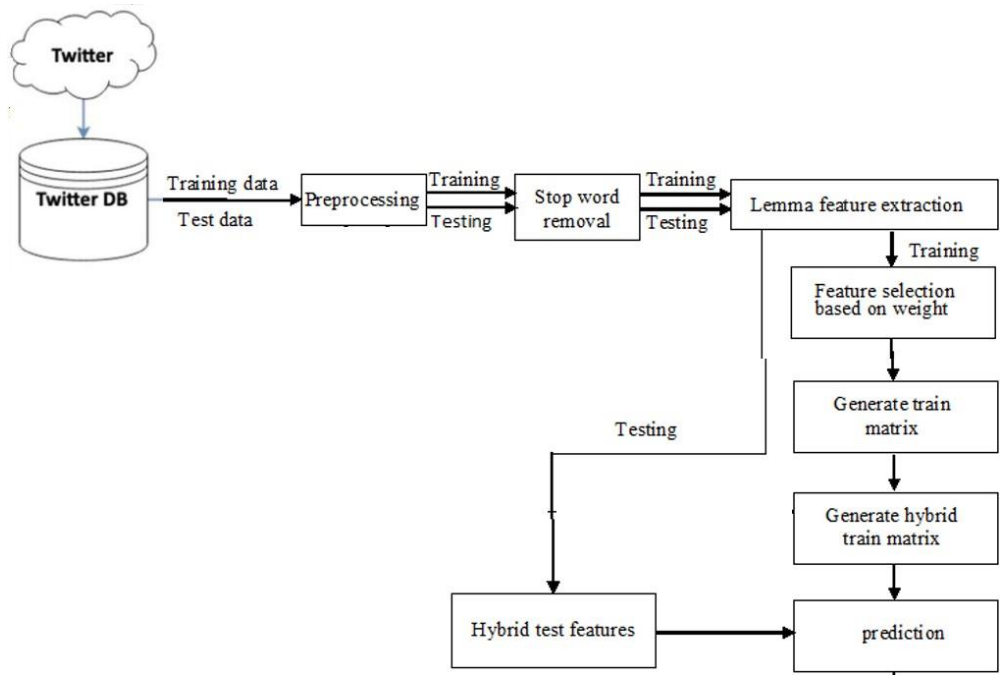

Fig. 1: proposed model for evaluation of large text data using supervised machine learning classifiers

Pre-processing: At this stage, the both train and test two datasets are cleaned up by having any punctuation marks removed and any contractions such as "can't," "isn't," and others replaced with "cannot" and "is not." The phrases used for training and testing are first processed via a stemmer, after which the stop words are removed.

Feature extraction: During feature extraction, the features we extracted from large text such as TF-IDF, relational features, lemmas, Bigram features and dependency-based features from text data. The combines feature set generate the hybrid feature vector that called as hybrid feature set. Hybrid feature selection generates a training matrix using lemma and rule-based features. The training matrix of both feature classes are concatenated. The multiple classification label is established using the same testing technique as the basic system.

Feature selection: Once feature extraction has been done, using few quality thresholds, we optimized a feature set called feature selection. The weighted term

frequency technique has been used to optimize features and forwarded to the training module.

Classification: Finally, the system detects each transaction, and predict the class label using a supervised classification technique. We carried out various machine learning as a supervised classification algorithm which used for train and test the classifier. Here class labelled data is present at the beginning. ANN, NB, RF and SVM algorithm for machine learning is the algorithm applied on entire data. The results have been evaluated according to the confusion matrix and generated the accuracy, precision, recall and F1-Score etc.

Analysis: After completion of entire process analysis has done with the calculation of confusion matrix such as TP, TN, FP and FN. The accuracy the major parameter that used evaluation of system performance.

ALGORITHM DESIGN

This hybrid algorithm performs the classification task for input test data, and utilized the trained module as background knowledge.

Algorithm 1: Feature selection using lemmas method.

Input: TF[i]k is the set of selected features containing lemmas.

Output: A train matrix.

Step 1: for each feature/term f in TF

Frequenct(f) ← TF[f]k

if (Frequenct(f)

weightedList.add(f)

end for

Step 2: for each term t in weightedList calculate weight

if weight(t) >Thresholdk

add t in weight[t][w], where t is the term and w

is its weight in corresponding aspect category

kw.

end for

Here, threshold on weight is different for each aspect category.

