Sentiment Analysis using Few-Short Learning

Submitted by:

Archit Garg (9915103245)
Daksh Jain (9915103249)

Under the supervision of:
Dr. Mukesh Saraswat

MARCH 2019

Submitted in partial fulfilment of the Degree of Bachelor of Technology In Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

DECLARATION

We hereby declare that this submission is our own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: JIIT NOIDA

Date: 15/03/2019

Archit Garg (9915103245)

Daksh Jain (9915103249)

CERTIFICATE

This is to certify that the work titled **Comparing Study of Machine Learning Algorithms for Sentiment Analysis** submitted by **Archit Garg, Daksh Jain** of **B. Tech** of Jaypee Institute of Information Technology University, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree or diploma.

Signature of Supervisor

Name of Supervisor - Dr **Mukesh Saraswat**

Date: - 15/03/2019

ACKNOWLEDGEMENT

The completion of any inter-disciplinary project depends upon cooperation, co-ordination and combined efforts of several sources of knowledge. We are grateful to **Dr Mukesh Saraswat for** his even willingness to give us valuable advice and direction whenever we approached her with a problem. We are thankful to him for providing us with immense guidance for this project.

Signature(s) of Students

Archit Garg (9915103245)

Daksh Jain (9915103249)

SUMMARY

In today's world, everyone is expressive in one way or other. Many social websites and android applications whether being Facebook, WhatsApp or Twitter, in this highly advance and modernized world is flooded with views and data. One of the most global and popular platforms is Twitter. This is seen as the main source of sentiments where almost every enthusiastic or social person tends to express his or her views in form of comments. These comments not only express the people but also give the understanding of their mood. These sentiments can be utilized to solve the problem of text classification. People expression and their mood can be classified to mainly three components namely positive, negative and neutral. Sentiment analysis is an extremely old topic but very few have used few-short learning for the same thus in this work as a new contribution we have used very few data to classify sentiments thus adding a new technique for sentiment classification which would be computationally inexpensive and easy to implement. In our previous work we had applied 10 machine leaning algorithms on two datasets SS-tweet and HCR but were not able to obtain good results so to improve this time we applied hybrid deep learning model of CNN and Bi-LSTM and thus obtained a great jump in our results.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

Chapter 1: INTRODUCTION

1.1 GENERAL INTRODUCTION

Sentiment Analysis is a well classified problem of text processing in natural language processing. We have several ways in which we can use machine learning / Deep Learning algorithms on textual data to get some insights from it and thus we have used some of them. With an ongoing increase in social media interaction, deep learning algorithms provide several ways to get some results from the data. As the data volume increases each year, it becomes harder for companies to process and store information. Deep learning is itself a motivating subject in which we can use most awesome features of it to solve some real world problems. We have thus used its power to classify twitter tweets among negative, positive and neutral. Purpose of the project is to classify twitter tweets as positive, negative or neutral we have used four datasets SS-Tweet and Health Care Reform (HCR), US Airline dataset and Whatsapp Chat dataset. Main purpose of this project is to classify sentiments of a text using few short learning , there are many paper published for sentiment analysis but many of these used millions of data to train rather we have used maximum of 3000 sentences to train our model which makes it easy to compute and is time efficient with good results . In our previous work we applied 10 machine learning algorithms on 2 datasets – SS tweet and HCR but we didn’t got a good accuracy so in order to improve it now we have applied a hybrid model of CNN and Bi-LSTM on 4 datasets SS-tweet, HCR, US airline and Whatsapp chat dataset and thus we have seen a good improvement in our results

1.2 RELEVANT OPEN PROBLEMS

- Sentiment Analysis
- Few short learning
- Training Deep Learning models on normal machines (without external GPU)
- Text Classification

