Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7922979

IMPROVED SVM- FUZZY SENTIMENT ANALYSIS USING BERT ALGORITHM

Mr.S. Niresh
Ph.D-Research Scholor, PG and Research Department of Computer Science, Government Arts and Science College, Kangeyam, Thirupur DT, Tamilnadu,
India, e-mail:ks.niresh@gmail.com.

Dr C.Sathya
Assistant Professor, PG and Research Department of Computer Science, Government Arts and Science College, Kangeyam, Thirupur DT, Tamilnadu, India,
e-mail:sathyavenkateswaran@gmail.com.

Abstract- Categorizing the tweets are difficult based on their simplicity and use of standard or non-standard slang. There have been several types of research that have discovered highly precise sentiment data classifications, but very few of them have been validated on Twitter data. Techniques of sentiment analysis in the past have mostly dealt with interpreting binary or ternary emotions in unilingual texts. Furthermore, emotions become apparent in writings that are written in more than one language. Today's society increasingly relies on the usage of emotions and summarise key points in written communication, such as tweets on the Covid Vaccine. Earlier machine learning methods focused exclusively on text classification, image classification, or emotion classification that ignores the majority of emotions. Based on the examination of text and emotions in tweets on the Covid Vaccine Sentiment Analysis (SA), this paper propose improved SVM- Fuzzy using BERT technique together to enhance the performance measures. The proposed Improved SVM- Fuzzy method is to improve the performance measures (accuracy) and reduce the error rate based on feature extraction and classification in sentiment analysis. In experimental circumstances, the proposed model outperformed state-of-the-art approaches in tweet emotion identification and text recognition by 97.3%.

Keywords: BERT, SVM- Fuzzy, Twitter Data, Sentiment Analysis, Emoji.

I. INTRODUCTION

SA is a Natural Language Processing (NLP) technique that helps determine if an opinion is positive, negative, or neutral [1]. Some of the uses include spotting online trends, assessing reviews, and monitoring brand and product markets based on user feedback feelings. Sentiment analysis often follows a broad structure that involves gathering raw data, cleaning the data to eliminate noise, translating all pre-processed data to a computationally appropriate format, and labeling the training data [2].

Furthermore, sentiment analysis algorithms are classified into three types: rule-based, automated, and hybrid. Manually generated rules underpin rule-based algorithms, while machine learning methods underpin automated algorithms, and hybrid algorithms integrate both rule-based and automatic approaches [3].

Social networks have risen tremendously, and numerous social services have drawn millions of people in a short period. As a result, social networks play an important part in people's lives such as speak and post about their daily activities to generate an unprecedented quantity of original information. Any of the lucrative domains where this data is gathered is SA and Emotional Analysis (EA) [4]. Twitter, for example, gets a lot of different tweets every day. Many intellectuals are interested in how to classify the polarities of these tweets, and they want to come up with a better way to classify both polarity and the emotional responses of users [5].

According to the findings of prior research, the majority of the methods used constrained binary categorization, which can be summarised as either positive or negative, instead of inclusive analysis or what is known as fine-grained classification. The nomenclature was so vague, that it was necessary to do an analysis that included both associates and inclusive. The requirement might perhaps be satisfied by merging the methodology of Fuzzy logic with that of NLP. Even though Fuzzy logic techniques may improve efficiency with various weightages and capabilities, resulting in a more accurate analysis [6], according to the research that is currently accessible, Fuzzy logic approaches do not see the widespread application in the field of SA.

In this paper, we take a glance at how BERT and SVM, which are the most prevalent pretrained linguistic predictions based on Transformer, perform in terms of accuracy in the classification of the activities of sentiment analysis and emotion recognition [7]. To test their performance, we offer two BERT-based models for Covid Vaccine tweets text including emoji categorization with enhanced SVM- Fuzzy. The remainder of the paper focuses on data obtained from microblogging networks, namely Twitter [8].