Step 3: A matrix X[i][j] is generated, where i = {t1, . . ,tn}

and j={k1, . . ,k5} for 5 aspect categories. Each row in matrix is weight of X[t] in k1 . . k5 class

Step 4: Generate a binary train matrix MatB[i][j].

for each term Mat[ti] in Mat[i][j]

for j from k1 to k5

if Mat[ti][j] >0.0 then

MatB[ti][j]=1

else

MatB[ti][j]=0

end if

end for

end for

Algorithm 2: Classification Algorithm

Preprocessing: Input test sentences are preprocessed and lemma features are extracted.

Input: Binary train matrixMatB[i][j], Lemma[i]={lemmas for each sentence i in test dataset}.

Output: Aspect label prediction for each test sentence.

Step 1:

for each sentence i from test dataset

for each lemma L in lemma [i]

for each term ti in MatB

if lemma L in lemma[i] = = term in MatB

test[L][1..5]=MatB[t][1..5]

copy binary vector from MatB to test vector

for lemma L.

end if

end for

end for

end for

Step 2: Calculate conditional probability of each aspect category.

Step 3: Return aspect category corresponding to highest probability score.

4. Results And Discussion

The column of the matrices shows the instances in an actual class, whereas the row depicts the predicted instances in a class. The name comes from the fact that the approach makes it easy to see whether two groups are confused. In supervised learning, an uncertainty matrix is a simple tool for evaluating outcomes. It's used to describe the test result of a prediction model. Column of

the matrix elaborates the instances in a predicted class, whereas each row depicts those classes in a class diagram. Four independent experiments were performed to test the discriminant function for various dataset formats.

Table 1: Confusion matrix table

Confusion matrix		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

The calculation strategy for confusion matrix in implemented after experimental analysis which is defined in below section:

+9666: It gives negative prediction for actual negative label classes,

FP: It gives positive as prediction to all negative classes,

FN: It given negative to all positive classes,

TP: It gives positive prediction for positive classes.

The accuracy (Eq. 1) is the percentage of accurate predictions out of an overall number of projections. The equation is used to measure it:

$$Acc=(TP+TN)/(TP+TN+FP+FN) \quad (1)$$

The recommended accuracy has been estimated using the equation described, and it achieves about 97.23% precise forecasts, which is better than all other methods. To compare the outcomes of different experiments, the F1-score was used as an assessment measure in this analysis. Convergent and discriminant validity are used to measure the F1 score (Eq. 2). TP stands for a positive result, FP for false positive, and FN for the negative test in Eqs. (3) and (4).

$$F1=2\times(precision\times recall)/(precision+recall) \quad (2)$$

$$precision=TP/(TP+FP) \quad (3)$$

$$Recall=TP/(TP+FN) \quad (4)$$

The proposed implementation has been carried out in an open-source environment for Windows, and the Java Platform has been utilised since open source is readily available. The data from the Twitter online application has been extracted using the Twitter API, which is open to the public. In order to improve the accuracy of the system's categorization, we generate a variety of data chunks and then use three distinct machine learning algorithms. For the purpose of data splitting, three distinct methods of cross-validation, namely 5-fold, 10-fold, and 15-fold individually, have been used.

Dataset Description

The implementation of machine learning procedures and research techniques for text content to obtain data is web mining. Due to unprecedented growth in digital textual information from websites, Google initiatives such as Google Scholar and Google N-gram, and popular media platforms such as Twitter, emotion analysis has gained a lot of attention in recent years. Data from Twitter is a rich source that can be used to gather data on any subject imaginable. These data could be used in various use cases, such as detecting patterns related to a particular keyword, assessing keyword rankings, and collecting feedback on emerging technologies.

To implement the proposed system, we have used the Twitter dataset from the Twitter account using provides API. This process determines the random data collections from Twitter accounts with selective attributes. Below we have mentioned the entire process of data extraction using API. For dataset, first, create an account on Twitter than from developer.twitter.com\app we can generate consumer-key and consumer-secret-key.

Twitter consists of 3 different APIs such as Standard, Premium, and Enterprise API. Twitter has around 22 attributes that contain some categorical attributes as well as some that are numerical. The data is an imbalance; it is mandatory to filter those entire where it contains misclassified instances. The table 2 older version of the Twitter dataset, while Table 3 having a newly updated dataset with additional attributes.

