Abstract

Sentiment analysis and opinion mining are one of those most active research areas in the field of Natural language processing.They deal with the study of people's opinion on different aspects and their sentiment and attitudes using written language.This process is also used in web mining ,text mining and data mining areas.It has great importance in business field and society as a whole. Sentiment analysis of text helps in identifying the expressions and emotions in text and classifies them accordingly. Also social media is generating huge amount of data in terms of tweets,blog posts,status updates etc.Sentiment analysis of these generated data is very helpful to know about crowd opinion on a particular topic.

Twitter sentiment analysis helps to know about the the emotion present in a tweet. Users share their daily lives, post their opinions on everything such as brands and places. Companies can benefit from this massive platform by collecting data related to opinions on them Sentiment analysis deals with identifying and classifying opinions or sentiments expressed in source text. Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and misspellings . So we propose a hybrid model built upon deep learning techniques to classify tweets based on the emotional content present in them , i.e, whether they infer positive sentiment or negative sentiment.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As the internet has evolved, the popularity of internet users has been increasing tremendously and people are engaging themselves in many of the social media platforms to share their knowledge,opinion regarding various issues.Thus the social media posts harbour plenty of significant information about the positive and negative attitudes of people.The availability of the real time opinions can make a revolution in computational linguistics.Because of this, Sentiment analysis of social media posts is gaining attention from researchers.

In order to perform text analysis various tools have been developed as advancements in Natural Language Processing and Machine learning took place and continue to be enhanced with different techniques.The tools provided by natural language processing and machine learning provide a way to interact with large volumes of text and extract insights.There are lot of techniques developed in order to get the sentiment hidden inside a block of text.

For training a model which is capable of taking a piece of text , extract features describing the sentiment and predicting the emotion present in it, requires huge quantity of structured data. But the data available from the internet is unstructured and requires lot of preprocessing before feeding them to train a model.Also human languages are very ambiguous that the same sentiment can be expressed in different ways and in different contexts. Social media documents such as twitter posts can contain user handles,urls, and some other redundant information for the model, in the same time they contain some useful information like emoticons and emojis which provide a great insight regarding the sentiment embedded in the sentence. Also people tend to use short forms and slang words which has to be taken care of. Another challenge is misspelled words which may significantly increase the count of features upon which,the model has to be trained.

We have proposed a hybrid model comprising of CNN and BiLSTM layers for classifying the tweet as either positive or negative. CNNs are best suited to get the n-gram effect while LSTM is capable of identifying long term dependencies.We compare the performance of stand alone CNN and BiLSTM with that of the hybrid model on sentiment140 dataset[5].

1.1 Problem Definition

Twitter is one of the highly used social media where people share their opinions regarding variety of aspects.Using different tools available for text analysis,the sentiment present in the tweet can be derived.The aim of the project is to develop such a system which takes twitter post as input.It must be able to process the tweet and remove redundant information and consider only those parts which contain information about the emotional content.

As popularity of performance of hybrid models in many of the tasks are increasing we apply a hybrid model developed with CNN and BiLSTM to classify these processed tweets as either positive or negative based on their content.

1.2 Motivation

Machine learning tools combined with natural language processing are capable of building effective tools to gather significant information about the sentiment embedded inside the posts created by people online. They can be potential solution for various problems associated with wide range of operations ranging from product creation and development by companies to creation of various governmental regulations and services.

Sentiment analysis has many applications and can add benefits to business and organizations.It can be used to get valuable insights about how people feel about the products or services being provided to them.System capable of analyzing sentiment from the text can be applied to identify potential product advocates and social media influencers.It can be used to detect potential negative threads emerging online regarding the business and helps to take appropriate actions for business improvements.

1.3 Objectives

The important steps of developing proposed system of sentiment analysis involves,

- Collection of dataset of twitter posts containing both posts and respective sentiment value.
- Processing the tweets to remove redundant information using natural language processing techniques.
- Developing and training a CNN model having an architecture to mimic n-gram technique.
- Developing and training a BiLSTM model on the dataset.
- Developing and training a hybrid model comprising of CNN and BiLSTM for classification
- Visualizing and analyzing the performance of all the three models on real time input and retuning the model based on results for improvisation.
- Deploying the trained model on the internet and getting results for real time input through a web application.

1.4 Report Outline

The first chapter of this report corresponds to the introduction to this project .i.e., the problem statement of the project, the main motivation behind this project and objectives of the project. The second chapter corresponds to the literature survey which analyzes the researches already done. The third chapter corresponds to the existing techniques or methods which are already present with respect to this project. The fourth chapter corresponds to the methodology used in this project and also the flow chart corresponding to it. The fifth chapter corresponds to the results obtained in this project, comparison of the work with base paper work and also the applications of the project. The sixth chapter corresponds to the conclusion about the project and the recommendation to the future work.

Chapter 2

Literature Survey

This section provides relevant methods and theories ,and also the existing research has been discussed and already existing methodologies are analyzed.

Yu Qing L et al., in their research[10] have worked upon text sentiment analysis on sentiment140 dataset which was created by Stanford University Students and explored different techniques such as Count Vectorizer, TF-IDF Vectorizer for preprocessing , word2vec and ELMo representations and tested a Logistic Regression model built upon these features.They achieved an F1 score of about 77% using these techniques.

Research[11] from Usman Naseem et al.,explores transformer based Deep Intelligent Contextual Embedding for Twitter Sentiment analysis on airlines datasets.They used DICE method for representing tweets so as to preserve contextual relations and trained Bidirectional LSTM. According to the results obtained it was concluded that DICE method can handle different language complexities.

Illanthenral et al., proposed a system[9] based on neutrosophy which factors in the concept of indeterminancy.They used Single Valued neutrosophic set (SVNS),triple refined indeterminate neutrosophic set (TRINS) and MRNS have been used.They used tweets related different scenarios as the dataset and it was seen that MRNS gave the best accuracy compared to other two techniques.

The research[18] by Ashima et al., provides a detailed survey of popular deep learning models that are increasingly applied in sentiment analysis. It presents taxonomy of sentiment analysis and discuss the implications of popular deep learning architecture. It explores different approach based sentiment analysis such as document level,multimodal,multidomain etc.It discusses the performance of differnt recurrent neural networks,CNN,GRUs with word embeddings.

Short term Sentiment Analysis [15] by Chao Song et al.,proposed Word2PLTS invovling probablistic linguistic terms sets. They used Support Vector Machines as the classifier for peroforming predictions. The datasets used were Standford Sentiment140 and Movie Reviews.87% was the highest accuracy achieved.

Study of Literatures's early impact with sentiment analysis[7] by Saeed et al.,uses twitter posts to get insights about the articles published.The dataset used was from Altmetric.The method used was TF-IDF with Support Vector Machine model.They found that adding lexical information improved the performance of the model.

Oliver and others researched about different deep learning approaches[6].The research explores about different hybrid models including CNN and Bi-GRU with attention mechanism and the accuracy achieved was around 85%.

Targeted Sentiment analysis by Chenquan and others[4] studies Sparse attention based seperable dilated convolutional neural network. The datasets included twitter dataset and Reviews dataset. SA-SDCNN model achieved an accuracy of about 75% on twitter dataset and 88% on reviews dataset.

Gonzalo et al.,[12] consider Bayesian network classifiers to perform sentiment analysis on two datasets in Spanish: the 2010 Chilean earthquake and the 2017 Catalan independence referendum.n this paper, bag-of-words (BOW) technique is used to convert training tweets into a numeric representation resulting in a term document matrix (TDM).Overall, the best performance was obtained by the SVM classifier followed by the BF TAN model with 19 edges.The SVM was the only classifier obtaining above 80% of accuracy in both datasets (81.2% and 82.9%).

SentiVec[21] by Luyao Zhu et al. makes use of sentiment context vector via kernel optimization.It was aimed at integrating the statistical information and sentiment orientation into sentiment word vectors and propagating, updating the semantic information to all the word representations in a corpus.The first phase uses a supervised learning while the second phase uses object-word-to-surrounding-words reward model and context-to-object-word reward model.The highest accuracy achieved using these methods was around 85%.

Khe Foon H et al.[8] made use of gradient boosting trees for predicting student satisfaction with MOOCs for sentiment analysis.The dataset was constructed by the collection of metadata from randomly selected MOOCs from Class Central.Gradient boosting was found be best model for the task and it achieved an accuracy of about 77% considering the structure of MOOCs.

Li Yang et al., proposed SLCABG[19], which is based on the sentiment lexicon and combines Convolutional Neutral Network(CNN) and attention based Bidirectional Gated Recurrent Unit.The dataset used was the data of book reviews collected from Dangdang using web crawler.Using the techniques mentioned they achieved an accuracy of about 93%.

Research[3] from Muhammad Asif and et al., aims at analysing extremism in social media from textual information.Using Linear Support Vector classifier incorporated text views were classified as high extreme,low extreme,moderate and neutral based on level of extremism with an accuracy of 82%.