The main reasons for this preference are the wide selection of tweets (as opposed to, say, Facebook posts, which have various data policies) and the fact that such data are typically difficult to analyze due to the presence of slang, typos, and acronyms (e.g., "btw" for "by the way") and thus represent a beneficial benchmark for text classifiers [9].

The remaining part of the paper is organized as follows: the following section presents some related works on sentiment analysis using SVM- FUZZY, BERT, and emotion recognition. Section 3 proposes the architecture of models employed for both functions, while Section 4 reveals the outcomes of the experimental analysis of the models on real-world tweet datasets. Section 5 concludes by drawing conclusions and discussing future work.

II. RELATED WORKS

Emoji Prediction

A more entertaining kind of sentiment analysis is called emoji prediction. Can you determine what your pals are feeling just by reading their text messages? Are they in a good mood? Would it be possible for you to attach the proper smiley face to each text message that you receive? If that's the case, it seems like you get their point of view.

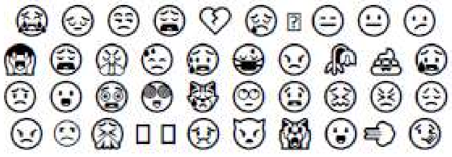
In this article [18], we construct something that is known as a classifier, which is an algorithm that learns to link emojis with phrases. Even though there are a lot of complicated technological aspects, the basic idea behind the classifier is rather straightforward. To begin, we gather a huge number of phrases from Twitter conversations that include emojis and then analyze them. After that, we analyze specific aspects of those phrases (words, word pairings, etc.) to prepare our classifier to link certain aspects with the smileys that are already known to us [19].

“For instance, if the classifier notices the word "happy" popping up in a lot of phrases that also include the smiley face symbol, they'll know to look out for it😂, It will acquire the ability to categorize such messages as 😂.” On the other hand, the word "happy" can be followed by the phrase "not," in which case we shouldn't rely on simply individual words to be associated with certain smileys [20]. For this reason, we also look at lexical items, while in this specific situation, we would find that the term "not happy" is significantly strongly related to sadness, outweighing the "happy" portion of the word. This is something that we discover when we look at lexical items. The classifier has been taught to consider all of the possible word sequences that can be found in a phrase to identify which type of smiley would best characterize the text. This is done to classify the sentence. Even if the idea behind it is straightforward, we may be able to improve our performance on this job by a significant amount if we are provided with a large number of text that is related to recognized smileys.

The following are three sample statements, along with the emojis that the classifier believes best represent them:

😂	Vaccination is available free of cost.
😢	I don't like it
😱	My neighbor didn't take the vaccine

Table 1: Emojis and their Sentimental tendencies.

Sentimental Tendencies	**Emojis**
Happy	
Sad	

Despite the lack of consensus on which emotions are the basic emotions of humans, the scientific world is increasingly focused on the particular topic of emotion detection. Table 1 illustrates emojis and their emotional characteristics. For code-switched emotion prediction, a multilingual wireless network technologies model has been developed. The remote supervision used a Gated Recurrent Unit (GRU) network to automatically construct a dataset for emotion identification and train a fine-grained emotions detection system. Another technique focuses on learning improved representations of emotional circumstances by pre-training neural models using billions of emoji instances on social media. Recently, a BERT method has been presented. Without any significant task-specific architectural alterations, our improved SVM-Fuzzy with the BERT model demonstrated state-of-the-art performance across a variety of NLP tasks.

Support Vector Machine- Fuzzy Model for Sentiment Analysis

The opinions and input of the community have always proven to be the most important and useful resource for businesses and organizations. As social media becomes more popular, it offers the door for unprecedented study and assessment of different elements for which companies previously had to depend on unorthodox, time-consuming, and error-prone approaches [10]. Sentiment analysis is a broad discipline that entails the successful categorization of user-generated material into predetermined polarity. Lexical approaches that accomplish classification using a dictionary-based annotated dataset and Hybrid technologies that combine machine learning with lexicon-based algorithms [11].

