HYBRID MODEL FOR SENTIMENT ANALYSIS OF TWITTER DATA

1. Harbhajan Singh, Department of Computer Science and Applications, Khalsa College, Amritsar, India
E-mail: hsrandhawakca95@gmail.com

2. Vijay Dhir, Professor, Department of CSE, Sant Baba Bhag Singh University, Jalandhar, India
E-mail: drvijaydhir@gmail.com

Abstract

Nowadays, sentiment analysis is the most prevalent method adopted by many companies to conduct real-time surveys based on the information available on different social media platforms. Sentiments are internal feelings and emotions of a person towards an entity in the real world. Sentiment analysis is a study about capturing the various sentiment information supplied in natural language writings. In the initial years, most of the research was dedicated to extracting sentiment traits through analyzing lexical and syntactic variables. However, as a vast range of disciplines have begun leveraging the potential of machine learning based algorithms for solving different types of problems, consequently, in the area of sentiment analysis, the idea of machine learning is also not nascent anymore. Many researchers have proved their proficiency in analyzing the opinion of internet users. The current paper proposes a hybrid model to analyze and classify the tweets using a very popular dataset, i.e., Sentiment140 Twitter dataset and the diabetes dataset. The outcome procured from the proposed model showed significantly better results than the other existing machine learning and deep learning based models.

Keywords: Deep Learning, Machine Learning, Sentiment Analysis, Sentiment140 dataset, Twitter, Diabetes Dataset.

1. Introduction

In today's era, users prefer social media platform to express their opinion about products, services, incidents, policies, etc.[1]. These opinions or sentiments on social media hold significant importance for companies and policymakers, as these opinions serve as a base for further decision-making processes[2]. Sentiment analysis is a method for extracting a person's emotions and attitudes about a product, entity, or political or personal issues from web data. This data contains a wealth of information about the user's viewpoints and opinions. The need to analyze the sentiment laden information has been realized in a variety of areas, such as assisting manufacturers in collecting and predicting consumer attitudes toward their products and assisting political organizations in understanding public opinion. For the last few years, people have utilized social media platforms, like Facebook, Twitter and many more, to share their opinion toward specific things. Among such social media platforms, Twitter has acquired the attention of the majority of online users for expressing their attitudes and emotions.

Hitherto, a plethora of studies conducted in the different areas claimed and further proved the efficiency of a single machine learning method for producing reliable results. The utilization of these methods was observed in several domains, such as Internet of things and smart cities, traffic prediction and transportation, cyber security and threat intelligence, Healthcare and COVID-19 pandemic, and for product recommendations[3]–[5]. However, the idea of integrating two or more distinct techniques was then suggested to improve the performance by leveraging the merits of the considered techniques[6]. In the myriad of existing studies, researchers have tested the effect of integrating different machine learning techniques and consequently got improved results. A similar effect was also observed in Convolutional Neural Network (CNN) technique, and the researchers got significantly better results when they combined the above technique with other machine learning methods. The CNN technique has many advantages like it demands less processing time than other machine learning algorithms. Besides this, in case of CNN the necessity of hyper parameters and supervision is minimum. In [7], the authors combined CNN and SVM methods to get a more accurate result for classifying remote sensing images.

On the grounds of the results of the existing studies, the current research has considered the CNN technique and merged it with different machine learning and deep learning methods to probe their efficiency in the arena of sentiment analysis. In the current paper, four hybrid models have been created by combining CNN and Naïve Bayes (NB), CNN and Support Vector Machine (SVM), CNN and Random Forest (RF), CNN and Long term Short term memory (LSTM) and CNN and FCDNN. The results obtained from these hybrid models have also been compared with the existing models.

The main contributions of the research include:

- To introduce a hybrid model in the field of sentiment analysis.
- To compare the proposed model's classification performance with some existing techniques.

The remaining paper is organized as follows:

Section 2 encompasses a review of the studies depicting the current trends in the field of sentiment analysis. Section 3 comprises the methods backing the experiments executed in the current study. Section 4 contains the results procured from the experiments and Section 5 contains the conclusions.

2. Review of Earlier Works

A plethora of studies has been conducted in the sentiment analysis domain. This section outlines the work conducted to analyze the sentiment expressed by the internet users.

Wagh*et al.* suggested an approach for sentiment analysis by combining the natural language toolkit with machine learning. Natural language toolkit is the most prevalent approach for classifying the entities into two or more groups. The aforementioned combination was utilized by the authors to gauge the sentiment's polarity in data available on the social sites. The data for experiment was collected from Twitter [8]. Subsequently, the selected sentiments were classified using Naive Bayes and multinomial classifiers. Finally, the evaluation metrics such as precision, recall, accuracy, and f-measure were used to gauge the classification ability of the developed classifier.[9].

Zhang *et al.* proposed a quantum inspired sentiment representation model which basically prioritized both semantic and sentiment of the word[10]. This model works by extracting phrases that contains adjectives or adverbs. The extraction process was performed through part of speech tagger. The proposed model's performance was evaluated using the Obama-McCain Debate dataset and Sentiment140 Twitter dataset. The proposed model showed significant improvement over other existing vector techniques.

Garcia *et al.* introduced a state of art technique to ameliorate the accuracy of decision support systems. The information processing was automated via the social platform and enhances the accuracy of sentiment analysis support systems [11].

Chidananda*et al.* (2019) proposed a model leveraging the potentials of N-gram and log function to analyze sentiments of Twitter data. The outcome obtained from the proposed model showed significant improvement, in terms of accuracy, as compared to existing models[12].

Batra *et al*. Predicted the stock price movement on the basis of tweets. The main idea was that the emotional state of the people strongly affects stock prices and their movement. The machine learning technique was utilized to extract the feelings associated with the product, person, or organization. The dataset utilized for sentiment detection was obtained from the StockTwits and the yahoo finance market index data. The SVM method is used to generate the sentiment score of the tweets. The SVM classifier assigned optimistic and bearish characterizations to the tweets. Using this categorization, one can forecast stock movement for the next day. The findings from the method showed a positive correlation between people's feelings and market data[13]. In [14], authors utilized NB, SVM, and K-nearest neighbor, for classifying the user views based on their opinion about movie. In [15], the authors proposed a machine learning based model for analyzing the sentiments of users. The tweets utilized for the experiment were in Malayalam language.

All of the above studies have developed the classification models by using a single machine learning algorithm. However, the current study has adopted the notion of hybrid approach where two algorithms are combined to build a final hybrid classification model.

3. Methodology

This section describes different methods and techniques which backed the entire experiment process. Firstly, we describe the dataset and then explain the feature selection and classification techniques utilized during the experiment. And, the last subsection contains the evaluation measures adopted in the current paper to depict the ability of the proposed hybrid model.

3.1 Datasets

The potential of the proposed model has been validated through the most prevalent dataset, i.e., Sentiment140[1]Twitter dataset[16] and diabetes dataset. The sentiment140 dataset was produced by Alec Go*et al*. This dataset has been used in a number of studies conducted in the domain of sentiment analysis. Figure.1 depicts the first five rows of the

[1]https://www.kaggle.com/datasets/kazanova/sentiment140

Sentiment140 dataset. The Sentiment140 dataset contains around 1.6 million tweets and for the current study, we randomly chose 1620 positive and 1600 negative tweets [17][18].

The diabetes dataset has been created by extracting the tweets related to diabetes from Twitter using Apache Flume. These tweets were stored in Hadoop Distributed File System (HDFS). Figure 2 shows the extracted tweets along with their IDs [19][20].

```
In [ ]:  1  dataset.head()
Out[11]:
```

	target	ids	date	flag	user	text
0	0	2327192646	Thu Jun 25 08:02:13 PDT 2009	NO_QUERY	quiz_master	Was having dinner with parents downstairs in D...
1	0	2327193206	Thu Jun 25 08:02:16 PDT 2009	NO_QUERY	djcampos	Blah 5am still up daang I got deep problems
2	0	2327193455	Thu Jun 25 08:02:17 PDT 2009	NO_QUERY	RKF	@jenspeedy I would suggest avoiding 360 Living...
3	0	2327193641	Thu Jun 25 08:02:18 PDT 2009	NO_QUERY	AnaHertz	@alexbroun I didn't convince myself I was fat ...
4	0	2327193806	Thu Jun 25 08:02:18 PDT 2009	NO_QUERY	yenafer	@spotzle @jstarrh check on sunscreen, snacks, ...