Murtadha Ahmed and et al.,proposed a novel approach[1] to build domain dependent sentiment dictionary, SentiDomain. It invovles a weak supervised model aiming at learning a set of sentiment clusters embedding.It also proposes an attention based LSTM model to address aspect level sentiment analysis.Datasets used were from SemEval 2014.TD-LSTM model was able to classify restaurant related data with an accuracy of 82.6%.

Multi label sentiment analysis by Tao and Fang[16] studies transfer learning models based on BERT and XLNet.The dataset was collected from Yelp related to restaurant reviews and another dataset was wine reviews. LSTM and CNN were used as baseline models.66% was the accuracy achieved using these technique for Yelp dataset, where as for wine reviews it was around 79%.

Kai Shuanga et al., proposed an Feature Distillation network[13] capable of dealing with the challenge of aspect sensitive sentiment and reducing the introduction of noise irrelevant to the given aspect.It consists of Joint Projection module which is built upon Bi-LSTM and Feature Distillation module.The datasets used were from SemEval 2014.82% was the highest accuracy achieved using the techniques mentioned.

Fang Yao and Yan Wang proposed a neural network based domain specific sentiment analysis[20] of tweets.The proposed method uses Domain adversarial Neural network which is a type of RNN.The proposed fully connected neural network achieved an F1-score of 82% on twitter sentiment dataset.

Comparative analysis and review of sentiment analysis[14] over Social media as done by Nikhil Kumar S and et al.,studies different approaches. It compares different researches done in the field, different datasets , different machine learning and deep learning techniques and their performance on those datasets.

Citation identification using Sentiment analysis[2] by Hanan and et al.explores text analysis with SVM,Multinomial Naive Bayes,KNN and other classifiers on benchmark dataset D1. The average f-measure achieved was 83%.

Multi modality Sentiment analysis by Guorong Xiao et al. [17] is based on Hierarchical Attentions and CSAT-TCN with MBM network.It aims to capture contextual information.The dataset used was CMU-MOSI.80% was the highest accuracy achieved using this technique.

In summary , to achieve higher accuracy in text classification task training a on large dataset is necessary so that model get exposed to different types of inputs. Without training a model on a huge dataset it is very much difficult to achieve higher efficiency as the model has to face diverse inputs.

Chapter 3

Existing Methods

There are number of methodologies employed in field of sentiment analysis.These approaches can mainly classified into machine learning-based, lexicon-based and hybrid methods.Machine learning techniques are one of the mostly once.

Machine learning based methods use labelled dataset to train a model and then employing that trained model for classifying text.SVM and Naive Bayes are popular algorithms for such classification in the field of Machine learning. While the lexicon based approach does not need any prior training in order to mine the data. It uses a predefined list of words, where each word is associated with a specific sentiment.. Finally in the hybrid approach, the combination of both the machine learning and the lexicon based approaches has the potential to improve the sentiment classification performance.The advantages of Machine learning approach are the ability to adapt and create trained models for specific purposes and contexts .

Another popular technique is n-gram.There are uni-gram,bi-gram, tri-gram and so on.They capture information from multiple words at the same time.By combining multiple methods of among the mentioned techniques better performance can be obtained.

Chapter 4

Methodology

This section deals with the steps and methods involved in developing an Sentiment Analysis System.These methods aim to develop a model to classify tweets as positive(1) or negative(0). The steps are explained in detail in following sections.

- **Dataset collection**

  The tweets dataset used for building the analysis tool is Sentiment140[5] which was built using Twitter API. It has following fields,

  1. the polarity of the tweet (0 = negative,4 = positive)
  2. the id of the tweet (e.g 2087)
  3. the date of the tweet (e.g Sat May 16 23:58:44 UTC 2009)
  4. the query (lyx). If there is no query, then this value is NO_QUERY.
  5. the user that tweeted (e.g robotickilldozr)
  6. the text of the tweet (e.g Lyx is cool)

  Among the mentioned fields.the only two fields of our interest are the text content and corresponding polarity of the text/tweet.The wordclouds for the words appeared in positive and negative tweets are shown in Figure 4.1 and Figure 4.2 respectively.

Figure 4.1: WordCloud : words in positive tweets

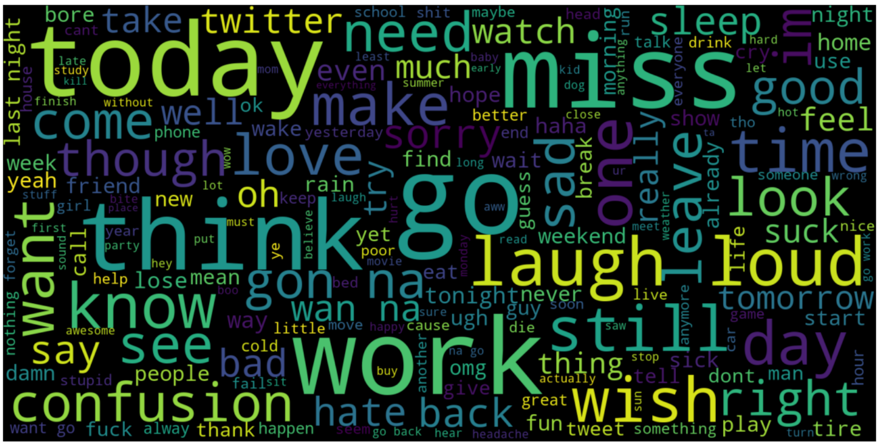


Figure 4.2: WordCloud : words in negative tweets

- **Text Preprocessing**

  The diversity in structure of a tweet is one of the challenging part as it poses a problem to model accuracy. The tweet has to be processed before feeding into the model. There are many unwanted entities inside a tweet considering sentiment analysis, for example urls,user handles,symbols.They provide no contribution towards the sentiment prediction process.

  Emojis play an important role in identifying the emotion present in the tweet. It makes the process much simpler. So emojis are replaced with words representing them. Shortforms are also replaced with their full forms.Thus important steps involved in this context are

  - Removing URLs (eg:,https://somewhere.com)
  - Removing user handles (eg: @username)
  - Replacing emojis with words
  - Replacing shortforms with words

  Stopwords removal is also known be one of the important step. But the model used in proposed system,t LSTM has to capture time dependency between words.So this step is ignored.

- **Vectorization of text**

  The preprocessed text has to be converted into vectors / tensors in order to feed them to a neural network. There are different approaches in Natural language processing techniques. We achieved this conversion by using a embedding layer, which learns tensor representation during training process. The input to embedding layer must be a number. So the words in dataset were given a index number and a embedding layer was trained for 5 epochs. The obtained results are as shown Figure 4.3 and Figure 4.4 using tensorboard visualization tool using PCA technique.The dimension of each vector representation is 728. The number of words for which the embedding layer was trained is 10,416. Thus the embedding matrix has a dimension of 10,416 X 728.