1.3 INTEGRATED SUMMARY OF LITERATURE STUDIED

Research paper name	Author	Summary
Few-short learning for short text classification[1]	Leiming Yan & Yuhui Zheng & Jie Cao	In this paper authors have used few-short learning for short text classification i.e. amount of data used is very less and they have applied several machine learning and deep learning techniques to classify text and got maximum accuracy of 88% from Siamese CNN. Datasets used by them are SS-Tweet, HCR Sentiment-B 2013 and Multigame.
Weakly-supervised Deep Embedding for Product Review Sentiment Analysis [2]	Wei Zhao, Ziyu Guan , Long Chen, Xiaofei He, Fellow, IAPR, Deng Cai, Beidou Wang and Quan Wang	In this paper authors have used amazon reviews and classified them as positive or negative, they have used a total of 12000 reviews to train. They have used CNN and LSTM separately and obtained an accuracy of 87% and 88% respectively.
Age Groups Classification in Social Network Using Deep Learning [3]	Rita G. Guimaraes, Renata L. Rosa, Denise De Gaetano, Dem ostenes Z. Rodr ıguez, Senior Member, IEEE, and Graca Bressan	In this paper authors have used twitter dataset and classified them on basis of age as Teenager, Adult and General. They have used various deep learning and machine learning algorithms and have obtained an accuracy of 90%.
Convolutional Convolutional Recurrent Deep Learning Model for Sentence Classification [4]	ABDALRAOUF HASSAN and AUSIF MAHMOOD	In this paper authors have used IMDB review dataset and SS-Tweet dataset to classify sentences into positive, negative and neutral, They too have used a combined model of CNN + LSTM and got an accuracy of 90% and 49% respectively but amount of data used by them is more.

Exploring Deep Recurrent Convolution Neural Networks for Subjectivity Classification [5]	XUEJUN ZHANG SHAN HUANG , JIANQIANG ZHAO , XIAOGANG DU1 , FUCUN HE	In this paper authors have used SS-Tweet dataset to classify sentences to positive, negative and neutral. They have used separate model of CNN and BI-LSTM and obtained an accuracy of 82% and 79% respectively.
Sentiment analysis in a cross-media analysis framework [6]	Yonas Woldemariam	This paper presents the integral analysis of sentiments data of a linear sequence of specialized modules into the cross-media analytical frameworks which are actually open source. The modules or pipeline consists the following parts: chat room cleaner, NLP and sentiment analyzer. The author in his experiment used the Apache-Hadoop framework along with it lexicon-based forecast algorithm which is based on lexicon_sentiments. The sentiments are also based on a library of Stanford coreNLP that uses Recursive Neural Tensor Network (RNTN) model. The whole performance evaluation displays that RNTN performs better the lexicon-based by 9.88% on accuracy. Also, they showed that lexicon-based predicts efficiently on classifying positive comments.
Apply word vectors for sentiment analysis of APP reviews [7]	Xian Fan; Xiaoge Li; Feihong Du; Xin Li; Mian Wei	Vector representations have been utilized for Natural Language processing tasks. Authors here targeted on utilizing the efficiency of word vector representations for providing the solution to sentiment analysis problem. Three tasks, of retrieval of sentiment words, the polarity of sentiment words identification, and forecasting text sentiment, have been given the primary importance. They scrutinized the

		potency of vector representations over unique text data and checked the quality of vectors depending upon different domains. The representations have been also used to calculate various vector-based features to provide and check effectiveness. They state that they have achieved F1_score and the accuracy to be 85.77% and 86.35% respectively for text sentiment analysis for APP reviews.
A Deep Neural Network Model for Target-based Sentiment Analysis [8]	Siyuan Chen; Chao Peng; Linsen Cai; Lanying Guo	They propose a deep neural network model with the combination of combining convolutional neural network and regional long short-term memory (CNN-RLSTM). The main aim is of performing target based analysis of sentiments. The authors took the approach as a time saving as it can minimize the time of learning for neural network model through a regional LSTM. Also, the CNN-RLSTM utilizes a sentence-level CNN for retrieving sentiment features of the complete sentence and regulates the transfer of data through various weighted matrices. These matrices particularly deduce the sentiment poles of various targets in the sentence itself. The experiment is carried out on multi-domain datasets of two languages from SemEval2016 and auto data. They concluded that their approach gives better performance than SVM and some other neural network models.

Sentiment analysis: Arabic sentiment lexicons [9]	Khaled S. Sabra; Rached N. Zantout; Mohamad A. El Abed; Lama Hamandi	Various lexicons are there to perform the task, for recognizing sentiment from a text written in a natural language with respect to the entity it is mapping to, in English utilizing WordNet. In this paper, they presented a unique process to build a sentiment lexicon for the Arabic language. They used semi-supervised training on the WordNet and matching them with an Arabic database.
A Sentiment Analysis Method of Short Texts in Microblog [10]	Jie Li; Lirong Qiu	Old approaches for sentiment analysis of short text lack the dependence of emotion words and modifiers and simply collect the sentiment of the sentence to seek the sentiment of short text. In this paper, they managed to mitigate the difficulties through sentiment structure and the sentiment computation norms. In the paper, the proposed approach shows how the dependency parsing deduces the sentiment structure with the relational migration and adjusted distance, which provides good contribution to knowing the sentiment of short text. The sentiment of short text is gathered as per the distinct influence of mappings between the modifier and the emotion word. Their experiment results ensure the effectiveness of the approach they actually proposed for mitigating the problems through sentiment structure.
Geo-Spatial Multimedia Sentiment Analysis in Disasters [11]	Abdullah Alfarrarjeh; Sumeet Agrawal; Seon Ho Kim; Cyrus Shahabi	Some sentiment analysis of disaster-related posts in social media can contribute to the state knowledge and deeper learning of the dynamics of disaster events by recognition of the polarity of sentiments given by people. But according to the authors, there still exist