Figure 1 : First five rows of the Sentiment140 dataset using .head() function

```
Administrator: Command Prompt - hive
hive> select * from diabetes_data_twitter;
OK
1363556299254988804  RT @incmnszmx: Si tienes #diabetes evita las hipoglucemias haciendo comidas peque?as y frecuentes, consume una colaci?n antes de dormir. ToÓ
1363556322881335297  RT @WrittenByHanna: No but Nick Jonas really had me thinking diabetes was gonna take him out at any second when I was in middle school ??
1363557030179581953  RT @blah_baa: @BorisJohnson @Jeremy_Hunt @MattHancock @BBCNews #Ridge @ChangiAirport @ZeroCovid_UK @ZeroCovAlliance @Parents_Utd @UKActiontÓ
1363560783620276233  The resverstrol in red wine is one of the polyohenols that is good for your gut and weight-loss#diabetes #foodÓ https://t.co/rJqHoN7fMJ
1363560822149050368  Agree with diabetes being a common cause of CKD but not with HTN...it is probably vastly overrated as an etiology oÓ https://t.co/XjDUGzE2uX
1363560852318793729  Beans are powerhouse of fiber/protein/complex carbs/phytonutrients..very satisfying and still low in calories..greaÓ https://t.co/b8n6YYSpgv
1363557063809400833  #diabetes symptoms diabetes type 2 https://t.co/EoTZ6m9rhP type 1 diabetes 31:52
1363557112366981129  RT @Fact: Giving up alcohol for just a month can improve liver function, decrease blood pressure and reduce your risk of liver disease andÓ
1363557132759629831  never seen a man so dramatic
1363557166054055939  So is diabetes now racist too?  Or do we have to wait for the news to catch up?
1363557182462132229  RT @American_Heart: February is American Heart Month. There is no better time to evaluate Cardiovascular risk in patients with T2D. DownloaÓ
1363557255279476736  RT @Jazzymykai__: Broooo i stg
1363557260467834880  RT @finite_alright: cool. remember when coca-cola exasperated a water shortage by extracting more than 300,000 gallons of water per day forÓ
1363557451782623238  We hope that this work will allow us to improve patient care and also to simplify certain screening tests Ó https://t.co/x1Q1RTVHwc
1363557295456727041  BorisJohnson I would like to see a road map for things such as: Anorexia &amp as well as obesity and diabetesÓ https://t.co/efGlggmtlr
1363556344876462087  Top Selling Diabetes Care:OneTouch, Accu-Chek, FreeStyle and morehttps://t.co/7x7j5nN1Wi #ebaysellerÓ https://t.co/ycGFOjIag8
1363557451782623238  We hope that this work will allow us to improve patient care and also to simplify certain screening tests Ó https://t.co/x1Q1RTVHwc
1363557468991873028  RT @parthaskar: And here is the update from @NHSDigital #GestationalDiabetes #Shielding Thank you from @NHSDiabetesProg to @DHSCgovuk @NiÓ
1363557491347398664  @7VENKMEN omg same here! my grandpa had diabetes, cholesterol and heart problems too?? same here feel free to sendÓ https://t.co/9xPplN1Mx8
1363557612013387776  RT @LEAD_Coalition: Prediabetes linked to worse brain health https://t.co/NYW1de7aoy #diabetes #Alzheimers #dementia @LindaLeeKing @CycliÓ
1363557612013387776  RT @LEAD_Coalition: Prediabetes linked to worse brain health https://t.co/NYW1de7aoy #diabetes #Alzheimers #dementia @LindaLeeKing @CycliÓ
1363557635602198529  Iâd cry my eyes out during his a little bit longer speech ??
1363557643495874563  @irsievan @Rad_Demo Mental psychosis diabetes
1363557666258358275  Not to be an agent of doom but all I see in this video is Diabetes
1363557670876241922  ¶Banting sold the patent rights for insulin to The University of Toronto for $1, claiming that the discovery belongÓ https://t.co/ChmTc5aZVn
1363557708482322434  Cant finish them all in one sitting. Sobra yung sugar baka may diabetes na ko bukas. But I love them all kasi lahatÓ https://t.co/1l1hgub08Y
1363556351725592577  RT @GHargraveWrites: @michelleptweets @WrittenByHanna The fact that she hid her diabetes from Luca in the movie to the point that she almosÓ
1363557761045303297  #diabetes @AmDiabetesAssn @DiabetesUK
1363557791991033858  Hi yãall- keep on telling the hubby to lose a few pounds- doctors say his BMI is dangerously close to being in theÓ https://t.co/onpbMG4zis
1363557805240770560  RT @OlveiraPaco: packaging infames El tipo se ocup? de los programas para pacientes de Diabetes y HIV De la prevenci?n del C?ncer G?nito-MaÓ
1363557843165732877  Another part of a heart healthy lifestyle that can help you manage your diabetes is to focus on what youâre eating.Ó https://t.co/3e1SmhJeFl
1363557871401766912  Nevergetsick World's Greatest Healing Miracle of All Time÷100% Scientifically Proven to Cure Cancer, Diabetes TyÓ https://t.co/mzajYETdqO
1363557880620793862  Diabetes is tough @wedsgp1
1363557914892509186  RT @MarcLobliner: .@CocaCola helps create an obesity and diabetes epidemic in black communities then rallies against being white.Sounds aÓ
1363557930168176640  Poor countries like ours who have richer countries crappy food foisted on us see the highest growth rates of NCDs lÓ https://t.co/MckoNtvJkK
1363557940972634112  @CancerWarrior8 Oh, I know she is going to start screaming at people in a few hours. The second someone brings up diabetes.
1363557970882289666  Hi Darren, youâre speaking to a young adult who got covid just before the first lockdown and now has diabetes becaÓ https://t.co/Y9noHgp4Lh
1363556434080849927  @robkhenderson All of this coincides with the rise of heart disease, diabetes, and obesity...this sounds like sometÓ https://t.co/sqQnMUjm7U
1363557986812231680  RT @thewayoftheid: Thereâs a character named Suga Feet bc ppl get diabetes every time he hits the floor
1363558012267294721  @WrittenByHanna FR? I was like ¶my baby has diabetes?÷ ??
```

Figure 2: Tweets extracted through Apache Flume

3.2 Pre-processing and feature selection technique

The pre-processing of the extracted tweets is a necessary task before starting the actual classification process. In the current paper, we applied a set of pre-processing tasks, such as removal of stopwords, removal of punctuation, stemming, and lemmatization. The pre-processing starts with tokenization, where the sentences were split up into words. The stopword feature discovers and then eliminates the commonly used words, such as “a”, “an”, “the”, “we” and many more, which do not contribute significantly to sentence comprehension. Stemming and lemmatization were utilized to produce the root forms of the words obtained so far [21][22].

In order to identify the opinion, it is necessary to extract only those relevant features from the sentence, which contribute significantly to shed off the true opinion of the internet user. In this paper, feature extraction has been performed by using a prevalent method, i.e. Term Frequency - Inverse Document Frequency (TF-IDF). This method determines the need of a specific term in a given document. Jones *et al*. pointed out that the terms which frequently occur in a sentence, but rarely occur in the entire document are more informative. The TF-IDF method can be formulated as below [23][24].

$$tfidf(t,d,D) = tf(t,d) * idf(d,D)$$

Where $$tf(t,d) = 1 + logf_{t,d}$$

$$idf(d,D) = log \frac{|D|}{\{d \in D : t \in D\}}$$

In the above formula, *tf(t,d)* indicates how often term *t*occurred in the whole document d. i*df(d,D)*helps identify terms observed in only a handful of documents. Therefore, with the help of *tfidf(t,d,D)*, it is possible to identify the terms which were observed in a small number of documents, but their occurrence in those documents was frequent. Therefore, it helps to identify the terms observed in a few documents, but at the same time, their occurrence in those documents was too high [25][26].

3.3 Sentiment classification

The following subsections contain the outline of machine learning techniques considered in the current paper to develop the classification model [27].

3.3.1 Naive Bayes

Naïve Bayes relies on the Bayes theorem of probability analysis. Naïve Bayes scheme calculates the posterior probability of reaching a final decision output. First, the frequency table and likelihood table are calculated in the large training data table. Then, after prediction is made for different categories based on various attributes. The main applications of Naïve Bayes include text classification, spam filtering, etc. Naive Bayes has also been adopted in a number of existing studies to calculate sentiment analysis [28].

3.3.2 Support vector machine

The SVM model works by locating the decision surface at an appropriate position that classifies the points of two or more heterogeneous groups. This model is preferably used for n-dimensional data sets. Each input data set is classified based on values of features and put into one of the groups made from the training data set. SVM segregates n-dimensional space into groups using a hyperplane. The hyperplane specifies the decision boundary between groups of data elements. Each group contains dataset elements having high similarity. The hyperplane creates segregation between two sets of data elements. Each group has extreme points which define the edge of that group. The hyperplane is optimal if there is a maximum difference between the hyperplane and extreme points. After the machine is constructed, any new data value should be classified into one of the groups. The SVM approach has a number of applications in the real world, such as face recognition, text categorization, and image classification etc [29].

3.3.3 Random Forest

In the family of supervised learning techniques, the random forest algorithm comes out as an important and result-oriented algorithm. It helps to solve problems of classification and regression analysis. On the ground of information provided in the existing studies about random forest, we have divided the overall process into three steps. In the first step, subsets of data sets are randomly selected to decision trees. In the next i.e. second step, different decision trees are built. After that, in the last step, majority voting is performed to form the final combined result, which is then used to predict the final outcome. The random forest technique is suitable for binary classification, multinomial classification, and regression purposes.

3.3.4 Deep learning techniques

LSTM and CNN are two deep learning methods used in the current paper for sentiment analysis. The LTSM method can reserve the content both for a longer or shorter period. In addition, the LSTM method possesses advanced memory compared to its predecessor, i.e. RNN, which suffered from the vanishing gradient problem.

The first application of CNN took place in the domain of computer vision. The CNN method has emerged as an efficient method for recognizing images. The working of this method is motivated by the visual cortex, an essential part of the human brain. The main reason behind the widespread adoption of the aforementioned method is considered to be its ability to produce good results even in the case of a lower number of neurons. If the same work needs to be done through the predecessor of CNN, such as ANN, the same requires a lot of neurons. CNN method consists of convolutional layers, pooling layers, and fully connected layers.

3.3.5 Hybrid model

The proposed approach uses a hybridized form of the techniques at the level of features to obtain semantic and linguistic information. At the classifier level, CNN has been hybridized with SVM, NB, and other classifiers because CNN improves feature mapping in non-linear space and machine learning enhances the learning performance of classifiers [7]. Semantic and linguistic features are hybridized to improve sentiment domain learning and reduce feature noise using a hybridized classifier strategy such as CNN and the machine learning approach.