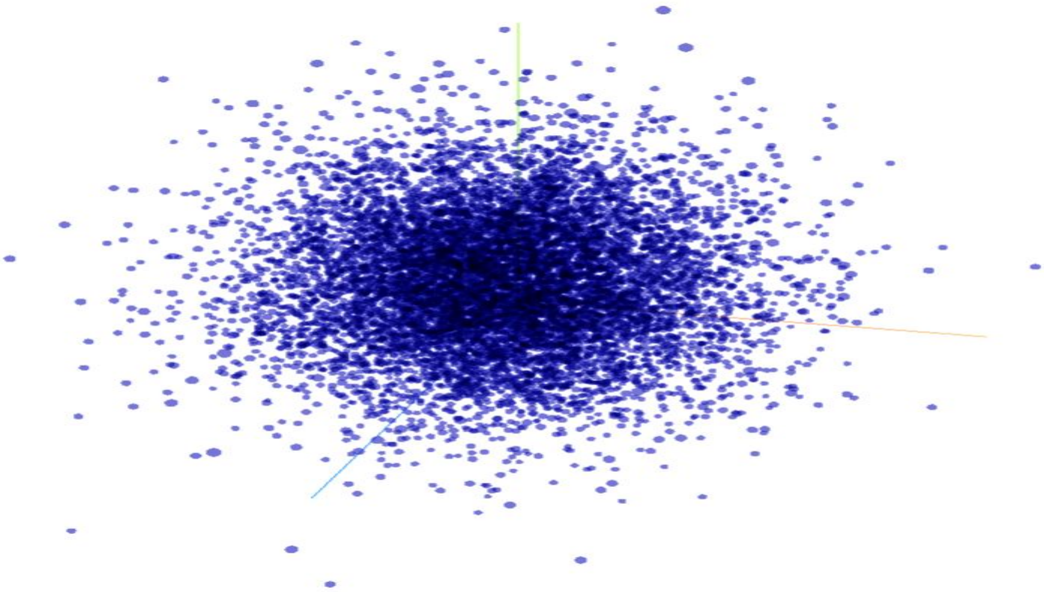


Figure 4.3: Embedding representation

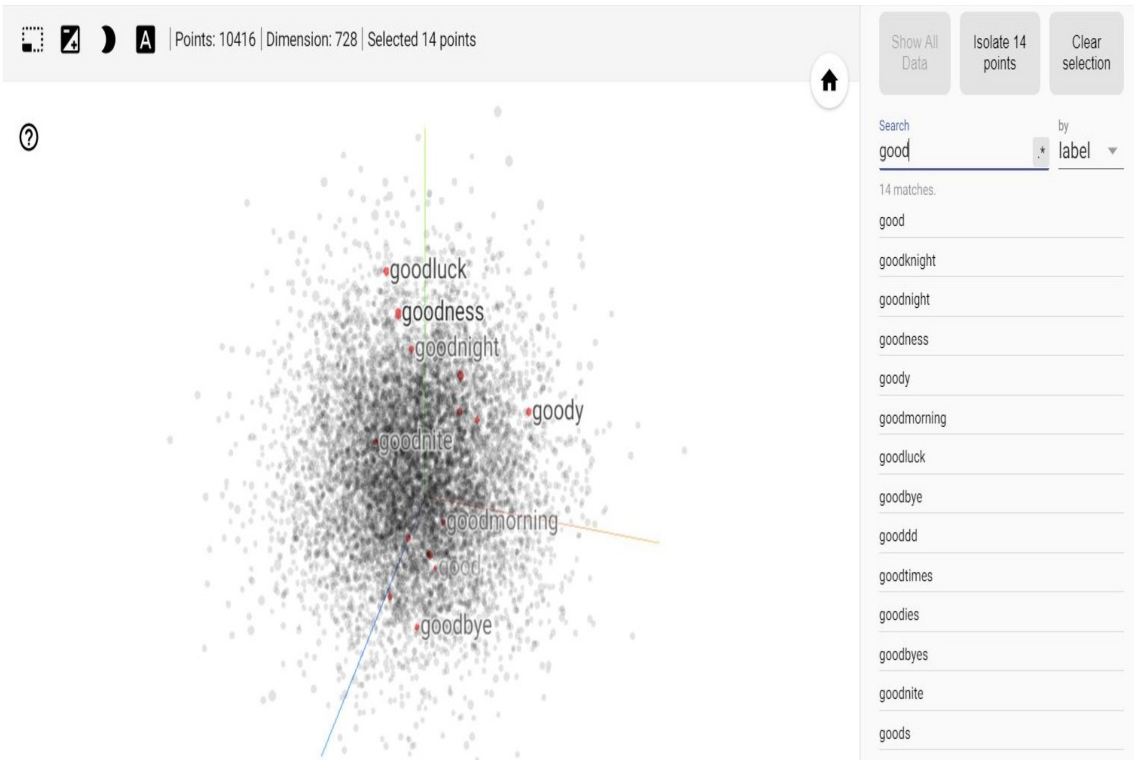


Figure 4.4: Embedding representation : related to word *good* highlighted

- **Proposed Neural Network Architectures**

  We propose 3 model architectures, CNN,LSTM and a hybrid model of CNN & LSTM.The respective details are as follows.

  - **Convolutional Neural Networks**

    Convolutional neural network is different from a normal neural network because it operates over a volume of inputs. Each layer tries to find a pattern or useful information of the data.

Convolution is a mathematical combination of two relationships to produce a third relationship. It joins two sets of information.Kernels/filters slide over input data the convolution to extract features.. This is important in feature extraction. There are some parameters associated with that sliding filter like how much input to take at once and by what extent should input be overlapped.Stride represents Size of the step filter moves every instance of time. Filter count is the Number of filters we want to use.

For text classification the behaviour of CNN can be used to get N-gram features.The architecture of the proposed CNN model used is shown in Figure 4.5. The proposed architecture has 4 kinds of 2D convolutional filters having heights 2,3,4,and 5 respectively.16 filters of each kind extracts features and passed through a 1D Maxpooling layer to get 4 tensors each having dimension 16 which are then passed through fully connected layers having number of neurons 64,32,4,1 respectively.The padding used was of the order height-1 for each kind of filter.tanh activation function before maxpooling and everywhere else ReLU was used as activation function.In fully connected layer part batchnorm was applied before activation function.

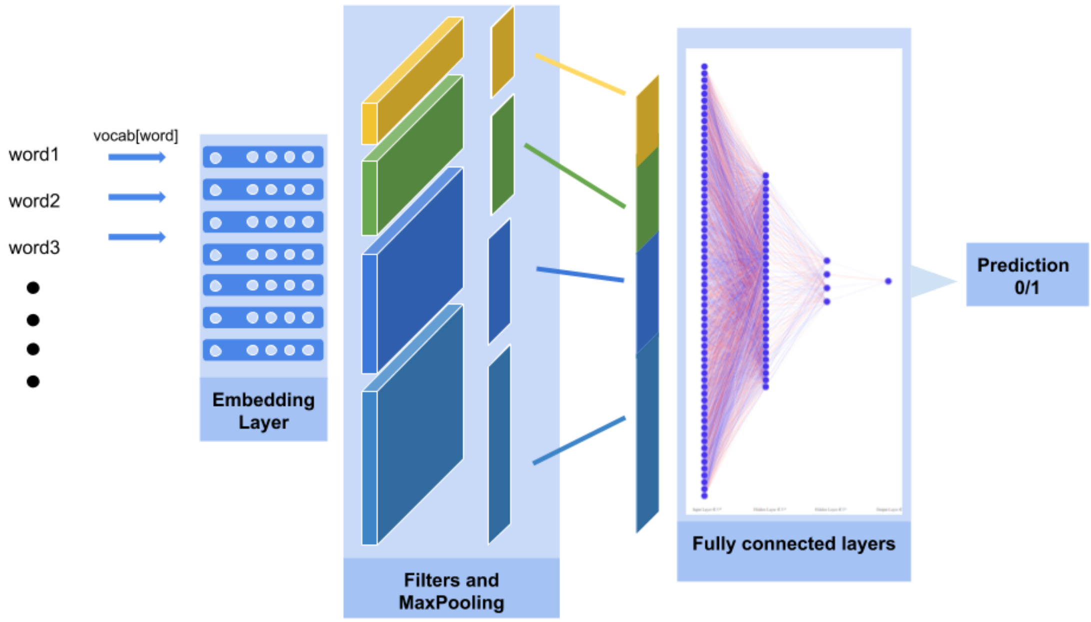


Figure 4.5: Proposed CNN based model architecture

– **Bidirectional Long Short Term Memory Neural Network (BiLSTM)**
A Bidirectional LSTM, or biLSTM, is a sequence processing model that consists of two LSTMs: one taking the input in a forward direction, and the other in a backwards direction. BiLSTMs effectively increase the amount of information available to the network, improving the context available to the algorithm (e.g. knowing what words immediately follow and precede a word in a sentence).

The architecture of the proposed LSTM model used is shown in Figure 4.6 containing two Bidirectional LSTM layers having hidden dimension of 256 taking 100 words represented in 728 dimension.Output from second bilstm is fed to fully connected layers having 64,32,4,1 neurons each.ReLU was used activation function and batch normalization was applied before every ReLU function.

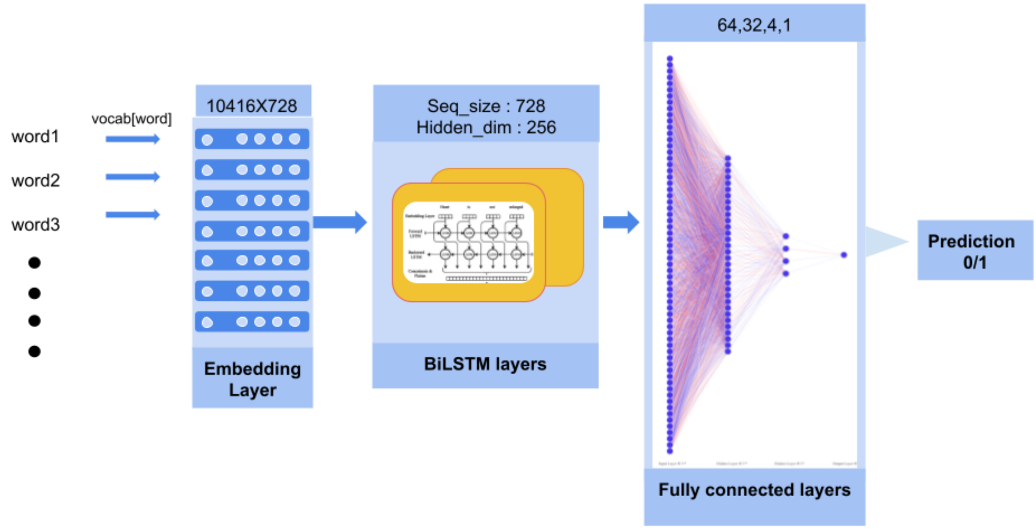


Figure 4.6: Proposed BiLSTM based model architecture

– **Hybrid Model using CNN and BiLSTM**
The hybrid model was developed based on CNN and BiLSTM. The output from these models was first passed through dense layers separately,then they were concatenated and passed through final fully connected layers.