This study [12] proposes a sentiment analysis technique that uses SVM to bring together disparate sources of potentially relevant information, such as different favorability metrics for sentences and adjectives and, when available, information about the text's subject. Models including the new characteristics are then coupled with unigram models that have already been proved to be successful,and lemmatized variants of the word embedding models.

Sentiment analysis is a classification problem since it categorizes the direction of both a Covid Vaccine tweets text as positive or negative. This article [13] shows experimental findings from training a sentiment classifier using SVM on benchmark datasets. To isolate the most classical traits, N-grams and various weighting schemes were applied. It also investigates Chi-Square weight features to identify the useful characteristics for categorization of emotions.

Fuzzy logic on otherhand works with sets that have either an objective or subjective meaning, such as "tall," "large," or "beautiful" for conventional logic concerned the assertions of unquestionable truth. This is an effort to simulate how individuals analyze circumstances and come to conclusions, which often involves depending on values that are ill-defined or erroneous rather than ultimate truth or false.The researchers proposed the hybrid work to predict outcomes with more accuracy [14].

BERT model for Sentiment Analysis

The primary focus of traditional sentiment analysis is to categorize the overarching feeling conveyed by a piece of writing; however, this does not take into account other crucial details, such as the entity, subject matter, or component of the text that the emotion is directed toward [15].

Aspect-based sentiment analysis (ABSA) is a more difficult process than traditional sentiment analysis since it requires the identification of both feelings and aspects. To solve the out-of-domain ABSA, this study demonstrates the possibility of employing the context word vectors from pre-trained word vectors BERT in conjunction with a fine-tuning strategy that utilizes extra produced text [16]. The integration of a BERT tokenizer rather than a standard BERT Tokenizer is the enhancement that will be made to the technique that has been suggested. A variety of experiments have been carried out with a broad range of subjects (dialect and standard) [17].

III. METHODOLOGY

3.1 Preprocessing

A phase known as pre-processing is done to prepare the data for the more advanced techniques of sentiment analysis. The standard manufacturing procedure consisted of the following steps: Typically, the original Twitter data contains unique emoticons.

Removal of Stop word: Sometimes, removal of stop words is employed to convey the true meaning of a text with minimal input (e.g. a, and, the, etc.). After that, every stop word gets eliminated.

Stemming: Additionally, modifications of a term's root, such as derivatives, are employed in the majority of phrases. Therefore, varieties of words are often reduced to their stems or basic forms. This is utilized for the Porter stemming algorithm's function.

Tokenization: The phrase is deconstructed into the individual token words that make it up. Both of these are approaches to cleaning and preparing raw data for further analysis.