Figure3 contains the visualization of the whole sentiment analysis process. The first step was the pre-processing of the tweets and this involves the application of different methods, such as removal of stopword, removal of punctuation, stemming, and lemmatization. After pre-processing, the next task was to identify the most important words which can contribute significantly to accurately represent the actual meaning of the sentence. This process was executed by TF-IDF. Then, the features extracted from the previous steps were further given as input to the hybrid classifier.

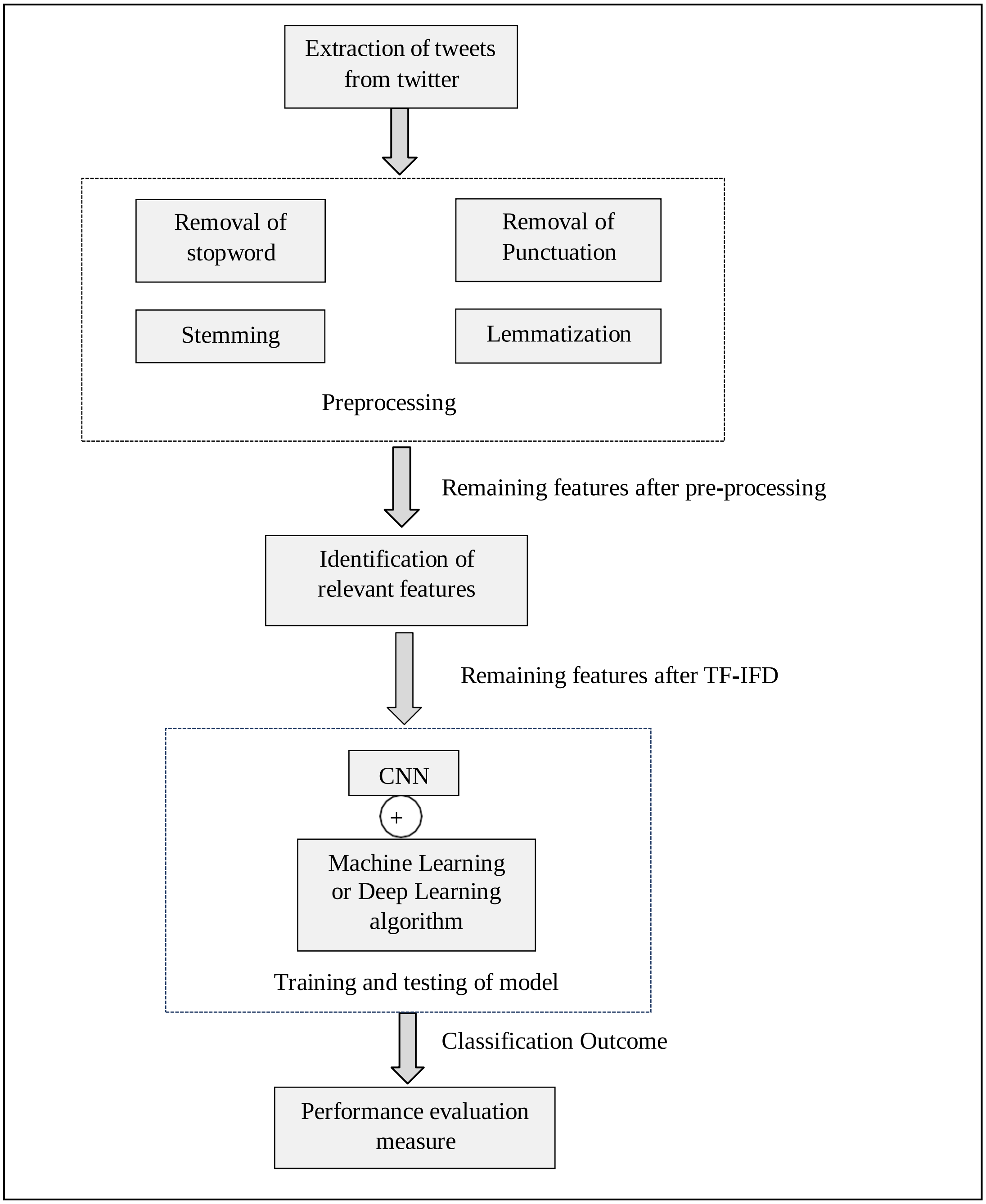

Figure 3: Flowchart of Proposed Hybrid Sentiment Representation (HSR) Model

These hybridized models are formed by combining CNN with Naïve Bayes, CNN with SVM, CNN with RF, and CNN with LSTM. Basically, the hybridization of different techniques enhances learning weights and makes classifier models more accurate. After this, the testing datasets were used by the developed hybrid model. At last, the procured classification results were expressed through the selected performance evaluation measures.

The following subsections describe the sequence followed during the creation of hybrid models.

In *hybrid Naïve Bayes*, the hybrid model incorporates two techniques CNN and Naïve Bayes. The model is created by combing the said methods in following manner.

CNN →Naïve Bayes

The CNN method extracts the key features from each sentence, whereas the Naive Bayes method can classify the sentence into one of the predefined category. The hybrid model constructed using CNN-Naïve Bayes method is explained in the Figure4. As mentioned previously, the relevant features are first collected from each sentence by using TF-IDF method. These vectors are then entered into our hybrid model. The Word2Vec method is utilized to create vectors out of the relevant words. The output of the Word2Vec is entered as the input to the first layer of CNN model. The first layer, convolutional layer has 256 filters. The next layer, max pooling layer, contributes to minimize the number of parameters by selecting the words seems appropriate among those produced by the filters.

The next layer, fully connected layer, ensures that every input of the input vector influences every output of the output vector. This layer sends the features extracted from the max pooling layer in a fully connected way to the sigmoid method. The sigmoid method is utilized in case of CNN model for the classification purpose. The variant of sigmoid method is softmax method, which is utilized when features are mapped to one of the three output classes, whereas the former one is preferred when the output classes are two. In our case, the sigmoid method can be considered for classification as we are mapping the sentiments into either of two classes, i.e., positive or negative. However, in the hybrid Naïve Bayes model, the sigmoid method is replaced by Naïve Bayes.

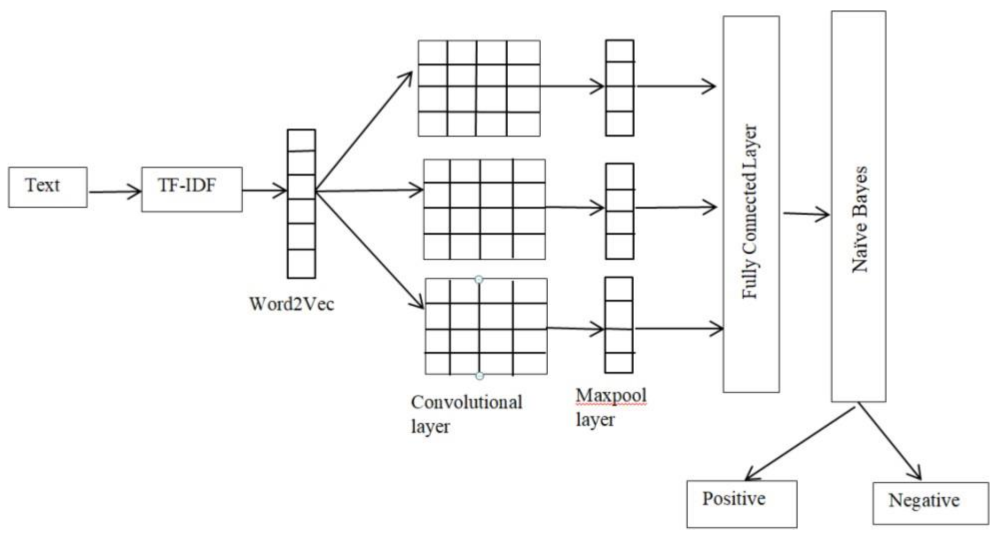

Figure 4: Hybrid Naïve Bayes model

In hybrid support vector machine, the hybrid model is constructed by following the following sequence

CNN →SVM

The hybrid SVM model is explained in Figure5. Here also, the CNN method is used to extract the features which can contribute significantly in revealing the true emotion of the user. This hybrid model incorporates CNN layer, pooling layer and fully connected layer as describe in the hybrid Naïve Bayes model. However in this model, the classification part is being performed by support vector machine.

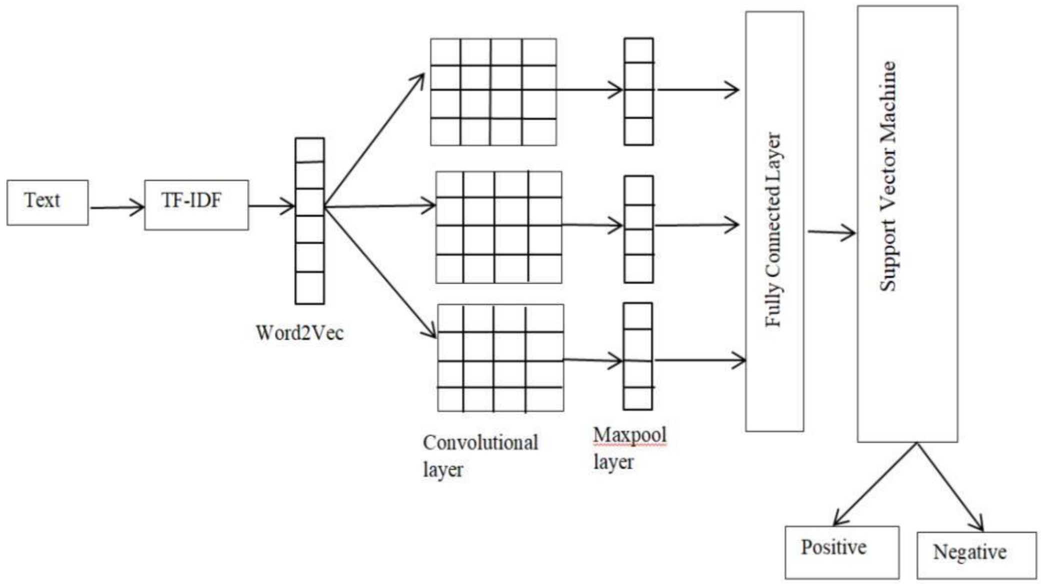

Figure 5: Hybrid Support Vector Machine Model

In hybrid random forest model, the model is arranged in the following way

CNN →RF

Figure 6 contains the detail design of hybrid random forest model. This model is nearly similar to the previous one with slight variation, i.e. in this model the classification is done with random forest method.