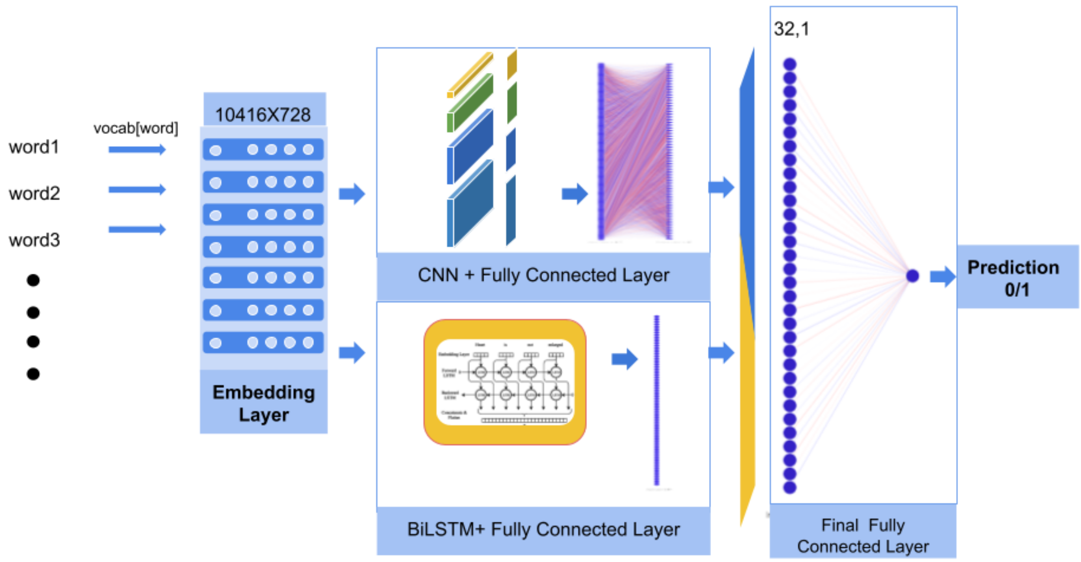


Figure 4.7: Proposed Hybrid model architecture

4.1 Flow Chart

The proposed system's flow chart is given in Figure 4.8 It starts with collecting and preprocessing the dataset.Then the preprocessed tweets are converted to tensors using embeddings.These embeddings are further used to train the model to classify the tweet as positive represented by 1 or negative represented by 0.

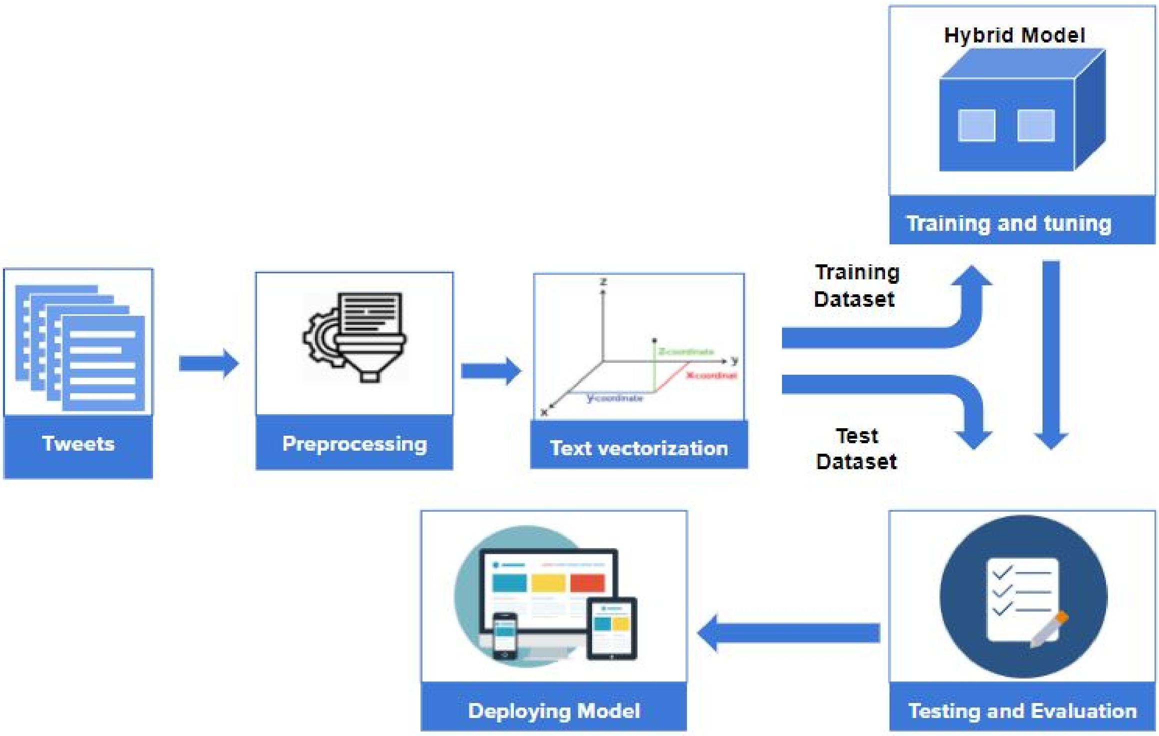


Figure 4.8: Flow chart of Proposed Sentiment Analysis System

The dataset once processed is split into training dataset and test dataset. The model once trained,is tested against test dataset for evaluating its performance. The model parameters are retuned as per requirement.Once the results seem satisfying the model is deployed to internet and can be used classify tweets through web application.

4.2 Mathematical Modelling

Mathematics behind Convolutional Neural Networks

The equations for Convolution layers are

$$conv(I,K)_{x,y} = \sum_{i=1}^{n_H} \sum_{j=1}^{n_W} \sum_{k=1}^{n_C} K_{i,j,k} I_{x+i-1,y+j-1,k}$$

More preciously, at the $\mathbf{l^{th}layer}$, we denote:

- `Input` : $a^{[l-1]}$ with size $(n_H^{[l-1]}, n_W^{[l-1]}, n_C^{[l-1]})$, $a^{[0]}$ being the image in the input
- `Padding` : $p^{[l]}$, `stride` : $s^{[l]}$
- `Number of filters` : $n_C^{[l]}$ where each $K^{(n)}$ has the dimension: $(f^{[l]}, f^{[l]}, n_C^{[l-1]})$
- `Bias` of the n^{th} convolution: $b_n^{[l]}$
- `Activation function` : $\psi^{[l]}$
- `Output` : $a^{[l]}$ with size $(n_H^{[l]}, n_W^{[l]}, n_C^{[l]})$

And we have:

$\forall n \in [1, 2, ..., n_C^{[l]}]$:

$$conv(a^{[l-1]}, K^{(n)})_{x,y} = \psi^{[l]}\left(\sum_{i=1}^{n_H^{[l-1]}} \sum_{j=1}^{n_W^{[l-1]}} \sum_{k=1}^{n_C^{[l-1]}} K_{i,j,k}^{(n)} a_{x+i-1,y+j-1,k}^{[l-1]} + b_n^{[l]}\right)$$
$$dim(conv(a^{[l-1]}, K^{(n)})) = (n_H^{[l]}, n_W^{[l]})$$

Thus:

$$a^{[l]} = [\psi^{[l]}(conv(a^{[l-1]}, K^{(1)})), \psi^{[l]}(conv(a^{[l-1]}, K^{(2)})), ..., \psi^{[l]}(conv(a^{[l-1]}, K^{(n_C^{[l]})}))]$$
$$dim(a^{[l]}) = (n_H^{[l]}, n_W^{[l]}, n_C^{[l]})$$

With:

$$\begin{aligned} n_{H/W}^{[l]} &= \left\lfloor \frac{n_{H/W}^{[l-1]} + 2p^{[l]} - f^{[l]}}{s^{[l]}} + 1 \right\rfloor ; s > 0 \\ &= n_{H/W}^{[l-1]} + 2p^{[l]} - f^{[l]} ; s = 0 \\ n_C^{[l]} &= number\ of\ filters \end{aligned}$$

The `learned parameters` at the l^{th} layer are:

- **Filters** with $(f^{[l]} \times f^{[l]} \times n_C^{[l-1]}) \times n_C^{[l]}$ parameters
- **Bias** with $(1 \times 1 \times 1) \times n_C^{[l]}$ parameters (broadcasting)

$$a_{x,y,z}^{[l]} = pool(a^{[l-1]})_{x,y,z} = \phi^{[l]}((a_{x+i-1,y+j-1,z}^{[l-1]})_{(i,j)\in[1,2,...,f^{[l]}]^2})$$
$$dim(a^{[l]}) = (n_H^{[l]}, n_W^{[l]}, n_C^{[l]})$$