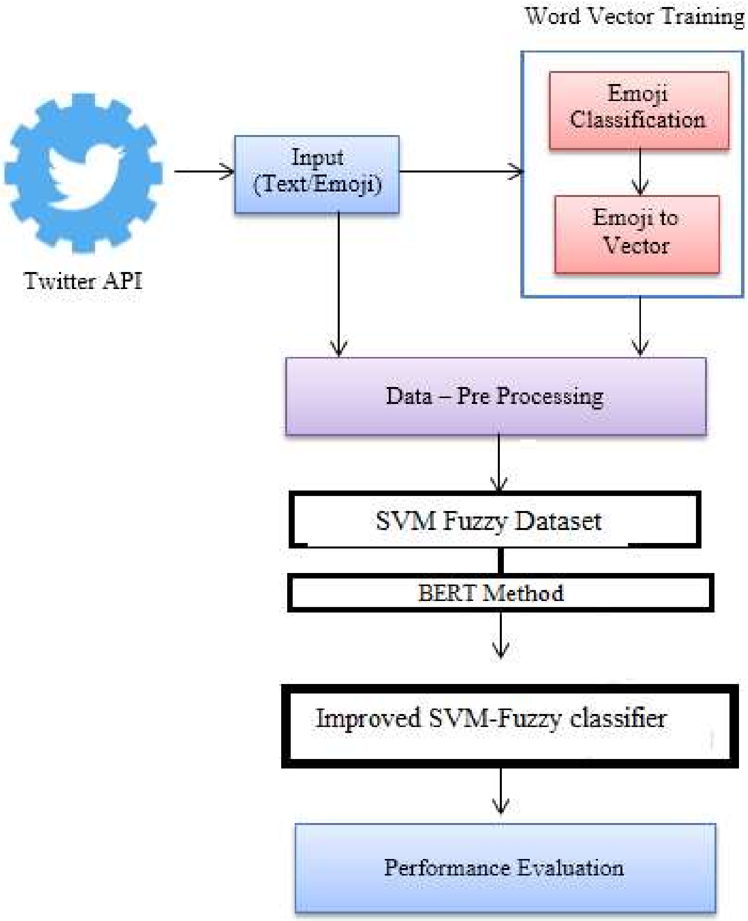


Figure 1: Proposed Architecture

3.2 BERT

BERT is developed to read the text of Covid Vaccine tweets to gather the context. The BERT model includes its method of feature extraction known as the BERT tokenizer, which utilizes BERT encoding to prepare input. The input data must be transformed appropriately so that each phrase may be delivered to a pre-trained algorithm to retrieve its embedding. BERT was trained on a huge text corpus, which offers the architecture/model the capacity to better comprehend language and learn a variety of data patterns, and to generalize effectively across several NLP tasks. Since it is bidirectional, the BERT training process involves the collection of data from the context token.

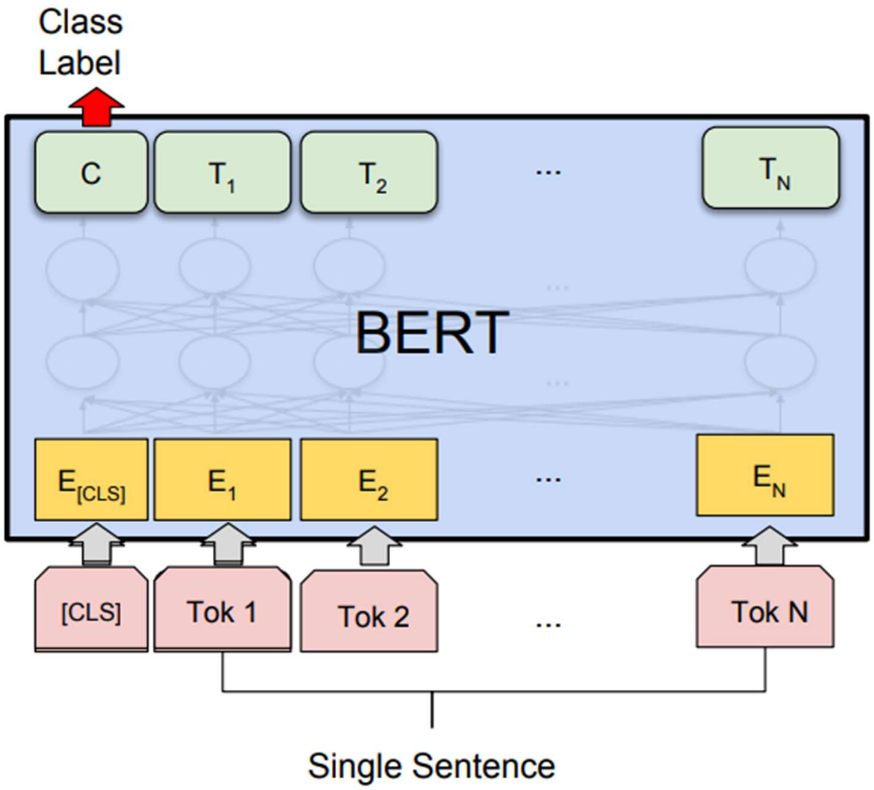


Figure 2. BERT Model (Source: https://www.geeksforgeeks.org/understanding-bert-nlp)

BERT's versatility to do various NLP tasks with state-of-the-art precision is one of its most significant characteristics as shown in fig 2.