In hybrid LSTM model, the following order was followed for its construction

CNN→LSTM

In Hybrid LSTM, the hybrid model is created by combining CNN with LSTM techniques as shown in Figure 7. Both of these are deep learning techniques. The techniques are combined in the following sequence. The key objective behind the integration of the CNN and LSTM is that the former technique is proficient in identifying short term dependencies, whereas the latter is efficient in remembering the long term dependencies. The real significance of the LSTM method can be noticed in the instance when the length of the tweet is very long. The main architecture of the CNN and LSTM model is depicted in Figure. The outcome retrieved from the convolutional layers is given as input to the max pooling layer. The features produced by the max pooling layer is then entered into LSTM, which incorporates three gates, forget update, and output. The words outsourced from LSTM are goes into sigmoid function for final classification.

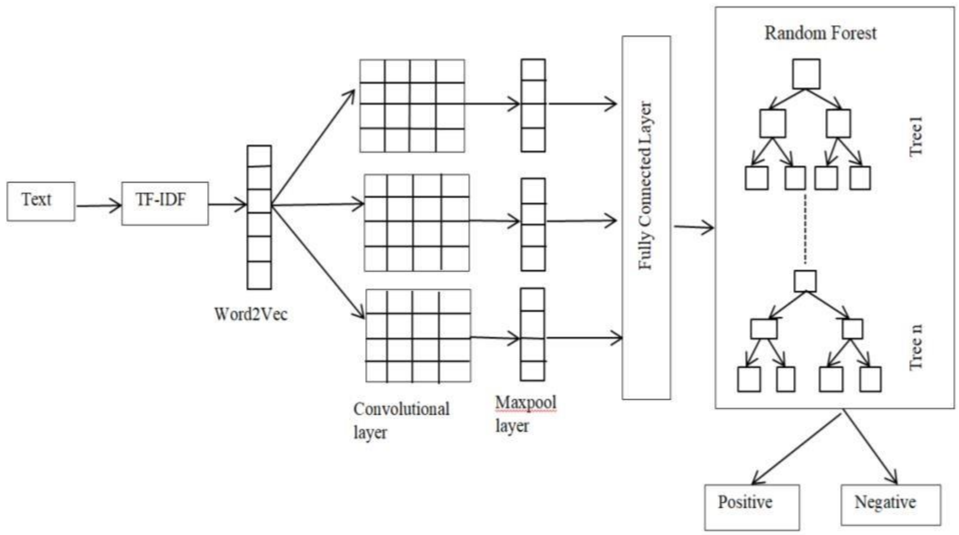

Figure 6: Hybrid Random Forest model

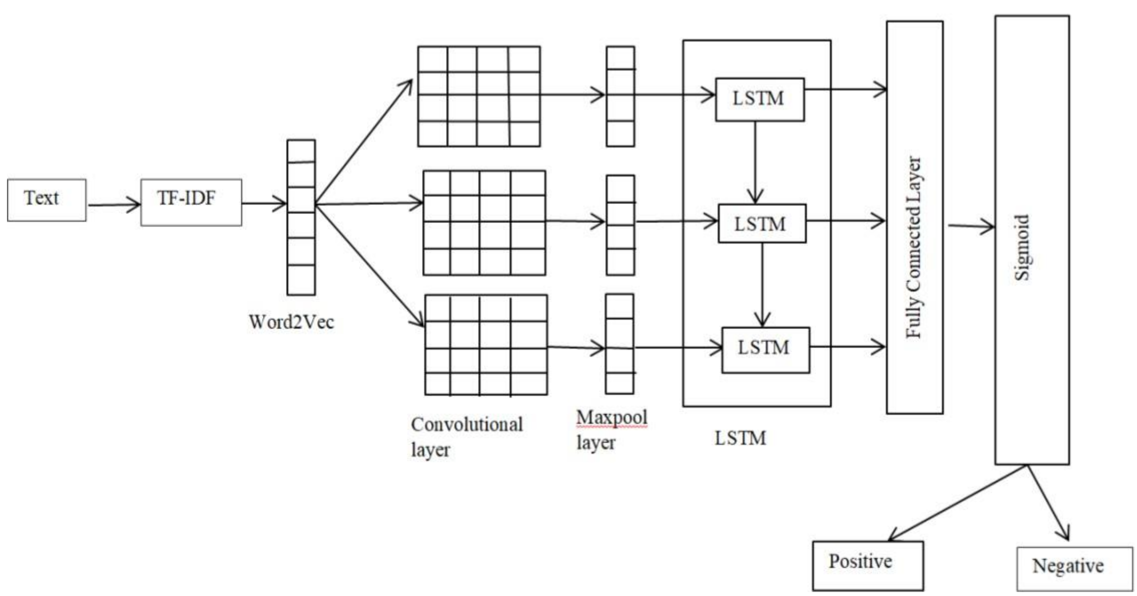

Figure 7: Hybrid LSTM model

In *hybrid FCDNN* model, the following order was followed for its construction

CNN→LSTM →SVM

Figure 8 contains the model constructed using CNN, LSTM, and SVM. In this model, we have utilized three instead of two techniques. The fully connected layer is the part of CNN layer. We have already included this layer in the previous models as a part of CNN technique. The output from the LSTM is entered in fully connected way into SVM method. The initial steps are same as that described in the hybrid LSTM model, however, instead of feeding output into sigmoid function, we have used the SVM method.

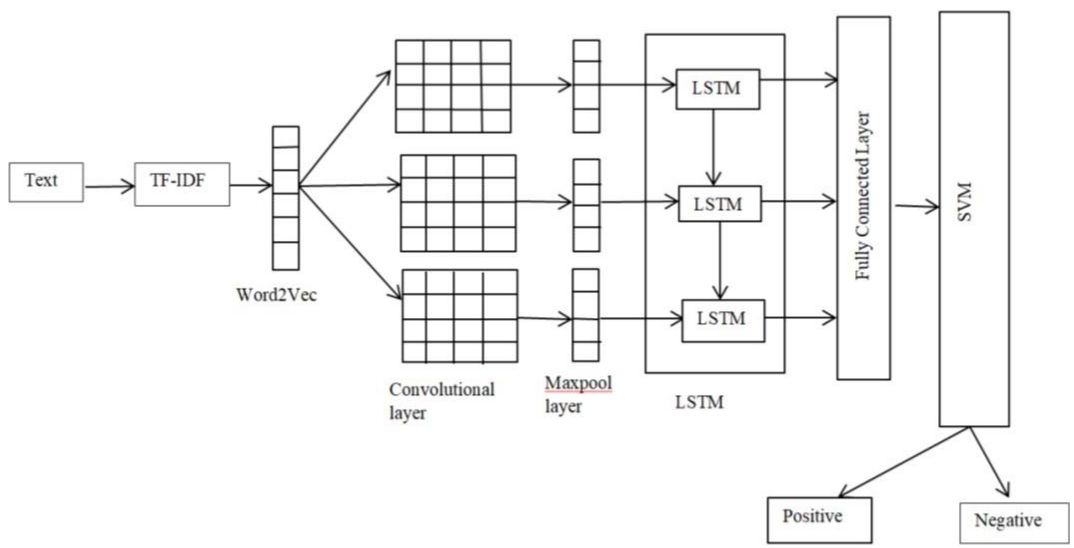

Figure 8: Hybrid FCDNN model

3.4 Performance evaluation metrics

To gauge the effectiveness of the proposed hybrid technique, it is necessary to select the appropriate performance evaluation method. The selected performance evaluation measures are accuracy, precision, recall, F1-score. However, calculating the selected performance metrics require the determination of the confusion matrix's indices. The confusion matrix is shown in Table 1. In addition, terms depicted in the confusion matrix are further helpful in computing the values of the performance measures chosen in the study.

Accuracy calculates the number of cases in which the classification model correctly puts the review under its correct category. The formula used to compute the Accuracy measure is given in Table 2.

Precision: The prediction index measures the number of correctly predicting positive reviews among the reviews classified as positive. The formula used to compute the precision measure is given in Table 2.

Recall: Recall calculates the number of cases in which the developed model correctly put the negative cases in its correct category. The formula used to compute the recall measure is given in Table 2.

F1-score: The F1-score is a metric utilized to determine the correctness of a test. This metric is computed using the accuracy and recall. The formula used to compute the F1-score measure is given in Table 2.