Table 2: New Twitter Dataset Extracted for Proposed System

Sr. No	Attribute Name	Description
1	NAME	On Twitter, the account holder's name is shown.
2	SCREEN-NAME	Account generated date time or timestamp with data year and month Token name or nickname for the account
3	CREATED-DATE	The total number of followers for each account user.
4	FOLLOWERS-COUNT	The total number of friends a person has on their account.
5	FRIENDS-COUNT	Account holder's longitude (this is subpart of location)

Sr. No	Attribute Name	Description
6	USER-LANGUAGE	The total number of Twitter groups in which the account participates is counted.
7	LISTED-COUNT	The user's chosen profile picture
8	PROFILE-IMAGE	The cumulative number of tweets sent by the user of the respective account.
9	STATUS-COUNT	Account holder's location records, such as city or country names
10	USER-LOCATION	The account holder's time zone
11	TIME-ZONE	The account's UTC offset, as determined by the TIMEZONE
13	UTC-OFFSET	The longitude of the location where the most recent tweet was sent
14	LOCATION-LATITIDE	The longitude of the location where the most recent tweet was sent
15	LOCATION-LONGITUDE	On Twitter, the account holder's name is shown.

The Table 2 shows extracted data from Twitter API that is open source available of the basic version; due to this reason, it having some limitations for runtime data extraction like some attributes could provide null values or dirty values.

Table 3: Proposed System Dataset Extracted from Twitter using Twitter API

Sr. No.	Attribute Name	Description
1	USER Name	The name displays of account holder on twitter
2	SCREEN NAME	The token name or nickname for the account
3	CREATED	Account created date time or timestamp with data year and month
4	PROFILE IMAGE	User account profile image
5	LOCATION	Location of users
6	USER LANGUAGE	Preferred Language selected by account owner
7	FRIENDS COUNT	Total number of friends
8	FOLLOWERS COUNT	Total number of followers
9	STATUS COUNT	The total number of tweets given by respective account user
10	LIST COUNT	The total number of twitter groups the account has belongs count
11	TWEET COUNT	Total number of tweets posted by user
13	UTC OFFSET	The total number of twitter groups the account has belongs count
14	LANGITUDE	User location longitude value
15	LATITUDE	User location latitude value
16	TWITTER POST	Last tweeter post by user
17	USER URL	User account profile URL
18	DESC DATA	Description of posted data
19	TIMEZONES	Time zone of current location (GMT)
20	GENDER	Gender of account holder
21	AGE	age of account holder
22	TWEETS	No of tweet posted by account holder
23	TOPIC	Tweet topic description
24	RETWEET	Retweet topic name
25	POSITIVE SENTIMENT	No of positive tweets

Sr. No.	Attribute Name	Description
26	NEGATIVE SENTIMENT	No of negative tweets

Table 3 describes the newly updated downloaded dataset from Twitter accounts. It contains around 26 attributes that are used for training and testing. Each attribute shows the identity of the respective user. The various cross-validation techniques help us effective classification accuracy for real-time data.

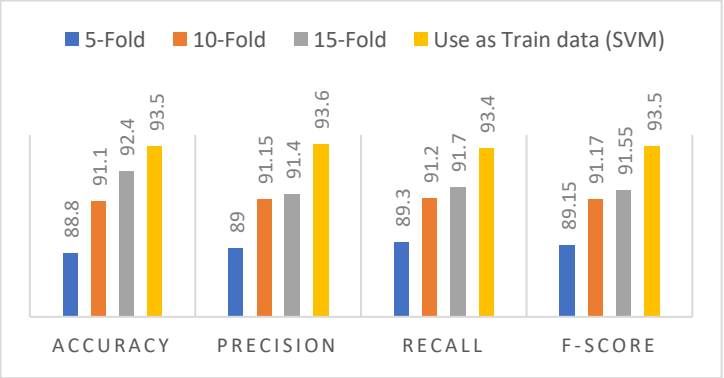

Fig. 2: classification accuracy with proposed system TF-IDF features and SVM classifier

Figure 2 describes a classification accuracy for Twitter data using a TF-IDF-based feature extraction method with an SVM classifier. The 0.25 threshold has been used for the selection of TF-IDF features. The four techniques are evaluated with various cross-validation, and it provides 93.50% is higher classification accuracy.