		some limitations and there is no standard and reliable method which can be adopted in disasters. Utilizing this problem of current state-of-the-art sentiment classifiers, the paper presents the framework for geospatial sentiment analysis of disaster-concerned social media data entity. The framework addresses three types of difficulties: the inaccuracy and discrepancy associated with different text and image sentiment classifiers, the geo-sentiment discrepancy among data entities in a confined geographical area, and noting manifold sentiments from multimedia data objects (i.e., text and image). They explored Twitter and Flickr datasets at the time of Hurricane Sandy and Napa Earthquake for evaluating the used framework and came up with the approach that presents a better comprehension of disaster eventuality
Sentiment analysis of a document using deep learning approach and decision trees [12]	Arman S. Zharmagambetov; Alexandr A. Pak	This paper depicts a recent approach to the aim of sentiment analysis of movie reviews by utilizing deep learning recurrent neural networks algorithms like decision trees. This technique depends on statistical models, which brief the machine learning algorithms. The productive area of their work is the application of Google's algorithm Word2Vec given by Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean in 2013. The core ideology of Word2Vec is the representations of words with the help of vectors in such a manner that semantic dependencies between words be intact as basic linear algebra operations. The competency of the mentioned

		algorithm is that the secondary is computational efficiency. The aim of this paper is about making use of Word2Vec model for text classification according to their sentiment type.

Table 1 Integrated summary of literature survey

1.4 PROBLEM STATEMENT

To classify twitter tweets into positive, negative and neutral classes on four datasets SS-Tweet, HCR, US Airline and Whatsapp dataset using few–short learning based hybrid model of convolution neural networks and bidirectional long-short term memory model.

1.5 OVERVIEW OF PROPOSED SOLUTION APPROACH AND NOVELTY

Usually we have seen various works on sentiment analysis but none of them have used little short learning technique for classification. We achieved good accuracy results rather considering short data, also the deep learning models applied are generally very computationally expensive but due to our short data we were able to apply them on normal machine.

Chapter 2: UPDATED LITERATURE REVIEW

[1] In this paper authors have used few-short learning for short text classification i.e. amount of data used is very less and they have applied several machine learning and deep learning techniques to classify text and got maximum accuracy of 88% from Siamese CNN. Datasets used by them are SS-Tweet, HCR Sentiment-B 2013 and Multigame. [2] In this paper authors have used amazon reviews and classified them as positive or negative, they have used a total of 12000 reviews to train. They have used CNN and LSTM separately and obtained an accuracy of 87% and 88% respectively. [3] In this paper authors have used twitter dataset and classified them on basis of age as Teenager, Adult and General. They have used various deep learning and machine learning algorithms and have obtained an accuracy of 90%. [4] In this paper authors have used IMDB review dataset and SS-Tweet dataset to classify sentences into positive, negative and neutral, they too have used a combined model of CNN + LSTM and got an accuracy of 90% and 49% respectively but amount of data used by them is more. [5] In this paper authors have used SS-Tweet dataset to classify sentences to positive, negative and neutral. They have used separate model of CNN and BI-LSTM and obtained an accuracy of