Table 1: Confusion matrix

		Actual Value	
		Positive	**Negative**
Predicted Value	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 2: Performance evaluation measures considered in the current study

Performance measure	Formulae
Accuracy	$\frac{TP + TN}{TP + FP + FN + TN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F1-score	$2 * \frac{precision * recall}{precision + recall}$

4. Results

This section encompasses the results obtained during the experiment. It is pertinent to mention that all of the experiments were implemented in Python programming language. As already mentioned in the previous section, Sentiment140 Twitter dataset has been frequently utilized in a number of research studies to test the efficiency of various sentiment analysis methods. The key advantages of the considered dataset are its conciseness and comprehensibility. In the current paper also, the same dataset has been utilized to test the proficiency of the developed hybrid model. Apart from this, the diabetes dataset has been also used to perform sentiment analysis by employing same hybrid model.

Hybrid Naive Bayes

As already mentioned the CNN model has been combined with different machine learning and deep learning approaches in order to generate a more efficient classifier. In hybrid Naïve Bayes model, CNN was combined with Naïve Bayes algorithm. On the execution of the hybrid Naive Bayes model, the accuracy obtained on the training and testing datasets was around 87 % and 76 % respectively.

Figure 9 depicts displays the accuracy, precision, recall, and F1- score and confusion matrix of the hybrid Naïve Bayes classifier:

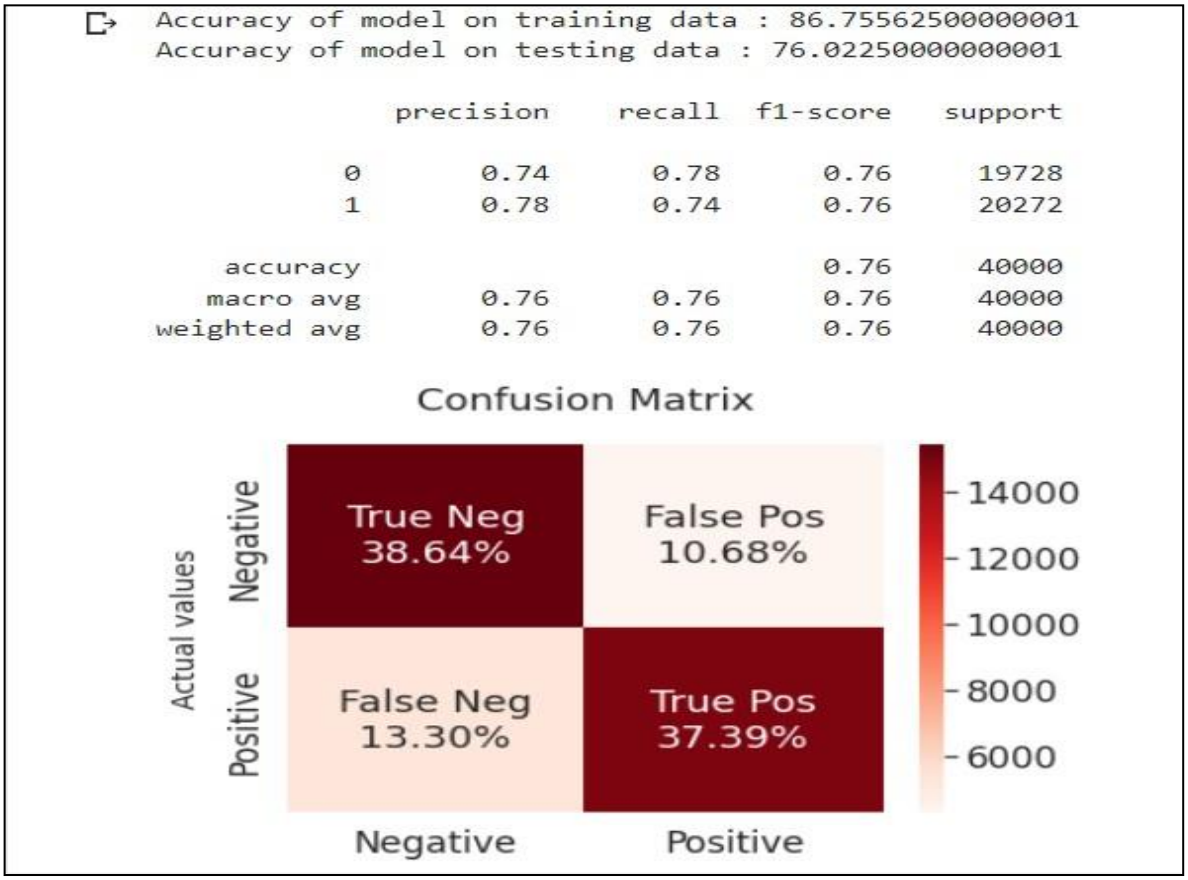

Figure 9: Accuracy results for Hybrid Naive Bayes

Hybrid SVM

Hybrid SVM was formed by using CNN approach followed by SVM. During the experiment, Hybrid SVM gives an accuracy of about 93 % on the training data and about 76 % on the testing data.

Figure 10 contains code displays the accuracy, precision, recall, F1- score, and confusion matrix of the Hybrid SVM classifier.

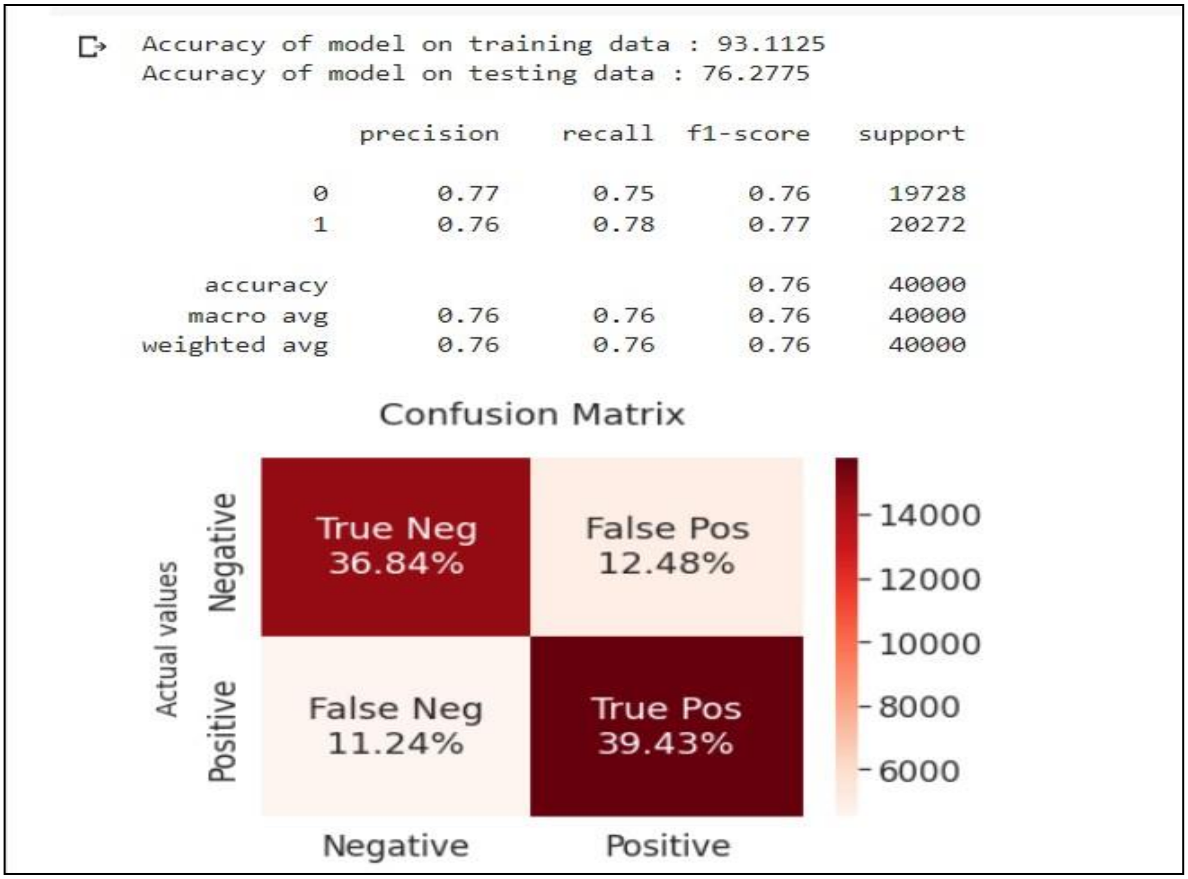

Figure10: Accuracy results for Hybrid SVM

Hybrid Random Forest

The Hybrid Random forest model is created by merging CNN and RF algorithms. This model gives an accuracy of 75 % on the training data and around 71 % on the testing data.

Figure 11 displays the results obtained for accuracy, precision, recall, F1- score, and confusion matrix on the execution of the Hybrid Random Forest classifier.