With

$$\begin{aligned} n_{H/W}^{[l]} &= \left\lfloor \frac{n_{H/W}^{[l-1]} + 2p^{[l]} - f^{[l]}}{s^{[l]}} + 1 \right\rfloor ; s > 0 \\ &= n_{H/W}^{[l-1]} + 2p^{[l]} - f^{[l]} ; s = 0 \\ n_C^{[l]} &= n_C^{[l-1]} \end{aligned}$$

For a fully connected layer following pooling layer

In general, considering the j^{th} node of the i^{th} layer we have the following equations:

$$z_j^{[i]} = \sum_{l=1}^{n_{i-1}} w_{j,l}^{[i]} a_l^{[i-1]} + b_j^{[i]}$$
$$\rightarrow a_j^{[i]} = \psi^{[i]}(z_j^{[i]})$$

The input $a^{[i-1]}$ might be the result of a convolution or a pooling layer with the dimensions $(n_H^{[i-1]}, n_W^{[i-1]}, n_C^{[i-1]})$.
In order to be able to plug it into the fully connected layer we `flatten` the tensor to a `1D vector` having the dimension: $(n_H^{[i-1]} \times n_W^{[i-1]} \times n_C^{[i-1]}, 1)$, thus:

$$n_{i-1} = n_H^{[i-1]} \times n_W^{[i-1]} \times n_C^{[i-1]}$$

The `learned parameters` at the l^{th} layer are:

- **Weights** $w_{j,l}$ with $n_{l-1} \times n_l$ parameters
- **Bias** with n_l parameters

Mathematics behind Long-Short Term Memory Networks

LSTM and BiLSTM cells are shown in Figure 4.9.The gate equations and output equations are given in Figure 4.10

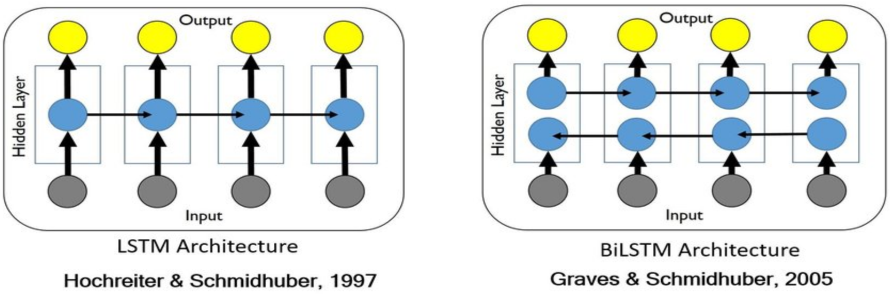
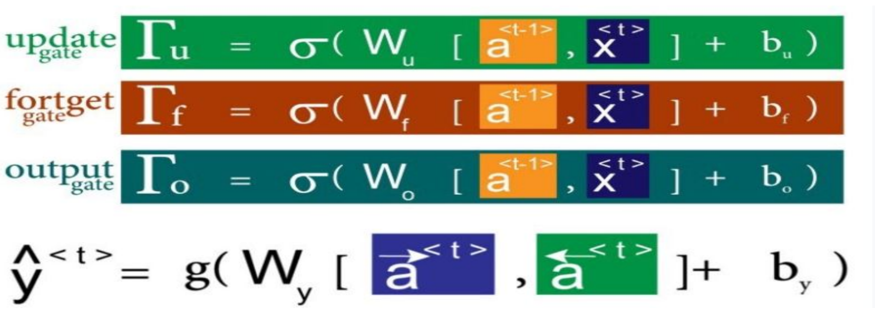


Figure 4.9: LSTM and BiLSTM cells



$$\Gamma_u = \sigma(W_u [a^{<t-1>}, x^{<t>}] + b_u)$$
$$\Gamma_f = \sigma(W_f [a^{<t-1>}, x^{<t>}] + b_f)$$
$$\Gamma_o = \sigma(W_o [a^{<t-1>}, x^{<t>}] + b_o)$$
$$\hat{y}^{<t>} = g(W_y [\overrightarrow{a}^{<t>}, \overleftarrow{a}^{<t>}] + b_y)$$

Figure 4.10: LSTM gates and output equations

Chapter 5

Result and Discussion

The metrics used to measure the performance of the model are classification accuracy and confusion matrix. All the models are trained with Training Dataset of size 89802 and validation dataset of 4989 tweets.The models were tested on 4990 tweets.

- **Accuracy score**

  It also called as Classification accuracy and it is the ratio of the number of correctly detected predictions to the total number of input samples fed to the system.

$$Accuracy = \frac{\text{Count of Correct Predictions by the model}}{\text{Total Number of test samples}}$$

- **Confusion Matrix**

  It is used to describe the performance of a classification model. The number of correct and incorrect predictions by the model are shown with corresponding classes and count values

Accuracy and loss curves of all the three models are shown Figure 5.1,Figure 5.2 and Figure 5.3.