3.3 Word2vec Training

Emojis are highly connected with the emotional tone of brief communications such as tweets. Using emojis to be one of the study objects in the emotional analysis of tweets allows for a more objective evaluation of the sentimental worth of tweets. To increase the accuracy of emotion categorization, one of the concerns that must be resolved is how graphical emojis should be utilized in conjunction with the text. emoji2vec is a vectorization technique. By converting emojis to vectors, they may be utilized in all aspects of language processing, in addition to words. Google open-sourced Word2vec in 2013 to convert textual words into standard formats that computers can interpret. Unsupervised learning of hidden data between words in unidentified training sets yields word vectors that retain syntactic and semantic links. The word vector is trained using the word2vec tool using the emoji2vec method. Following training on the processed corpus using the worf2vec program was generated.

Construct the description of the vector vec_i. w_1, w_2. . .$w_n$is a collection of word vector sequences that correlate towards the word sequences within the sample's description phrases. In this study, these word vectors are combined as emoticon description vectors. Then, the vector is described by the following Equation

$$vec_i = \sum_{n=1}^{M} w_n \qquad (1)$$

3.4 Improved SVM- Fuzzy Classification

Based on feature extraction, Fuzzy logic is used, followed by classification using a computational model. The key benefit of the suggested approach is its high precision and low error rate.

Fuzzification: The use of an appropriate membership function allows for the transformation of crisp inputs into Fuzzy inputs.

- **Input 1:**A range of [0 1] is used for the users' weights, and the trapezoidal function is used for the membership function, which has three levels: Light Impact (LI), Medium Impact (MI), and High Impact (HI).
- **Input 2:**The exact range for the polarization scores for tweets equals [1 1], and the triangle function has been chosen as the membership function for this range. This function has seven levels: Strongly Positive (SP), Weakly Positive (WP), Neutrality (N), Weakly Negative (WN), and Strongly Negative (SN).

Rule evaluation: The evaluation of Fuzzy sets often involves using several IF-THEN criteria. In this study, the preceding parameters of the principles of our Fuzzy system of inference are the user's degree of influence, which may be either LI, MI, or HI, and the polarity level of the tweet. Between all of our antecedent variables, we employ the logical AND conjunction to connect them.

Defuzzification: At this point, the defuzzification method known also as the center of gravity (CoG), which is the one that is employed the most often, is used. Table 2 shows the polarity score of different emotions from Twitter data. The outputs of all of the collected Fuzzy rules are then translated into a single crisp number, which indicates the polarities of the tweet when it has been completed. CoG is described as the following equation:

$$CoG = \int \frac{\mu_z(c).cdz}{\mu_z(c)dz} \qquad (2)$$

Table 2: Polarity Score of different emotions

Value	Polarity (P)
$0 \leq P \leq 0.5$	SP
$0.5 < P \leq 1$	WP
$-1 \leq P < 0.5$	WN
$-0.5 \leq P < 0$	SN

Proposed Algorithm:

```
Input: Text or Emoji from Twitter data.
Output: Sentiment Polarity scores (Strong Positive/Weak Positive/Strong Negative / Weak Negative/Netural)
Begin
Sentiment analysis ( ) ← Data
Input data passed into the BERT model for train and test
BERT model used to train in   SVM-FUZZY model.
calculate sentiment analysis score;
if score ( ) > 0 and <=0.5 then
            Sentiment analysis ( )←□Strong Positive
if score ( ) < 0.5  && score ( ) < =1 then
            Sentiment analysis ( ) ←□Weak Positive
if score ( ) < -0.5 &&score ( ) < =-0.1 then
            Sentiment analysis ( ) ←□Weak Negative
if score ( ) < =0 &&score ( ) < =--0.5 then
            Sentiment analysis ( ) ←□Strong Negative
else
            Sentiment analysis ( )← Neutral;
        else
        end
end
 validate the results( )
calculate accuracy ( );
```