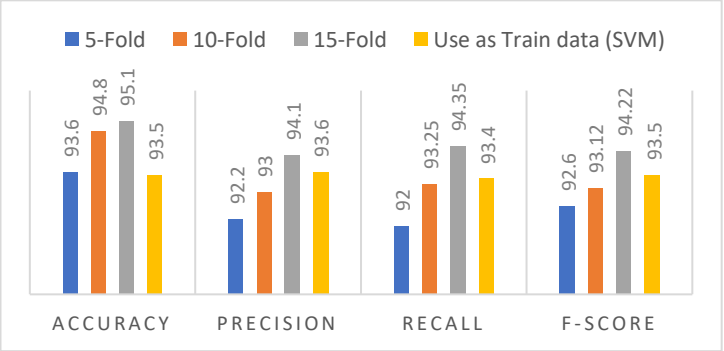

Fig. 3: classification accuracy with proposed system NLP features and SVM classifier

A classification accuracy for Twitter data is shown in Figure 3. This accuracy was achieved by combining an SVM classifier with a natural language processing NLP-based feature extraction approach. The 0.25 criterion was used throughout the selection process for NLP features. The four methods are analysed using several cross-validation methods, which results in a classification accuracy of 95.10%.

5. Conclusion

Tweets have served as a source of knowledge and a potential trigger for disease monitoring models during the COVID-19 epidemic, with the circumstances and people's reactions and emotions constantly shifting throughout this crucial time. Public health agencies may benefit from analysing tweets regarding the early detection of outbreaks. In this study, we suggested a survey of COVID-19 emotions expressed in retweets, which is a trustworthy and valuable source of data for analysing a lot of data and examining consumer activities. To cultivate the accuracy of system we used pre-processing approaches using NLP and other characteristics embedding techniques. We proposed combining NLP and TF-IDF to get the best performance for improving classification and computer vision model accuracy. When combined in one model, the syntactic presentation by TF-IDF and the semantic representations of text by relation have shown to be complementary in better collecting the information from tweets. Compared to other machine learning models, the research found that SVM performed better than the competition and achieved superior performance using our two-feature fusion strategy. Our solution, which combined the TF-IDF unigram, TF-IDF N-gram, and hybrid features embedding techniques, produced the best results with supervised machine learning classifiers out of all seven approaches examined. Additionally, we contrasted the effectiveness of our model using the top feature

approaches with results from earlier research. We concluded that the methods we suggested work a little better. This is because, in contrast to other evaluated techniques, the combination of two efficient document representation may location and provides words' lexicon and extract their attributes.

References

[1]. Li, X.; Zhang, J.; Du, Y.; Zhu, J.; Fan, Y.; Chen, X. A Novel Deep Learning-based Sentiment Analysis Method Enhanced with Emojis in Microblog Social Networks. Enterp. Inf. Syst. 2022, 1–22.

[2]. Balli, C.; Guzel, M.S.; Bostanci, E.; Mishra, A. Sentimental Analysis of Twitter Users from Turkish Content with Natural Language Processing. Comput. Intell. Neurosci. 2022, 2022, 2455160.

[3]. Zemberek, NLP Tools for Turkish. Available online: https://github.com/ahmetaa/zemberek-nlp (accessed on 20 September 2021).

[4]. Sitaula, C.; Shahi, T.B. Multi-channel CNN to classify nepali COVID-19 related tweets using hybrid features. arXiv 2022, arXiv:2203.10286.

[5]. Singh, C.; Imam, T.; Wibowo, S.; Grandhi, S. A Deep Learning Approach for Sentiment Analysis of COVID-19 Reviews. Appl. Sci. 2022, 12, 3709.

[6]. Parimala, M.; Swarna Priya, R.; Praveen Kumar Reddy, M.; Lal Chowdhary, C.; Kumar Poluru, R.; Khan, S. Spatiotemporal-based sentiment analysis on tweets for risk assessment of event using deep learning approach. Softw. Pract. Exp. 2021, 51, 550–570.

[7]. Kabir, M.; Madria, S. CoronaVis: A real-time COVID-19 tweets data analyzer and data repository. arXiv 2020, arXiv:2004.13932.

[8]. Taboada, M. Sentiment analysis: An overview from linguistics. Annu. Rev. Linguist. 2016, 2, 325–347.

[9]. Beigi, G.; Hu, X.; Maciejewski, R.; Liu, H. An overview of sentiment analysis in social media and its applications in disaster relief. In Sentiment Analysis and Ontology Engineering; Springer: Berlin/Heidelberg, Germany, 2016; pp. 313–340.

[10]. Sailunaz, K.; Alhajj, R. Emotion and sentiment analysis from Twitter text. J. Comput. Sci. 2019, 36, 101003.

[11]. Samuel, J.; Ali, G.; Rahman, M.; Esawi, E.; Samuel, Y. COVID-19 public sentiment insights and machine learning for tweets classification. Information 2020, 11, 314.