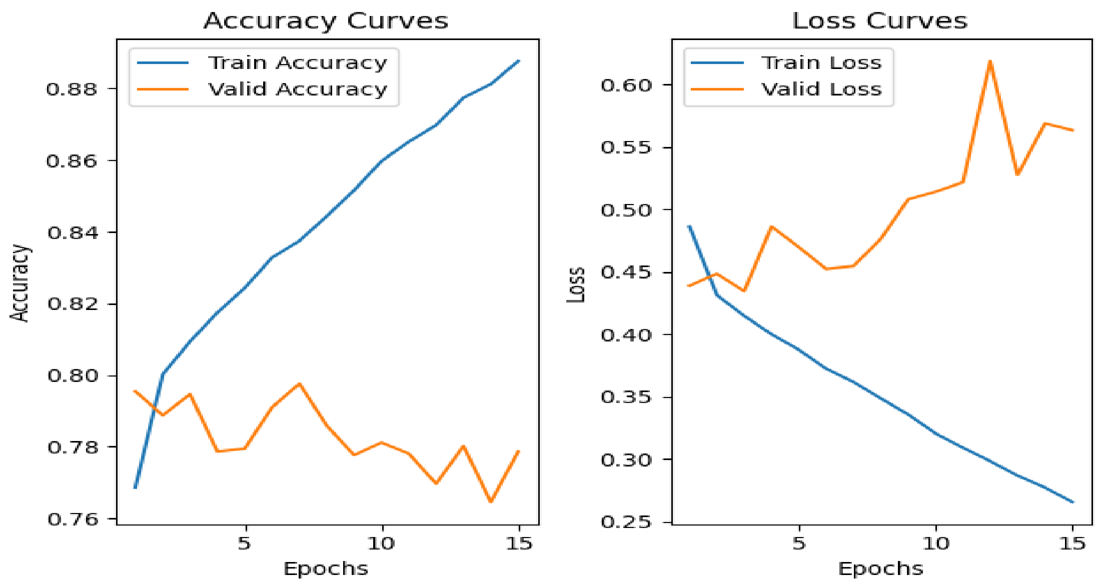


Figure 5.1: Accuracy and Loss Curves of proposed CNN based model

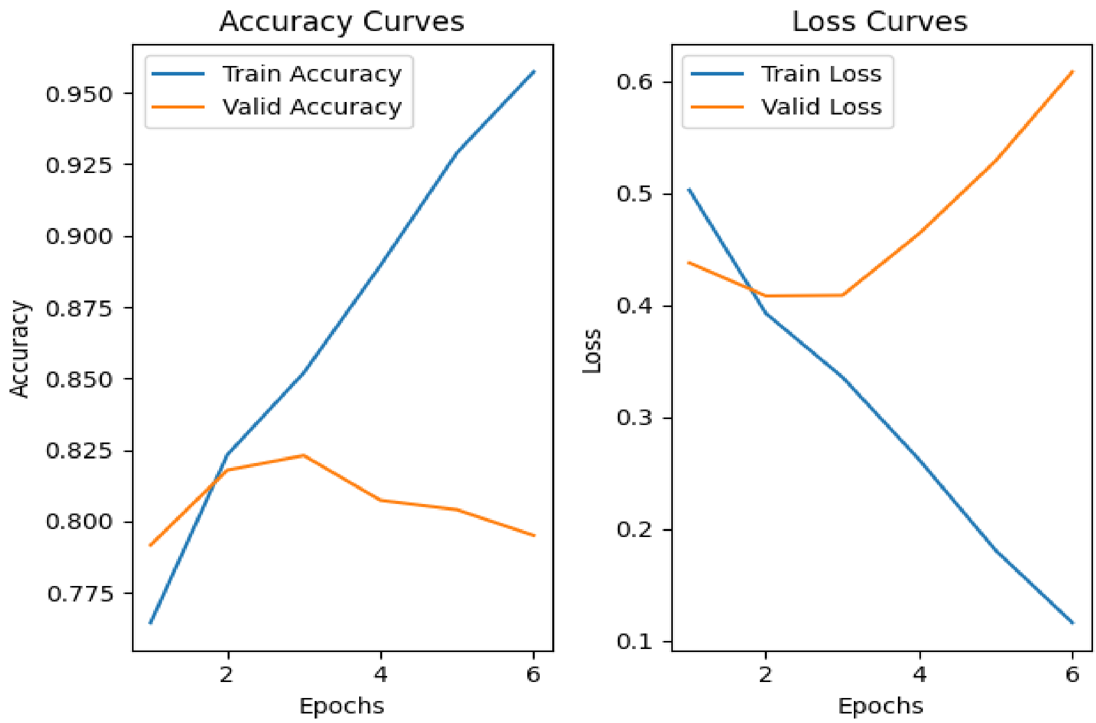


Figure 5.2: Accuracy and Loss Curves of proposed BiLSTM based model

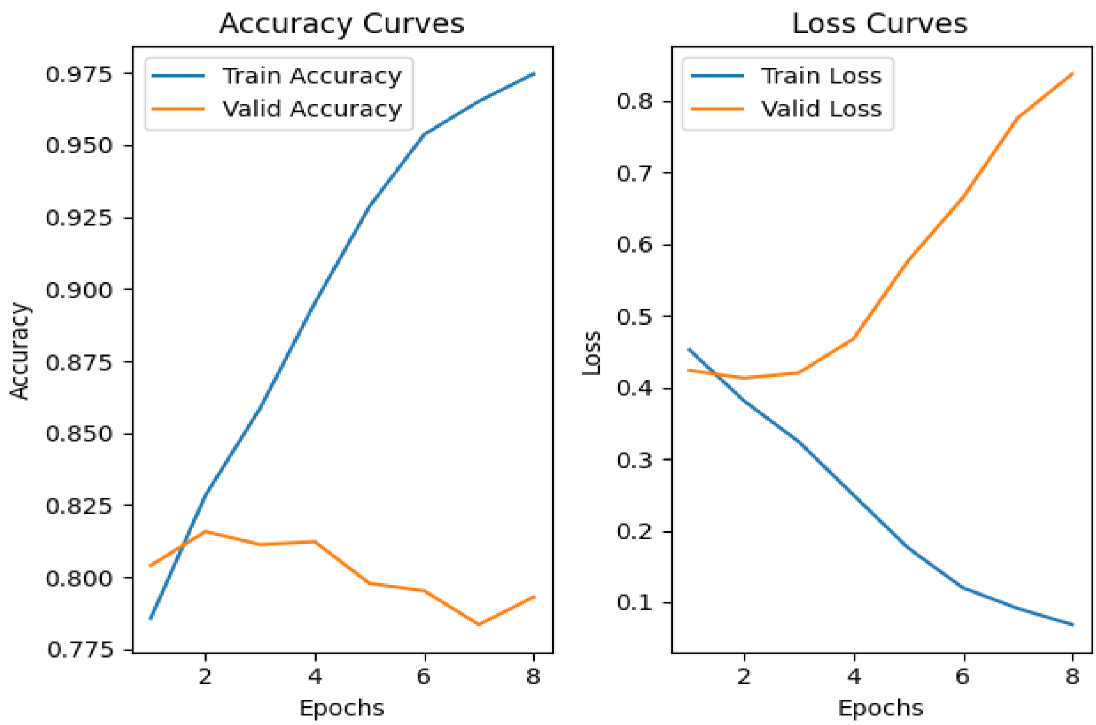


Figure 5.3: Accuracy and Loss Curves of proposed Hybrid model

As we can observe from the accuracy and loss curves, the models are getting overfit easily and optimal loss value can be obtained aroudn 3rd epoch itself.

The confusion matrices for all the three models are given in Figure 5.4,Figure 5.5 and Figure 5.6 respectively. The cnn based architecture achieved an accuracy of 79.28%. The lstm based architecture achieved an accuracy of 82.5%.

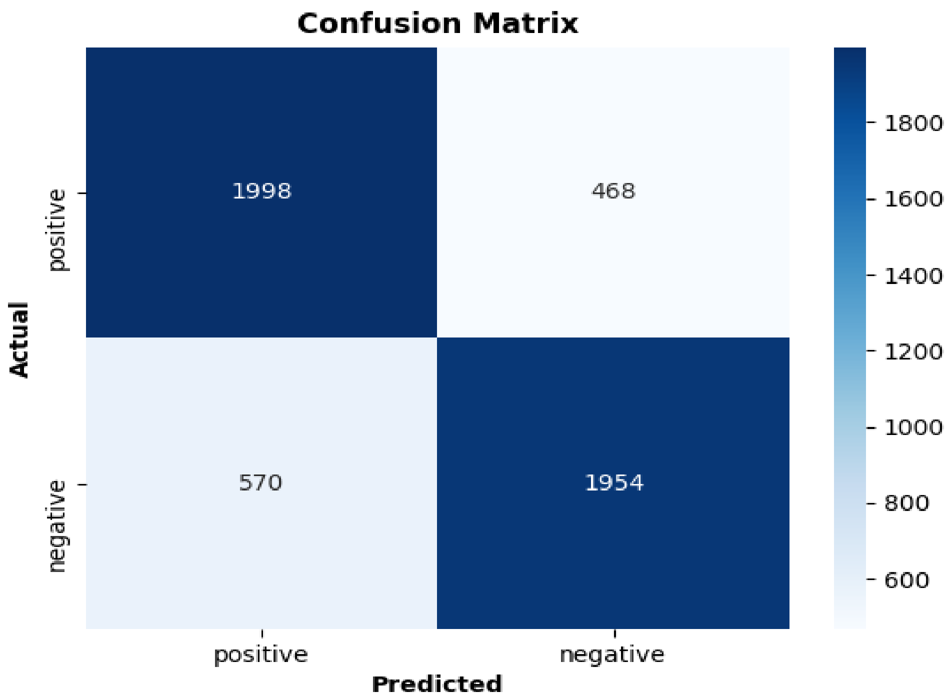


Figure 5.4: Confusion Matrix of proposed CNN based model

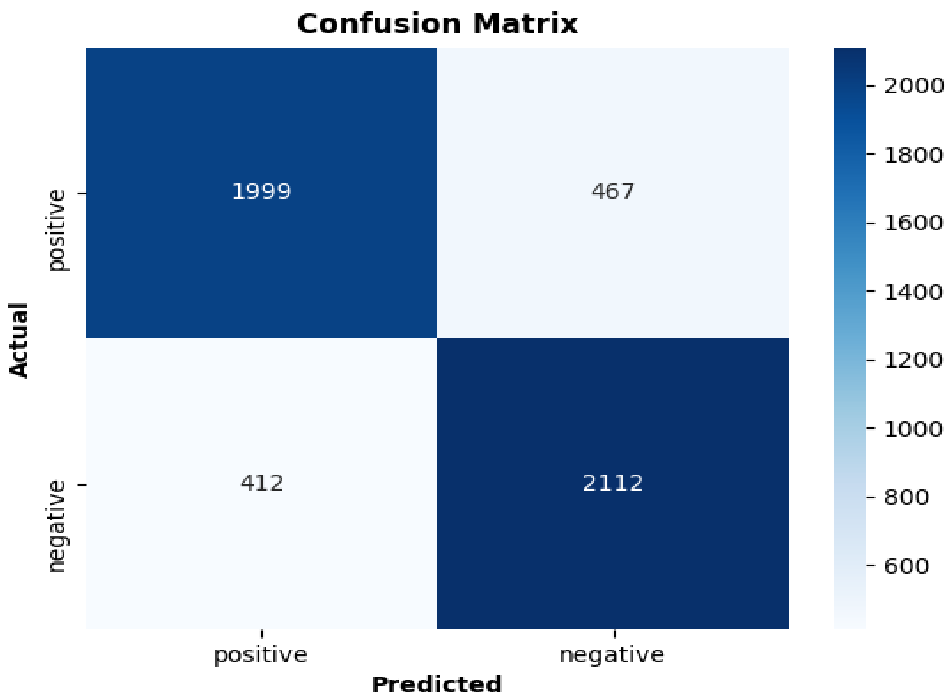


Figure 5.5: Confusion Matrix of of proposed BiLSTM based model

The hybrid architecture achieved an accuracy of 82.15%.