82% and 79% respectively. [6] This paper presents the integral analysis of sentiments data of a linear sequence of specialized modules into the cross-media analytical frameworks which are actually open source. The modules or pipeline consists the following parts: chat room cleaner, NLP and sentiment analyzer. The author in his experiment used the Apache-Hadoop framework along with it lexicon-based forecast algorithm which is based on lexicon_sentiments. The sentiments are also based on a library of Stanford coreNLP that uses Recursive Neural Tensor Network (RNTN) model. The whole performance evaluation displays that RNTN performs better the lexicon-based by 9.88% on accuracy. Also, they showed that lexicon-based predicts efficiently on classifying positive comments. [7] In this paper Vector representations have been utilized for Natural Language processing tasks. Authors here targeted on utilizing the efficiency of word vector representations for providing the solution to sentiment analysis problem. Three tasks, of retrieval of sentiment words, the polarity of sentiment words identification, and forecasting text sentiment, have been given the primary importance. They scrutinized the potency of vector representations over unique text data and checked the quality of vectors depending upon different domains. The representations have been also used to calculate various vector-based features to provide and check effectiveness. They state that they have achieved F1_score and the accuracy to be 85.77% and 86.35% respectively for text sentiment analysis for APP reviews. [8] In this paper they propose a deep neural network model with the combination of combining convolutional neural network and regional long short-term memory (CNN-RLSTM). The main aim is of performing target based analysis of sentiments. The authors took the approach as a time saving as it can minimize the time of learning for neural network model through a regional LSTM. Also, the CNN-RLSTM utilizes a sentence-level CNN for retrieving sentiment features of the complete sentence and regulates the transfer of data through various weighted matrices. These matrices particularly deduce the sentiment poles of various targets in the sentence itself. The experiment is carried out on multi-domain datasets of two languages from SemEval2016 and auto data. They concluded that their approach gives better performance than SVM and some other neural network models. [9] In this paper various lexicons are there to perform the task, for recognizing sentiment from a text written in a natural language with respect to the entity it is mapping to, in English utilizing WordNet. In this paper, they presented a unique process to build a sentiment lexicon for the Arabic language. They used semi-supervised training on the WordNet and matching them with an Arabic database. [10] In this paper old approaches for sentiment analysis of short text lack the dependence of emotion words and modifiers and simply collect the sentiment of the sentence to seek the sentiment of short text. In this paper, they managed to mitigate the difficulties through sentiment structure and the sentiment computation norms. In the paper, the proposed approach shows how the dependency parsing deduces the sentiment structure with the relational migration and adjusted distance, which provides good contribution to knowing the sentiment of short text. The sentiment of short text is gathered as per the distinct influence of mappings between the modifier and the emotion word. Their experiment results ensure the effectiveness of the approach they actually proposed for mitigating the problems through sentiment structure. [11] In this paper some sentiment analysis of disaster-related posts in social media can contribute to the state knowledge and deeper learning of the dynamics of disaster events by recognition of the polarity of sentiments given by people. But according to the authors, there still exist some limitations and there is no standard and reliable method which can be adopted in disasters. Utilizing this problem of current state-of-the-art sentiment classifiers, the paper presents the framework for geospatial sentiment analysis of disaster-concerned social media data entity. The framework addresses three types of difficulties: the inaccuracy and discrepancy associated with different text and image sentiment classifiers, the geo-sentiment discrepancy among data entities in a confined geographical area, and noting manifold sentiments from multimedia data objects (i.e., text and image). They explored Twitter and Flickr datasets at the time of Hurricane Sandy and Napa Earthquake for evaluating the used framework and came up with the approach that presents a better comprehension of disaster eventuality. [12] This paper depicts a recent approach to the aim of sentiment analysis of movie reviews by utilizing deep learning recurrent neural networks algorithms like decision trees. This technique depends on statistical models, which brief the machine learning algorithms. The productive area of their work is the application of Google's algorithm Word2Vec given by Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean in 2013. The core ideology of Word2Vec is the representations of words with the help of vectors in such a manner that semantic dependencies between words be intact as basic linear algebra operations. The competency of the mentioned algorithm is that the secondary is computational efficiency. The aim of this paper is about making use of Word2Vec model for text classification according to their sentiment type.

Chapter 3: ANALYSIS, DESIGN AND MODELING

3.1 DATASETS

3.1.1 HCR DATASET

HCR stands for Health Care Reform. The dataset contains tweets information which is annotated by language and computer class. The author of dataset originally included five labels of sentiments tweets namely positive, negative, neutral, and irrelevant and unsure. But the preprocessing of the dataset in our experiment is done after removing irrelevant and unsure tweets. The preprocessed dataset comprises 2394 tweets (470 neutral, 1382 negative, 542positive) [1]. We have used 1697 tweets for training, 189 tweets for validating and 629 tweets for testing.

3.1.2 SS-TWEET DATASET

SS -TWEET stands for Sentiment Strength Twitter Dataset. This dataset is annotated by the manual method. It contains 4242 tweets. The wall et al was the first person to build this dataset for evaluating SentiStrenth. He proposed re-annotation of tweets of this dataset with negative (a number 2), neutral (a number 0) and positive (a number 1), instead of sentiment strengths [1]. The dataset is used in our experiment after preprocessing and keeping only some labels. Mean positive, mean negative and classes (negative, neutral and positive) are main labels. We have used 2811 tweets for training, 313 tweets for validating and 1043 tweets for testing.