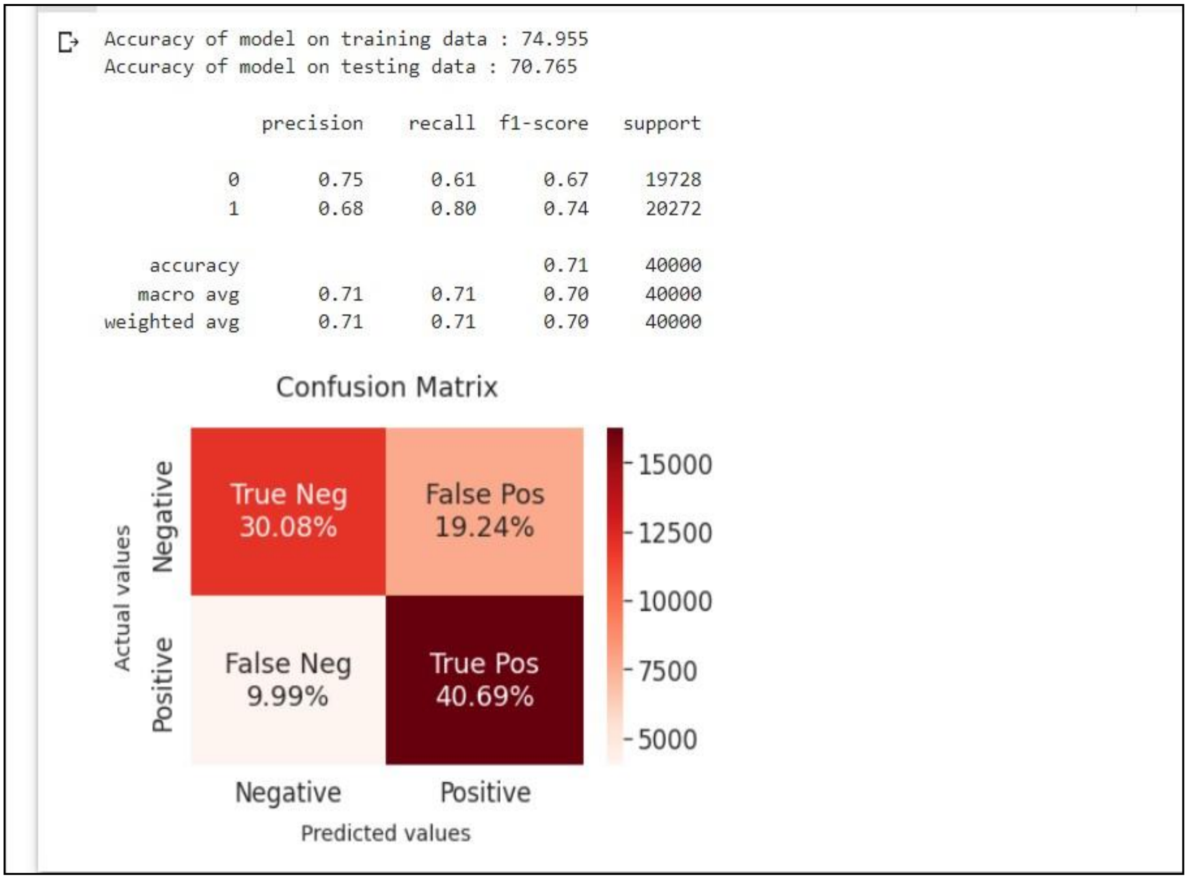

Figure11: Accuracy results for Hybrid Random Forest

Hybrid LSTM

The Hybrid LSTM was formed by combining CNN with LSTM. As a result, the Hybrid LSTM model gives an accuracy of around 83 % and performs far better than the traditional machine learning models mentioned above [17]–[19]. Figure 12 displays the accuracy results of the Hybrid LSTM after one epoch.

```
model.summary()

Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 32, 100)           10000000

 lstm (LSTM)                 (None, 64)                42240

 dense_6 (Dense)             (None, 64)                4160

 dense_7 (Dense)             (None, 1)                 65

=================================================================
Total params: 10,046,465
Trainable params: 10,046,465
Non-trainable params: 0
_________________________________________________________________

[80] num_epochs = 1
     history = model.fit(X_train, y_train,
                         epochs = num_epochs, validation_data = [X_test,y_test],
                         verbose = 1)

40000/40000 [==============================] - 2934s 73ms/step - loss: 0.4049 - accuracy: 0.8153 - val_loss: 0.3781 - val_accuracy: 0.8293

[81] model.evaluate(X_test, y_test)

10000/10000 [==============================] - 59s 6ms/step - loss: 0.3781 - accuracy: 0.8293
[0.37812715768814087, 0.829337477684021]

✓ 58s   completed at 7:26 PM
```

Figure 12: Model summary and accuracy results for Hybrid LSTM

Hybrid FCDNN

The Hybrid FCDNN model was built using CNN and FCDNN techniques. The model summary and accuracy result based on the Hybrid FCDNN model are outlined in Figure 13 and Figure 14respectively. This model gives an accuracy of about 82 %.

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 30)]              0

 embedding (Embedding)       (None, 30, 300)           73374900

 spatial_dropout1d (SpatialD  (None, 30, 300)          0
 ropout1D)

 conv1d (Conv1D)             (None, 26, 64)            96064

 bidirectional (Bidirectiona  (None, 128)              66048
 l)

 dense (Dense)               (None, 512)               66048

 dense_1 (Dense)             (None, 1)                 513

=================================================================
Total params: 73,603,573
Trainable params: 228,673
Non-trainable params: 73,374,900
_________________________________________________________________
None
```

Figure 13: Model summary for Hybrid FCDNN