Improved SVM- Fuzzy and BERT can be used to predict emoji among text on Twitter which is done broadly based on text. The results obtained from the individual SVM and BERT model suggest that improved SVM- Fuzzy performs better classification with the BERT model producing comparatively better word vector embeddings. An ensemble model would be developed to provide a fair result for text and emoji detection. In the Ensemble architecture, BERT is used for producing a vector embedding for all the sentences in the data set, and improved SVM- Fuzzy is used to classify sentences into respective classes based on the embedding vectors produced by BERT.

IV. EXPERIMENTAL ANALYSIS AND RESULTS

Both the sentiment analysis dataset and indeed the Tweet Sentiment Intensity datasets, which focus on emotion identification, were used in the evaluation of the suggested models. Experiments have been designed using the same criteria, and in addition, each dataset has been divided using a stratified sample method into a train (80%), dev (10%), and test (10%) set. In addition, we evaluated the BERT program in both its uncased and its cased forms. Experiments have been run on a laptop equipped with a 2x2.2GHz CPU, 8GB RAM, and an Nvidia GPU 740M graphics card. An accuracy with test data and correct instances for the different classifier is given in table 3.

Table 3: Accuracy with instances for sentiment classifier

Sentiment Classifier	Test Data	Correct instances	Accuracy %
Hybrid SVM- FUZZY	3000	2400	96
Improved SVM- FUZZY with BERT	3000	2400	97.3

Table 4 shows an accuracy comparison of BERT with improved SVM-Fuzzy with Hybrid SVM- Fuzzy on the presented dataset.

Table 4: Comparison between existing algorithm and the proposed algorithm

Version	Validation Accuracy	Test Accuracy
Hybrid SVM- Fuzzy	0.7775	0.96
Improved SVM- Fuzzy with BERT	0.9426	0.973

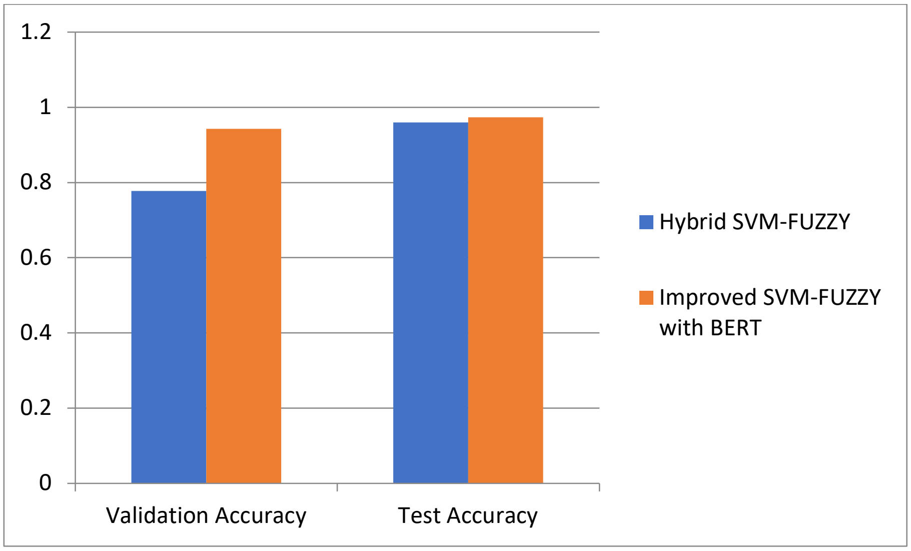


Figure 3. Algorithm Comparison for Validation and Testing Accuracy

The execution times described in the following subcategories apply to the hardware configuration used in these experiments. The classification accuracy and the F1 score were the two measures that were used in the evaluation of the model shown in fig 4 and table 5. Let the amount of information that originally belonged to the jth class but has been reclassified as the i-th class be denoted by the variable $x_{i\,j}$. Let's say that there are a total of n data, and let's call the amount of classes Ci.