[12]. Liu, R.; Shi, Y.; Ji, C.; Jia, M. A survey of sentiment analysis based on transfer learning. IEEE Access 2019, 7, 85401–85412.

[13]. Tyagi, P.; Tripathi, R. A review towards the sentiment analysis techniques for the analysis of twitter data. In Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering (ICACSE), Sultanpur, India, 8–9 February 2019.

[14]. Saura, J.R.; Palacios-Marqués, D.; Ribeiro-Soriano, D. Exploring the boundaries of open innovation: Evidence from social media mining. Technovation 2022, 102447.

[15]. Mackey, T.; Purushothaman, V.; Li, J.; Shah, N.; Nali, M.; Bardier, C.; Liang, B.; Cai, M.; Cuomo, R. Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: Retrospective big data infoveillance study. JMIR Public Health Surveill. 2020, 6, e19509.

[16]. Wan, S.; Yi, Q.; Fan, S.; Lv, J.; Zhang, X.; Guo, L.; Lang, C.; Xiao, Q.; Xiao, K.; Yi, Z.; et al. Relationships among lymphocyte subsets, cytokines, and the pulmonary inflammation index in coronavirus (COVID-19) infected patients. Br. J. Haematol. 2020, 189, 428–437.

[17]. Rajput, N.K.; Grover, B.A.; Rathi, V.K. Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. arXiv 2020, arXiv:2004.03925.

[18]. Muthusami, R.; Bharathi, A.; Saritha, K. COVID-19 outbreak: Tweet based analysis and visualization towards the influence of coronavirus in the world. Gedrag Organ. Rev. 2020, 33, 8–9.

[19]. Jelodar, H.; Wang, Y.; Orji, R.; Huang, S. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: Nlp using lstm recurrent neural network approach. IEEE J. Biomed. Health Inform. 2020, 24, 2733–2742.

[20]. Aljameel, S.S.; Alabbad, D.A.; Alzahrani, N.A.; Alqarni, S.M.; Alamoudi, F.A.; Babili, L.M.; Aljaafary, S.K.; Alshamrani, F.M. A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent COVID-19 outbreaks in Saudi Arabia. Int. J. Environ. Res. Public Health 2021, 18, 218.

[21]. Ghadeer, A.S.; Aljarah, I.; Alsawalqah, H. Enhancing the Arabic sentiment analysis using different preprocessing operators. New Trends Inf. Technol. 2017, 113, 113–117.

[22]. Imran, A.S.; Daudpota, S.M.; Kastrati, Z.; Batra, R. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. IEEE Access 2020, 8, 181074–181090.

[23]. Alam, F.; Dalvi, F.; Shaar, S.; Durrani, N.; Mubarak, H.; Nikolov, A.; Martino, G.D.S.; Abdelali, A.; Sajjad, H.; Darwish, K.; et al. Fighting the COVID-19 infodemic in social media: A

holistic perspective and a call to arms. arXiv 2020, arXiv:2007.07996.

[24]. Alqurashi, S.; Hamoui, B.; Alashaikh, A.; Alhindi, A.; Alanazi, E. Eating garlic prevents COVID-19 infection: Detecting misinformation on the arabic content of twitter. arXiv 2021, arXiv:2101.05626.

[25]. Naseem, U.; Razzak, I.; Khushi, M.; Eklund, P.W.; Kim, J. Covidsenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. IEEE Trans. Comput. Soc. Syst. 2021, 8, 1003–1015.

[26]. Basiri, M.E.; Nemati, S.; Abdar, M.; Asadi, S.; Acharrya, U.R. A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. Knowl.-Based Syst. 2021, 228, 107242.

[27]. Rustam, F.; Khalid, M.; Aslam, W.; Rupapara, V.; Mehmood, A.; Choi, G.S. A performance comparison of supervised machine learning models for COVID-19 tweets sentiment analysis. PLoS ONE 2021, 16, e0245909.

[28]. Nemes, L.; Kiss, A. Social media sentiment analysis based on COVID-19. J. Inf. Telecommun. 2021, 5, 1–15.

[29]. Vel SS. Pre-Processing techniques of Text Mining using Computational Linguistics and Python Libraries. In2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) 2021 Mar 25 (pp. 879-884). IEEE.

[30]. Kaur, H.; Ahsaan, S.U.; Alankar, B.; Chang, V. A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets. Inf. Syst. Front. 2021, 23, 1417–1429.