```
Epoch 12/15
625/625 [==============================] - 818s 1s/step - loss: 0.3904 - accuracy: 0.8224 - val_loss: 0.3923 - val_accuracy: 0.
8228
Epoch 13/15
625/625 [==============================] - 815s 1s/step - loss: 0.3900 - accuracy: 0.8228 - val_loss: 0.3922 - val_accuracy: 0.
8235
Epoch 14/15
625/625 [==============================] - 809s 1s/step - loss: 0.3891 - accuracy: 0.8228 - val_loss: 0.3944 - val_accuracy: 0.
8228
Epoch 15/15
625/625 [==============================] - 809s 1s/step - loss: 0.3887 - accuracy: 0.8235 - val_loss: 0.3918 - val_accuracy: 0.
8230
```

Figure 14: Accuracy results for Hybrid FCDNN after 15 epochs:

Zhang et al. proposed a quantum inspired sentiment representation model and compared its outcome with existing techniques such as SentiWordNet, Unigram, Bigram, Trigram, Doc2vector, PMI-IR, SentiStrength, and Simple addition [10]. The whole experiment was executed on Sentiment140 Twitter dataset. The results obtained by previous mentioned study along with the outcome yielded by our proposed hybrid models are shown in Table 3. In case of the NB method, all of the algorithm's results are close, except SentiWordNet, which only obtains an accuracy of 0.511. Based on the results, it can be observed that counting positive and negative words alone are insufficient to determine sentiments accurately.

Furthermore, the PMI-IR approach has depicted the lowest accuracy. Conversely, the proposed hybrid model has shown the highest accuracy. The best precision score goes to Trigram, while the highest recall and F1 values go to doc2vector. The outputs of the unigram, bigram, and trigram algorithms are nearly identical in terms of accuracy. In case of SVM classifier, we discover that the trigram method performs better. The RF classifier, on the other hand, does not produce satisfactory results. In case of RF classifier, Unigram produces superior classification outcome. As a result, it can be concluded that the performance of language models may be dependent on the classifiers. Furthermore, SentiWordNet and PMI-IR approach produced poorest outcome as compared to other algorithms. Accuracy, precision, recall, and F1-score based results obtained in the case of different machine learning methods are depicted in Figure 15, Figure 16, Figure 17, and Figure 18, respectively.

Table 3: Result of the proposed model and other existing techniques

Classifier	Algorithm	Accuracy	Precision	Recall	F1
NB	SentiWordNet	0.5111	0.5132	0.5149	0.5144
NB	Unigram	0.5581	0.5626	0.5539	0.5567
NB	Bigram	0.5564	0.5680	0.5561	0.5623
NB	Trigram	0.5546	0.6009	0.5326	0.5762
NB	Doc2vector	0.5548	0.5446	0.6483	0.5904
NB	PMI-IR	0.5205	0.5268	0.5387	0.5327
NB	SentiStrength	0.5436	0.5486	0.5444	0.5469
NB	Simple addition	0.5610	0.5771	0.5633	0.5715
NB	QSR model	0.5669	0.5698	0.5672	0.5684
NB	**Hybrid model**	**0.7603**	**0.756**	**0.754**	**0.7533**
SVM	SentiWordNet	0.5111	0.5132	0.5149	0.5144
SVM	Unigram	0.5548	0.5569	0.5600	0.5576
SVM	Bigram	0.5696	0.5630	0.5774	0.5679
SVM	Trigram	0.5728	0.5698	0.5752	0.5721
SVM	Doc2vector	0.5614	0.5600	0.6323	0.5939
SVM	PMI-IR	0.5205	0.5268	0.5387	0.5327
SVM	SentiStrength	0.5436	0.5486	0.5444	0.5469
SVM	Simple addition	0.5714	0.5765	0.5753	0.5761
SVM	QSR model	0.6567	0.6492	0.6548	0.6526
SVM	**Hybrid model**	**0.7627**	**0.768**	**0.7743**	**0.7833**
RF	SentiWordNet	0.5111	0.5132	0.5149	0.5144
RF	Unigram	0.5761	0.5815	0.5871	0.5842
RF	Bigram	0.5761	0.5780	0.6097	0.5934
RF	Trigram	0.5516	0.5577	0.5613	0.5595
RF	Doc2vector	0.5548	0.5565	0.6032	0.5789
RF	PMI-IR	0.5205	0.5268	0.5387	0.5327
RF	SentiStrength	0.5436	0.5486	0.5444	0.5469
RF	Simple addition	0.5754	0.5782	0.5818	0.5796
RF	QSR model	0.6283	0.6204	0.6678	0.6432
RF	**Hybrid model**	**0.7077**	**0.72**	**0.722**	**0.71**
LSTM	standard LSTM	0.6813	0.6772	0.6839	0.6811
LSTM	AT-LSTM	0.6929	0.6968	0.6922	0.6948
LSTM	**Hybrid model**	**0.8293**	**0.81**	**0.8**	**0.81**
FCDNN	FCDNN	0.5667	0.57	0.5625	.5654
FCDNN	FCDNN-QSR	0.5917	0.5886	0.5905	0.59
FCDNN	**Hybrid FCDNN**	**0.823**	**0.822**	**0.84**	**0.82**

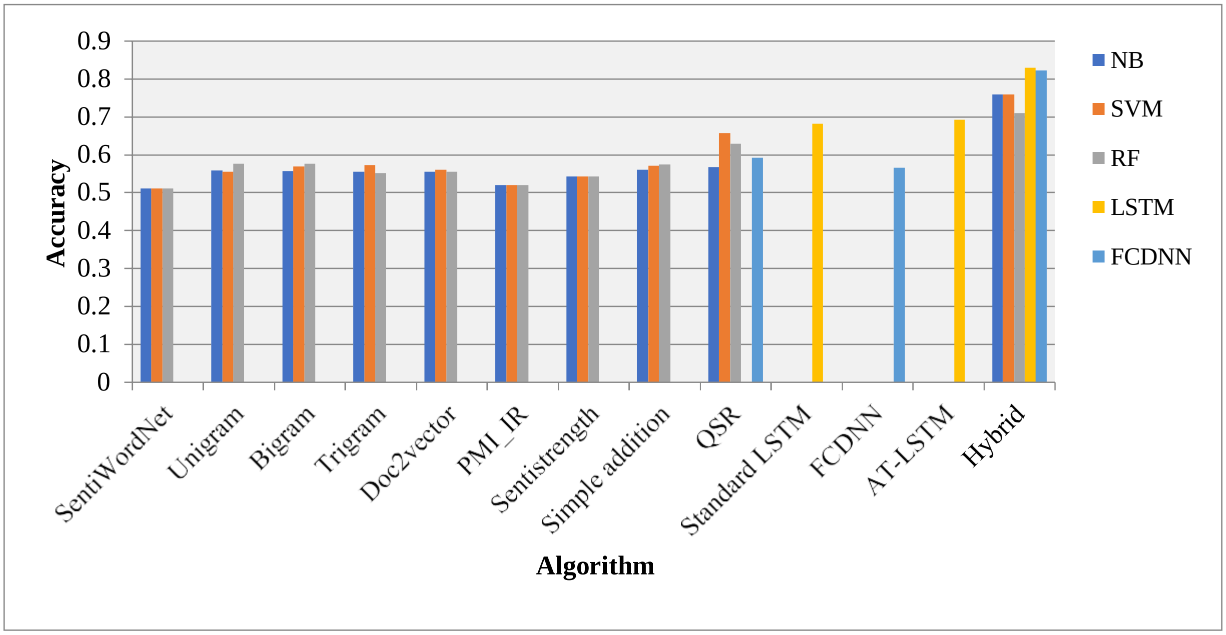

Figure 15: Accuracy results on Sentiment140 dataset

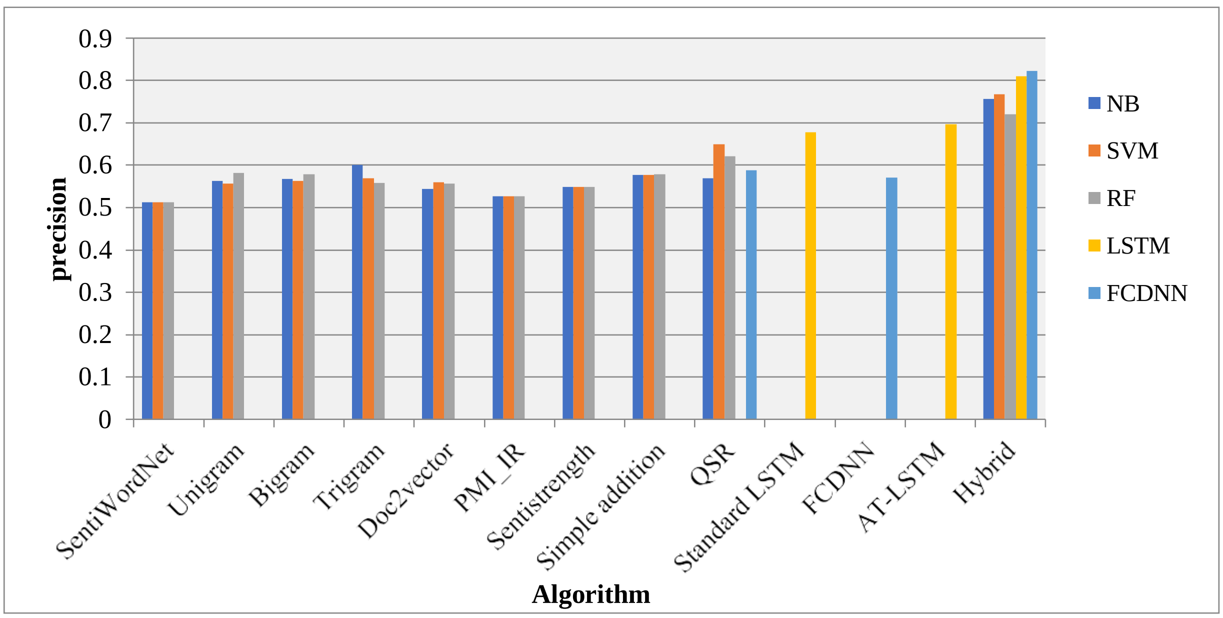

Figure 16: Precision results on Sentiment140 dataset

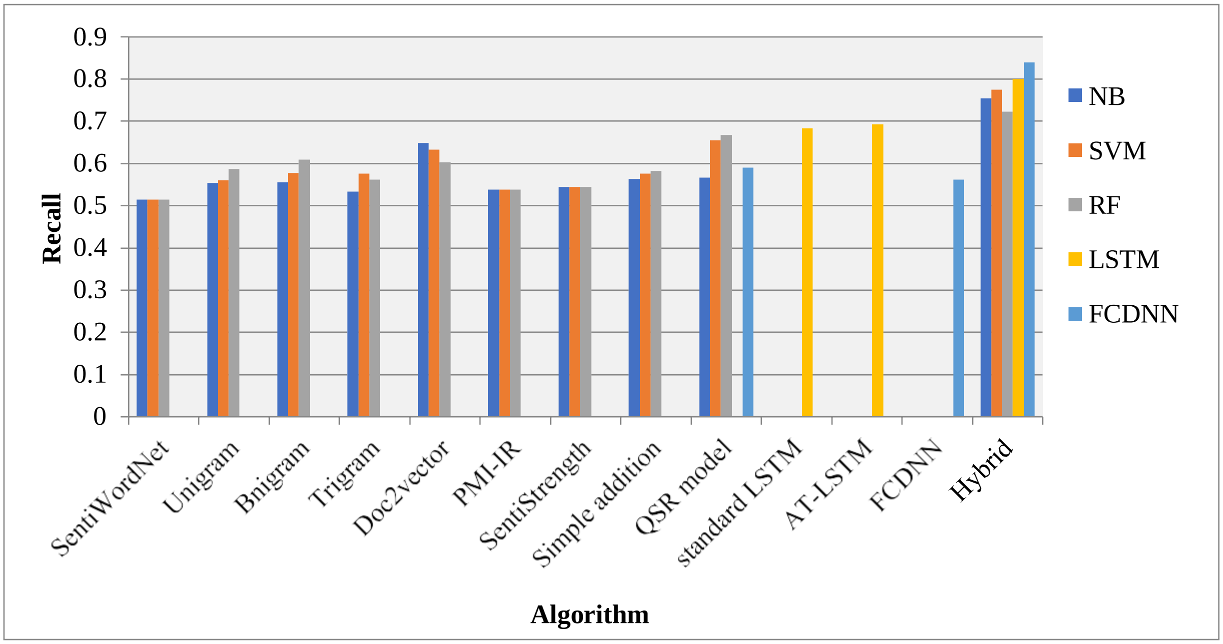

Figure 17: Recall results on Sentiment140 dataset

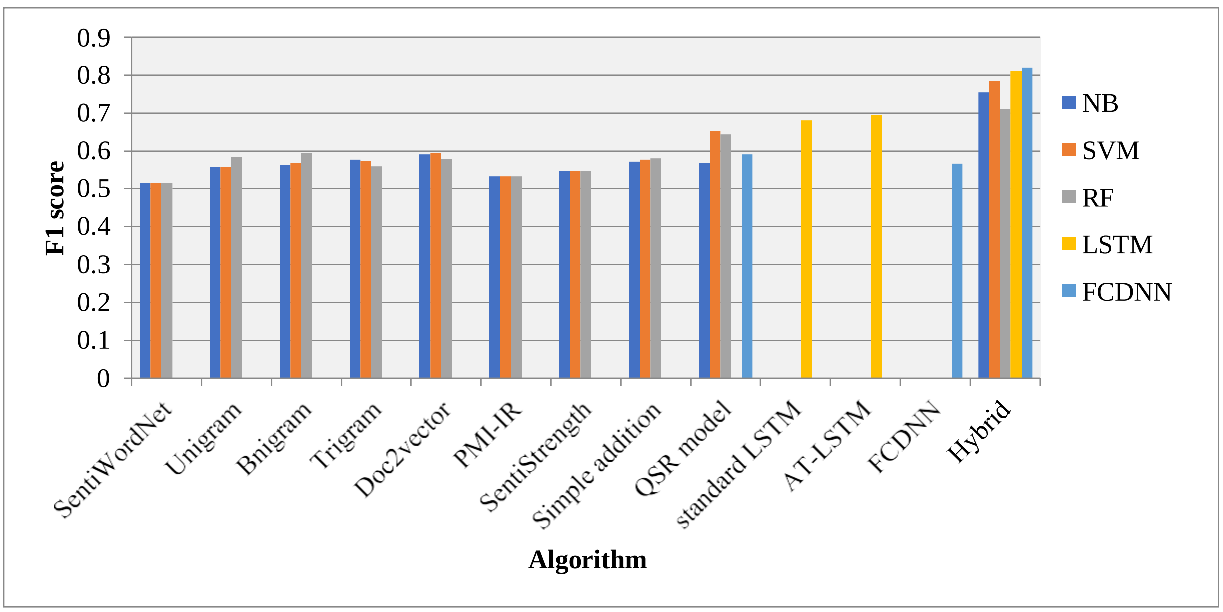

Figure 18: F1-score results on Sentiment140 dataset

The performance of proposed model on diabetes dataset

This section contains the results when hybrid model was applied on the diabetes dataset. The proposed hybrid model is compared with the models developed using existing machine learning and deep learning techniques. Results of the hybrid model while using diabetes dataset are shown in Table 4.

Table 4 Result of the proposed hybrid model on diabetes dataset

Technique	**Accuracy**	**Precision**	**Recall**	**F1**
NB	0.599	0.587	0.602	0.594
Hybrid NB	**0.788**	**0.759**	**0.825**	**0.791**
SVM	0.659	0.651	0.663	0.657
Hybrid SVM	**0.779**	**0.782**	**0.787**	**0.784**
RF	0.660	0.644	0.677	0.600
Hybrid RF	**0.737**	**0.701**	**0.828**	**0.759**
LSTM	0.697	0.706	0.741	0.723
Hybrid LSTM	**0.811**	**0.833**	**0.795**	**0.814**
FCDNN	0.610	0.624	0.603	0.613
Hybrid FCDNN	**0.823**	**0.820**	**0.840**	**0.833**

5. Conclusions

Sentiment analysis has become an essential task in data analytics due to the communication technology revolution. Having easy access to communicative devices, a large number of people respond to anything that occurs and interests them. In the current paper, the authors have presented a new hybrid model that outperforms the existing QSR model and other existing techniques. The QSR model had achieved an accuracy of 0.5669 in the case of Naïve Bayes classifier, whereas the accuracy achieved by the proposed hybrid model is 0.7603. Similarly, this difference of higher accuracy is exhibited in Support Vector Machine, Random Forest, Convolutional Neural Network, Long term Short Term Memory classifiers in similar conditions. The same hybrid model was applied to test the accuracy of results related to diabetic tweets. The proposed model also showed better accuracy, recall, precision, and F1 score than the existing models. Hence, it can be concluded that hybrid model is more efficient and dependable in analyzing sentiment140 dataset and diabetes dataset than the QSR model and other existing techniques. It can also be utilized in other spheres of data analytics to achieve better results in terms of accuracy, which would contribute to a better understanding, management, and policy formation regarding anything that requires data classification.