The accuracy that a classifier can attain may be calculated as:

$$accuracy = \frac{1}{n}\sum_{i=1}^{Ci} a_{ii}(3)$$

The recall and precision of the i-th class are calculated as follows:

$$Precision_n = \frac{a}{\sum_{j=1}^{Ci} a_{ij}} \qquad (4)$$

$$recall_n = \frac{a_{ii}}{\sum_{j=1}^{Ci} a_{ij}} \qquad (5)$$

The F1 score of the i-th class equals:

$$F_{1i} = 2x\,\frac{precision_{\,i}\,.recall_i}{i\,precision_i + recall_i} \,(6)$$

As a result, a classification model's F1 score is calculated by taking the average of F1i:

$$F_1 = \frac{1}{Ci}\sum_{i=1}^{Ci} F_{1i} \qquad (7)$$

Table 5: Comparison of Existing and Proposed Algorithms

Version	Precision	Recall	F1 score	Accuracy %
Hybrid SVM-FUZZY	0.95	0.98	0.94	0.96
Improved SVM-FUZZY with BERT	0.97	0.982	0.95	0.973

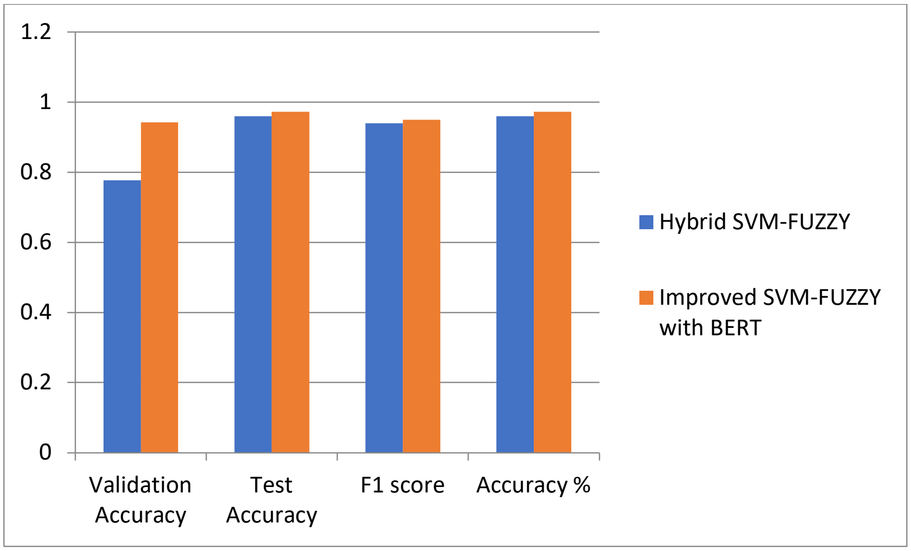


Figure 4: Comparative analysis of Performance measures for the existing and proposed system

V. CONCLUSION

The principal objective of this research is to provide a viable solution for Covid Vaccine Tweet sentiment data based upon that BERT language model. It is intended to be a two-stage pipeline, with the first phase requiring a variety of pre-processing techniques to transform Twitter jargon, including emojis and emoticons, onto plain text, and also the second step uses a variant of BERT to categorize tweets depending on their polarity. The categorizing of emotions has several uses. It is critical to provide an accurate and effective categorization method. Based on the examination of text and emoticons in tweets about the Covid Vaccine, this study proposes an algorithm and technique for sentiment analysis. We used the BERT algorithm to classify tweets, together with an Improved SVM- Fuzzy to analyze sentiment. The proposed model has been shown to outperform state-of-the-art methods in tweet emotion and text recognition by a margin of 97.3% in experimental settings.