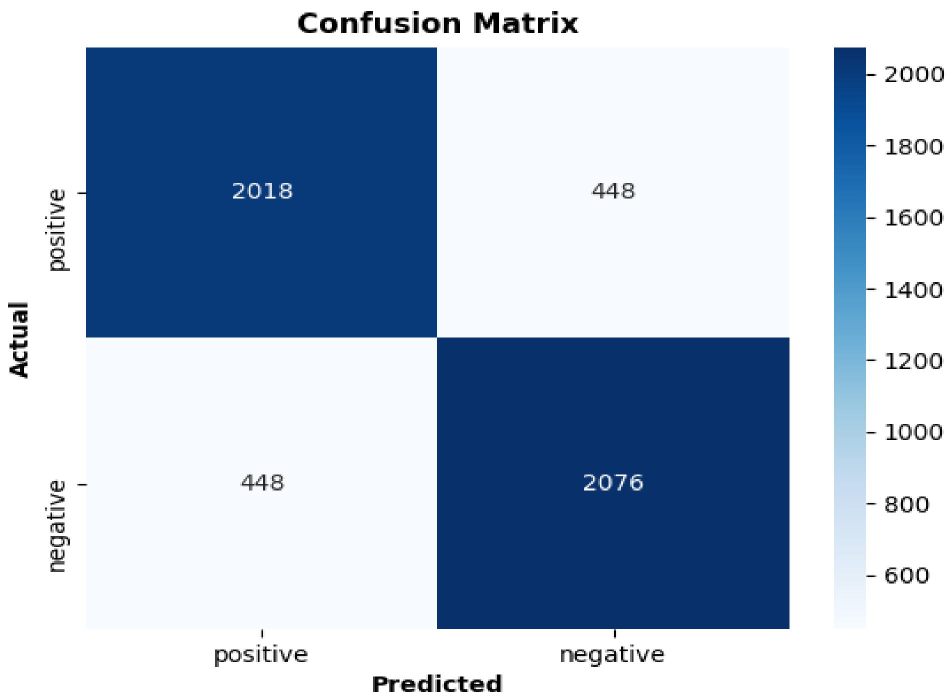


Figure 5.6: Confusion Matrix of of proposed Hybrid model

The web application was developed and depoyed using Github for frontend hosting and Heroku for server side.Frontend was developed using React JS and the model deployed using Flask framework. The images of web applications are shown in Figure 5.7,Figure 5.8 and Figure 5.9.

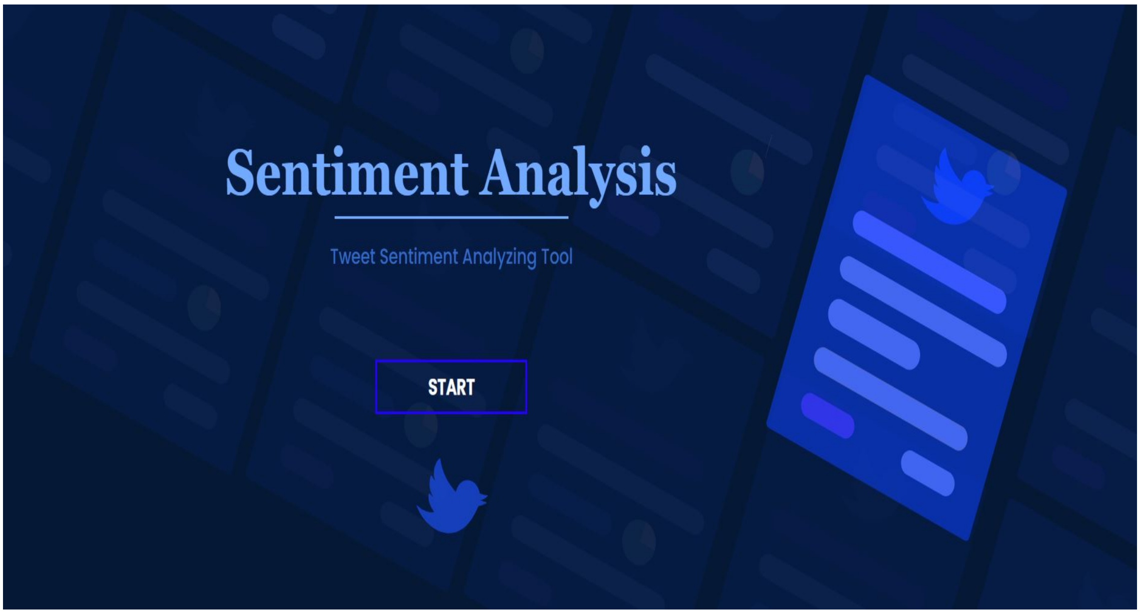


Figure 5.7: Web Application Landing Page

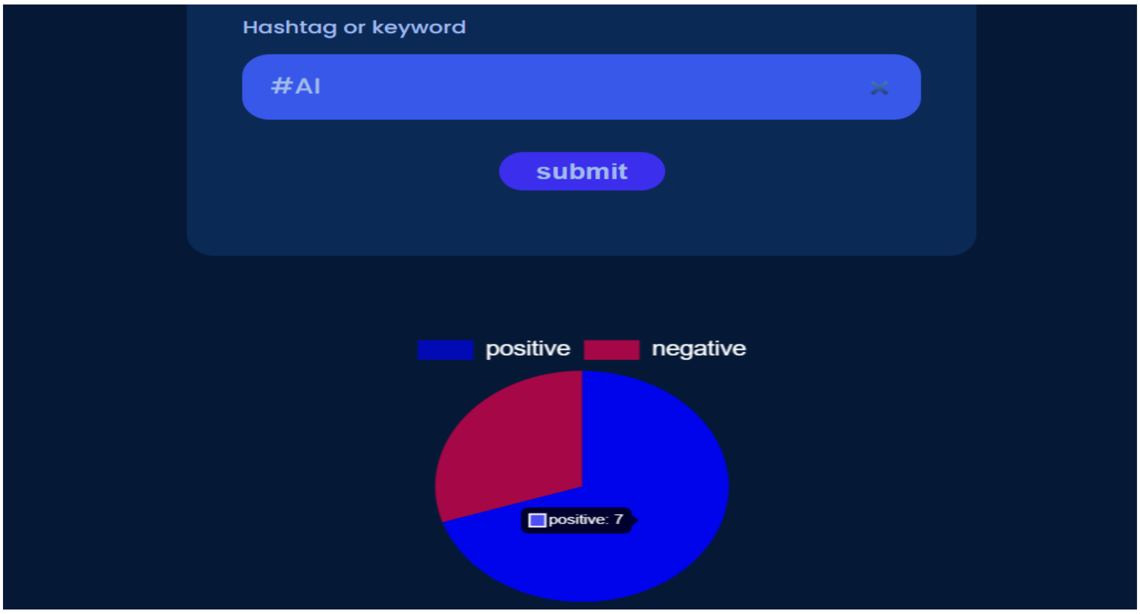


Figure 5.8: Search Results : #AI

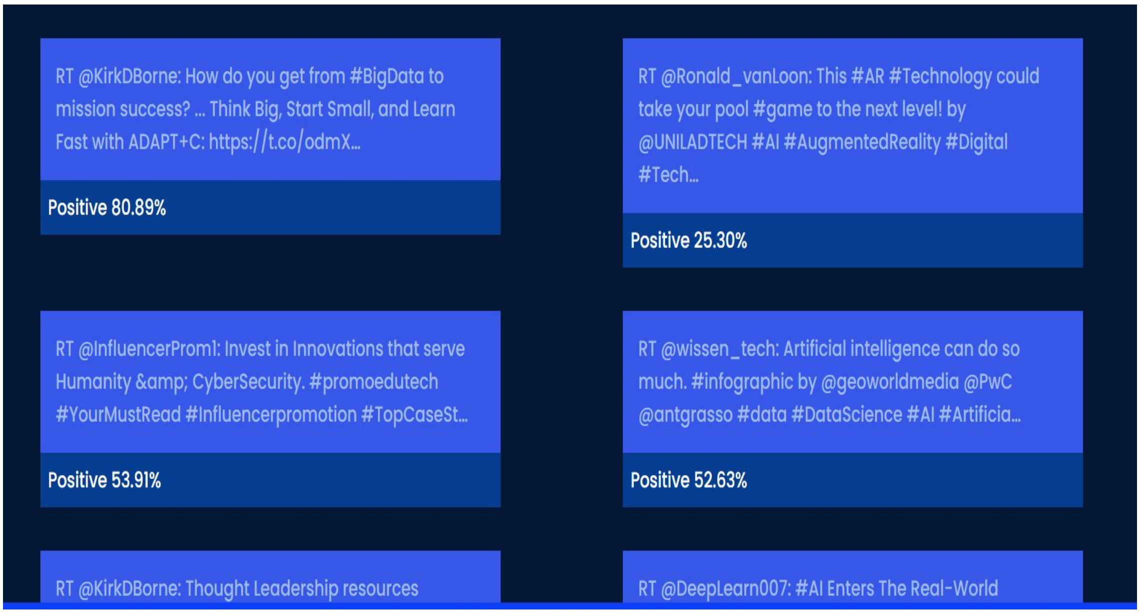


Figure 5.9: Search Results : #AI tweets

5.1 Comparison Table

The proposed work is compared with the work proposed by Lim and et al[10]. The table 5.1 shows the comparison of different aspects of our work with theirs.