References

[1] H. Medhat, W., Hassan, A., &Korashy, “Sentiment analysis algorithms and applications: A survey.,” Shams Eng. J., vol. 5, no. 4, pp. 1093–1113, 2014.

[2] H. Saif, Y. He, and H. Alani, “Semantic sentiment analysis of twitter,” in International semantic web conference, 2012, pp. 508–524. doi: 10.1007/978-3-642-35176-1_32.

[3] I. H. Sarker, Y. B. Abushark, F. Alsolami, and A. I. Khan, “IntruDTree: A machine learning based cyber security intrusion detection model,” Symmetry (Basel)., vol. 12, no. 5, pp. 1–15, 2020, doi: 10.3390/SYM12050754.

[4] A. Boukerche and J. Wang, “Machine Learning-based traffic prediction models for Intelligent Transportation Systems,” Comput. Networks, vol. 181, no. May, 2020, doi: 10.1016/j.comnet.2020.107530.

[5] S. F. Ardabili et al., “COVID-19 outbreak prediction with machine learning,” Algorithms, vol. 13, no. 10, 2020, doi: 10.3390/a13100249.

[6] I. Gupta and N. Joshi, “Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic,” J. Intell. Syst., vol. 29, no. 1, pp. 1611–1625, 2020, doi: 10.1515/jisys-2019-0106.

[7] W. Sun, X., Liu, L., Li, C., Yin, J., Zhao, J., & Si, “Classification for remote sensing data with improved CNN-SVM method.,” pp. 164507–164516, 2019.

[8] B. Gupta and S. Shukla, “An Analysis of Credibility of Content on Twitter,” Glob. Res. Dev. J. Eng., vol. 2, no. 5, pp. 98–101, 2017.

[9] P. A. K. Wagh, Bhagyashri, J. V. Shinde, “A Twitter sentiment analysis using NLTK and machine learning techniques,” Int. J. Emerg. Res. Manag. Technol., vol. 6, no. 12, pp. 37–44, 2018.

[10] Y. Zhang, D. Song, P. Zhang, X. Li, and P. Wang, “A quantum-inspired sentiment representation model for twitter sentiment analysis,” Appl. Intell., vol. 49, no. 8, pp. 3093–3108, 2019, doi: 10.1007/s10489-01901441-4.

[11] V. García-Díaz, J. P. Espada, R. G. Crespo, B. C. Pelayo G-Bustelo, and J. M. Cueva Lovelle, “An approach to improve the accuracy of probabilistic classifiers for decision support systems in sentiment analysis,” Appl. Soft Comput. J., vol. 67, pp. 822–833, 2018, doi: 10.1016/j.asoc.2017.05.038.

[12] S. S. Chidananda, HimadriTanaya, Debashis Das, “Sentiment analysis using n-gram technique.,” in Progress in Computing, Analytics and Networking., 2018, pp. 359–367.

[13] R. Batra and S. M. Daudpota, “Integrating StockTwits with sentiment analysis for better prediction of stock price movement,” 2018 Int. Conf. Comput. Math. Eng. Technol. Inven. Innov. Integr. Socioecon. Dev. iCoMET 2018 - Proc., vol. 2018-Janua, no. December, pp. 1–5, 2018, doi: 10.1109/ICOMET.2018.8346382.

[14] N. C. PalakBaid, Apoorva Gupta, “Sentiment Analysis of Movie Reviews Using Machine Learning Techniques,” Int. J. Comput. Appl., vol. 179, no. 7, pp. 45–49, 2017, doi: 10.1007/978-981-16-2377-6_34.

[15] S. S. and P. K.V., “Sentiment analysis of malayalam tweets using machine learning techniques,” ICT Express, vol. 6, no. 4, pp. 300–305, 2020, doi: 10.1016/j.icte.2020.04.003.

[16] A. Go, R. Bhayani, and L. Huang, “Twitter Sentiment Classification using Distant Supervision,” Processing, vol., pp. 1–6, 2009.

[17] D. Srivastava, R. Singh, and V. Singh, “Analysis of different Hybrid methods for Intrusion Detection System,” Int. J. Comput. Sci. Eng., vol. 7, no. 5, pp. 757–764, 2019, doi: 10.26438/ijcse/v7i5.757764.

[18] P. Kumar, S. Gupta, and R. Ratan, “New Ultrasound Image Watermarking Enhancement algorithm for cardiovascular disease,” Proc. IEEE Int. Conf. Signal Process. Control, vol. 2021-Octob, pp. 512–517, 2021, doi: 10.1109/ISPCC53510.2021.9609363.

[19] A. Sood, M. Hooda, S. Dhirn, and M. Bhatia, “An Initiative To Identify Depression Using Sentiment Analysis: a Machine Learning Approach,” Indian J. Sci. Technol., vol. 11, no. 4, pp. 1–20, 2018, doi: 10.17485/ijst/2018/v11i4/119594.

[20] Narayan, Vipul, et al. "Enhance-Net: An Approach to Boost the Performance of Deep Learning Model Based on Real-Time Medical Images." Journal of Sensors 2023 (2023).

[21] NARAYAN, VIPUL, A. K. Daniel, and Pooja Chaturvedi. "FGWOA: An Efficient Heuristic for Cluster Head Selection in WSN using Fuzzy based Grey Wolf Optimization Algorithm." (2022).

[22] Babu, S. Z., et al. "Abridgement of Business Data Drilling with the Natural Selection and Recasting Breakthrough: Drill Data With GA." Authors Profile Tarun Danti Dey is doing Bachelor in LAW from Chittagong Independent University, Bangladesh. Her research discipline is business intelligence, LAW, and Computational thinking. She has done 3 (2020).

[23] Narayan, Vipul, A. K. Daniel, and Pooja Chaturvedi. "E-FEERP: Enhanced Fuzzy based Energy Efficient

Routing Protocol for Wireless Sensor Network." Wireless Personal Communications (2023): 1-28.

[24] Faiz, Mohammad, et al. "Improved Homomorphic Encryption for Security in Cloud using Particle Swarm Optimization." Journal of Pharmaceutical Negative Results (2022): 4761-4771.

[25] Tyagi, Lalit Kumar, et al. "Energy Efficient Routing Protocol Using Next Cluster Head Selection Process In Two-Level Hierarchy For Wireless Sensor Network." Journal of Pharmaceutical Negative Results (2023): 665-676.

[26] Paricherla, Mutyalaiah, et al. "Towards Development of Machine Learning Framework for Enhancing Security in Internet of Things." Security and Communication Networks 2022 (2022).

[27] Pramanik, Sabyasachi, et al. "A novel approach using steganography and cryptography in business intelligence." Integration Challenges for Analytics, Business Intelligence, and Data Mining. IGI Global, 2021. 192-217.

[28] Srivastava, Swapnita, et al. "An Ensemble Learning Approach For Chronic Kidney Disease Classification." Journal of Pharmaceutical Negative Results (2022): 2401-2409.

[29] Sawhney, Rahul, et al. "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease." Decision Analytics Journal 6 (2023): 100169