3.1.3 US AIRLINE DATASET

US Airline datasets consists of 4040 sentences out of which we have trained on 2727 tweets, validated on 303 tweets and tested on 1010 tweets. Tweets are classified into positive, negative and neutral classes. Dataset is obtained from https://data.world/socialmediadata/twitter-us-airline-sentiment .

3.1.4 WHATSAPP DATASET

Whatsapp dataset consists of 1531 chats classified into Very Satisfied, Satisfied, Neutral Unsatisfied and Very Unsatisfied. We have used 1033 chats for training, 115 chats for validating and 383 chats for testing. Dataset is obtained from https://www.kaggle.com/weywenn/sentiment-analysis-multilanguage

3.2 OVERALL ARCHITECTURE

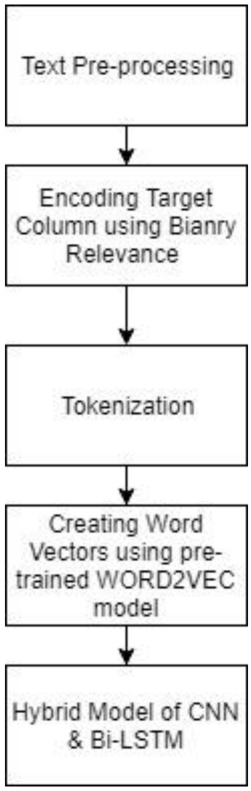

Fig 1: Flow Diagram

3.2.1 Text Pre-Processing

Text pre-processing includes techniques to remove unnecessary data from dataset. Using regular expressions we removed special characters, unwanted spaces, N.A values and numerical data

3.2.2 Encoding Target Columns using Binary Relevance

Next step is to encode our predicted column or target column so that we can convert them into classes for this we used a technique called binary relevance. In this technique all the classes are separated into different columns and the each sentence according to its class is assigned value 1 and for all others a value of 0.

3.2.3 Tokenization

Tokenization is an important step to convert sentences i.e. textual data into numerical form for this each word is assigned a unique number so that it can be processed into the model. Each sentence is converted to maximum of 100 tokens.

3.2.4 Creating word vectors using pre trained – WORD2VEC model

Next step is to create an embedding matrix which would be used by our hybrid model for weights calculation and changing them. Tokens generated form above step are used to create embedding matrix with tokens as rows and their corresponding feature vector is collected from pre-trained word2vec model thus this creates an embedding matrix with size (?,300)

3.2.5 Hybrid model of CNN and Bi-LSTM

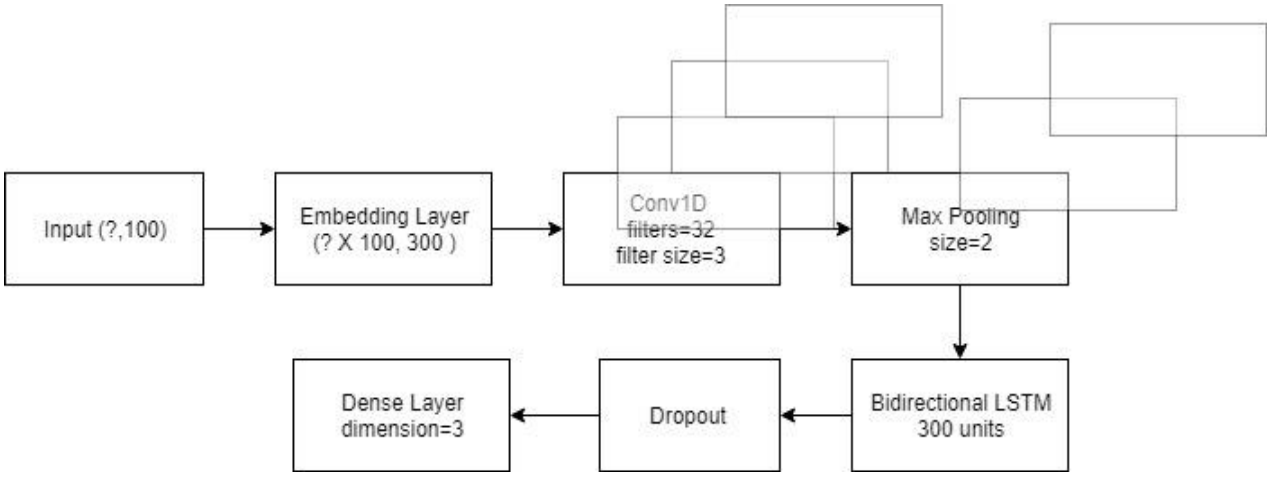

Fig 2: Proposed Model

3.2.5.1 Input Layer

Input Layer consists of sentences with maximum tokens length = 100. Each word in a sentence is converted to a token with a unique integer id and thus each sentence forms a vector of maxsize=100.