Table 5.1: Comparison Table

Aspect	Lim et al.(2020)[10]	Ours
Goal	Goal is to utilise sentiment analysis on text extracted from social media(twitter) content to classify whether theyhave a positive or negative context	Goal is to develop a hybrid model to classify tweets as positive or negative.
Dataset	Sentiment140[5]	Sentiment140[5]
Methodology	Embeddings from Language Models (ELMo) and Word2Vec for text vectorization and logistic regression for classification	Word embeddings for text vectorization and CNN-BiLSTM hybrid model for classification.
Performance	Model trained and validated against around 5000 tweets. Accuracy of 77% (word2Vec) and 73%(ELMo)	Model trained with 89802 tweets and tested against 4990 tweets. Accuracy of 82.15%

5.2 Applications

- In marketing field companies use it to develop their strategies,to understand customers' feelings towards products or brand how people respond to their campaigns or product launches and why consumers don't buy some products.
- In political field, it is used to track of political view, to detect consistency and inconsistency between statements and actions at the government level
- It can be used to predict election results. Sentiment analysis also is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.
- The sentiment analysis then represents an important element for any subject (policy makers, stakeholders, companies etc.) to perform different kinds of activities such as: predict financial performance , understand consumers' perception provide early warnings, define election outcomes etc.

In all of these examples, the sentiment input is whether a given consumer opinion has negative, positive or neutral polarity regarding the different target of interest . The large amount of these contents required the use of automated techniques for analyzing since manually it is not possible.

Chapter 6

Conclusion

6.1 Conclusion

The growth of internet has led to th explosion of data available for analysis.Sentiment analysis of these social media posts can get detailed insights about the users around a particular topic.

In this work we have developed a hybrid model using Convolutional neural networks and Bidirectional Long Short Term Memory Units.Text preprocessing plays an important role for proper model and training and also speeds up the process.The LSTM is capable of long term dependencies,while CNN is used to get n-gram.

The CNN based model got 79.28% accuracy where as the BiLSTM got 82.5% accuracy.The hybrid model which is a combination of these two achieved 82.15%. The small decrease compared to BiLSTM is only due the test dataset used and in actual real world environment, the hybrid model is much more reliable.

6.2 Future scope

Sentiment analysis is one of the popular field in natural language processing . Many deep learning techniques can be used to get much better results in this task in real world scenario.Some of the aspects which worth exploring for better performance are

- Training on a large dataset can improve performance as the model gets exposed to large number of words.
- Contextual based word representations can be used to get meaningful representation and thereby better performance of the model.

References

[1] Murtadha Ahmed, Qun Chen, and Zhanhuai Li. "Constructing domain-dependent sentiment dictionary for sentiment analysis". In: *Neural Computing and Applications* 32.18 (Mar. 2020), pp. 14719–14732. DOI: 10.1007/s00521-020-04824-8. URL: https://doi.org/10.1007%2Fs00521-020-04824-8.

[2] Hanan Aljuaid et al. "Important citation identification using sentiment analysis of In-text citations". In: *Telematics and Informatics* (2020), p. 101492.

[3] Muhammad Asif et al. "Sentiment analysis of extremism in social media from textual information". In: *Telematics and Informatics* 48 (May 2020), p. 101345. DOI: 10.1016/j.tele.2020.101345. URL: https://doi.org/10.1016%2Fj.tele.2020.101345.

[4] Chenquan Gan et al. "Sparse attention based separable dilated convolutional neural network for targeted sentiment analysis". In: *Knowledge-Based Systems* 188 (Jan. 2020), p. 104827. DOI: 10.1016/j.knosys.2019.06.035. URL: https://doi.org/10.1016%2Fj.knosys.2019.06.035.

[5] Alec Go, Richa Bhayani, and Lei Huang. *Twitter Sentiment Classification using Distant Supervision*. 2009. URL: http://help.sentiment140.com/home.

[6] Olivier Habimana et al. "Sentiment analysis using deep learning approaches: an overview". In: *Science China Information Sciences* 63.1 (Dec. 2019). DOI: 10.1007/s11432-018-9941-6. URL: https://doi.org/10.1007%2Fs11432-018-9941-6.

[7] Saeed-Ul Hassan et al. "Predicting literature's early impact with sentiment analysis in Twitter". In: *Knowledge-Based Systems* 192 (Mar. 2020), p. 105383. DOI: 10.1016/j.knosys.2019.105383. URL: https://doi.org/10.1016%2Fj.knosys.2019.105383.

[8] Khe Foon Hew et al. "What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach". In: *Computers & Education* 145 (Feb. 2020), p. 103724. DOI: 10.1016/j.compedu.2019.103724. URL: https://doi.org/10.1016%2Fj.compedu.2019.103724.

[9] Ilanthenral Kandasamy et al. "Sentiment analysis of tweets using refined neutrosophic sets". In: *Computers in Industry* 115 (Feb. 2020), p. 103180. DOI: 10.1016/j.compind.2019.103180. URL: https://doi.org/10.1016%2Fj.compind.2019.103180.

[10] Yu Qing Lim et al. "Text Sentiment Analysis on Twitter to Identify Positive or Negative Context in Addressing Inept Regulations on Social Media Platform". In: *2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. IEEE, Apr. 2020. DOI: 10.1109/iscaie47305.2020.9108706. URL: https://doi.org/10.1109%2Fiscaie47305.2020.9108706.

[11] Usman Naseem et al. "Transformer based Deep Intelligent Contextual Embedding for Twitter sentiment analysis". In: *Future Generation Computer Systems* 113 (Dec. 2020), pp. 58–69. DOI: 10.1016/j.future.2020.06.050. URL: https://doi.org/10.1016%2Fj.future.2020.06.050.

[12] Gonzalo A. Ruz, Pablo A. Henrıquez, and Aldo Mascareño. “Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers”. In: *Future Generation Computer Systems* 106 (May 2020), pp. 92–104. DOI: 10.1016/j.future.2020.01.005. URL: https://doi.org/10.1016%2Fj.future.2020.01.005.

[13] Kai Shuang et al. “Feature distillation network for aspect-based sentiment analysis”. In: *Information Fusion* 61 (Sept. 2020), pp. 13–23. DOI: 10.1016/j.inffus.2020.03.003. URL: https://doi.org/10.1016%2Fj.inffus.2020.03.003.

[14] Nikhil Kumar Singh, Deepak Singh Tomar, and Arun Kumar Sangaiah. “Sentiment analysis: a review and comparative analysis over social media”. In: *Journal of Ambient Intelligence and Humanized Computing* 11.1 (May 2018), pp. 97–117. DOI: 10.1007/s12652-018-0862-8. URL: https://doi.org/10.1007%2Fs12652-018-0862-8.

[15] Chao Song et al. “SACPC: A framework based on probabilistic linguistic terms for short text sentiment analysis”. In: *Knowledge-Based Systems* 194 (Apr. 2020), p. 105572. DOI: 10.1016/j.knosys.2020.105572. URL: https://doi.org/10.1016%2Fj.knosys.2020.105572.

[16] Jie Tao and Xing Fang. “Toward multi-label sentiment analysis: a transfer learning based approach”. In: *Journal of Big Data* 7.1 (Jan. 2020). DOI: 10.1186/s40537-019-0278-0. URL: https://doi.org/10.1186%2Fs40537-019-0278-0.

[17] Guorong Xiao et al. “Multi-modality Sentiment Analysis in Social Internet of Things based on Hierarchical Attentions and CSATTCN with MBM Network”. In: *IEEE Internet of Things Journal* (2020), pp. 1–1. DOI: 10.1109/jiot.2020.3015381. URL: https://doi.org/10.1109%2Fjiot.2020.3015381.

[18] Ashima Yadav and Dinesh Kumar Vishwakarma. “Sentiment analysis using deep learning architectures: a review”. In: *Artificial Intelligence Review* 53.6 (Dec. 2019), pp. 4335–4385. DOI: 10.1007/s10462-019-09794-5. URL: https://doi.org/10.1007%2Fs10462-019-09794-5.

[19] Li Yang et al. “Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning”. In: *IEEE Access* 8 (2020), pp. 23522–23530. DOI: 10.1109/access.2020.2969854. URL: https://doi.org/10.1109%2Faccess.2020.2969854.

[20] Fang Yao and Yan Wang. “Domain-specific sentiment analysis for tweets during hurricanes (DSSA-H): A domain-adversarial neural-network-based approach”. In: *Computers, Environment and Urban Systems* 83 (Sept. 2020), p. 101522. DOI: 10.1016/j.compenvurbsys.2020.101522. URL: https://doi.org/10.1016%2Fj.compenvurbsys.2020.101522.

[21] Luyao Zhu et al. “SentiVec: Learning Sentiment-Context Vector via Kernel Optimization Function for Sentiment Analysis”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2020), pp. 1–12. DOI: 10.1109/tnnls.2020.3006531. URL: https://doi.org/10.1109%2Ftnnls.2020.3006531.