REFERENCES

1. Nezhad, Z. B., &Deihimi, M. A. (2022). Twitter sentiment analysis from Iran about COVID 19 vaccine. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *16*(1), 102367.
2. Marcec, R., &Likic, R. (2022). Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. *Postgraduate Medical Journal*, *98*(1161), 544-550.
3. Swathi, T., Kasiviswanath, N., & Rao, A. A. (2022). An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. *Applied Intelligence*, 1-14.
4. Bibi, M., Abbasi, W. A., Aziz, W., Khalil, S., Uddin, M., Iwendi, C., &Gadekallu, T. R. (2022). A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. *Pattern Recognition Letters*, *158*, 80-86.
5. Loureiro, M. L., Alló, M., &Coello, P. (2022). Hot in Twitter: Assessing the emotional impacts of wildfires with sentiment analysis. *Ecological Economics*, *200*, 107502.
6. Mohammad, S. M. (2022). Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, *48*(2), 239-278.
7. Mohammad, S. M. (2022). Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, *48*(2), 239-278.
8. Boukabous, M., &Azizi, M. (2022). Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers. *Indones. J. Electr. Eng. Comput. Sci.*, *25*(2), 1131-1139.
9. Sivakumar, S., &Rajalakshmi, R. (2022). Context-aware sentiment analysis with attention-enhanced features from bidirectional transformers. *Social Network Analysis and Mining*, *12*(1), 1-23.
10. Maruthamuthu, A., Murugesan, P., &Muthulakshmi, A. N. (2022). A New Methodology to Arrive at Membership Weights for Fuzzy SVM. *International Journal of Fuzzy System Applications (IJFSA)*, *11*(1), 1-15.
11. AlBadani, B., Shi, R., & Dong, J. (2022). A novel machine learning approach for sentiment analysis on Twitter incorporating the universal language model fine-tuning and SVM. *Applied System Innovation*, *5*(1), 13.

12. Hidayat, T. H. J., Ruldeviyani, Y., Aditama, A. R., Madya, G. R., Nugraha, A. W., &Adisaputra, M. W. (2022). Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier. *Procedia Computer Science*, *197*, 660-667.
13. Jain, D. K., Boyapati, P., Venkatesh, J., & Prakash, M. (2022). An intelligent cognitive-inspired computing with big data analytics framework for sentiment analysis and classification. *Information Processing & Management*, *59*(1), 102758.
14. Hantsch, P., &Chkroun, N. (2022, July). connotation_clashers at SemEval-2022 Task 6: The effect of sentiment analysis on sarcasm detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 945-950).
15. S.Niresh, & C.Sathya (2022). Twitter Sentiment Analysis using Hybrid SVM – Fuzzy. Journal of the Asiatic Society of Mumbai, ISSN: 0972-0766, Vol. XCV, No.9. Page No. 20-30
16. Chandra, R., & Kulkarni, V. (2022). Semantic and sentiment analysis of selected bhagavadgita translations using BERT-based language framework. *IEEE Access*, *10*, 21291-21315.
17. Venugopalan, M., & Gupta, D. (2022). An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis. *Knowledge-Based Systems*, *246*, 108668.
18. Li, X., Zhang, J., Du, Y., Zhu, J., Fan, Y., & Chen, X. (2022). A Novel Deep Learning-based Sentiment Analysis Method Enhanced with Emojis in Microblog Social Networks. *Enterprise Information Systems*, 1-22.
19. Babu, N. V., &Kanaga, E. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: A review. *SN Computer Science*, *3*(1), 1-20.
20. Bansal, B., & Srivastava, S. (2019). Lexicon-based Twitter sentiment analysis for vote share prediction using emoji and N-gram features. *International Journal of Web Based Communities*, *15*(1), 85-99.