3.2.5.2 Embedding Layer

Embedding Layer consists of weights which are extracted from pre-trained word2vec model. Each word in sentences is matched with corresponding vector in pre-trained model and thus forms a vector called embedding vector with size – 300. Thus combining all words in the corpus an embedding matrix is formed which is used to train the model.

3.2.5.3 Conv1D Layer

First step is to introduce a 1D convolutional layer with 32 filters each of size (3,300) and thus each filter is used to produce a new matrix using convolution formulae defined below.

3.2.5.4 Max Pooling

In Max Pooling a filter of size (2,300) is used to produce a single output having the maximum value, thus the most important feature is considered.

3.2.5.5 Bidirectional LSTM Layer

Them a bidirectional long short term memory layer is used with 300 memory units in both directions. Using Bidirectional LSTMs, you feed the learning algorithm with the original data once from beginning to the end and once from end to beginning. There are debates here but it usually learns faster than one-directional approach although it depends on the task.

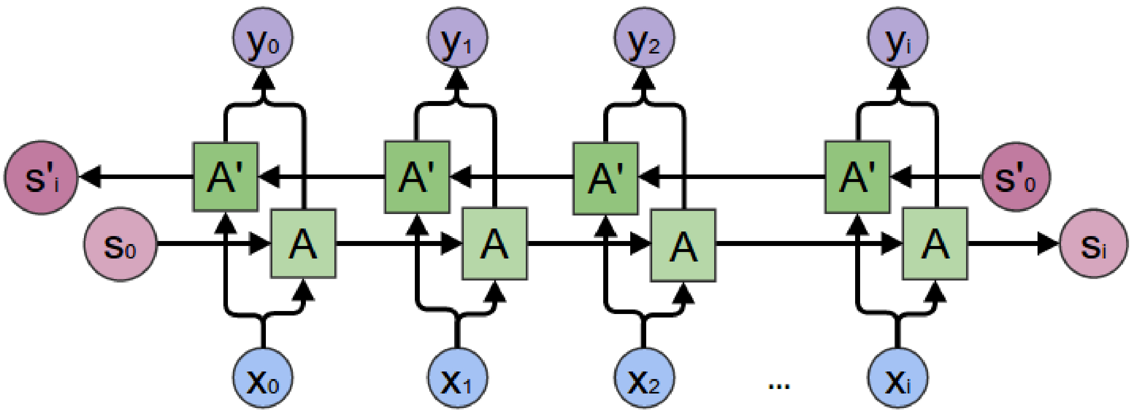

Fig 3 Basic model of a bidirectional -LSTM

3.2.5.6 Dropout

Dropout is used to drop 10% neurons created so as to avoid overfitting in the model.

3.2.5.7 Dense Layer

Dense Layer is used as an activation functional layer with activation function being rectified linear unit (ReLU) and sigmoidal.

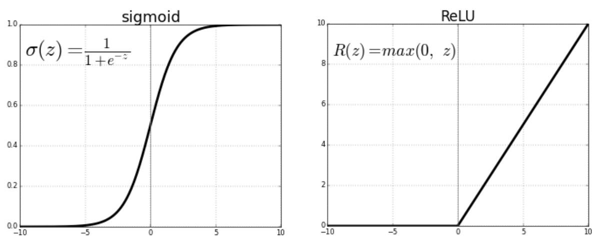

Fig 4 Activation Functions

Output dimensions used were 3 as output classes were positive, neutral and negative. Thus dense layer is an important functional layer which gives the required dimensional output.

3.2.6 Evaluation Parameters

Accuracy: Ratio of correctly predicted observations to the total number of observations is known as the accuracy of the model. A model is considered best with the highest accuracy. The formula calculating accuracy is as follows:

$$Accuracy = \frac{(TP+TN)}{TP+TN+FP+FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

3.3 DESIGN DOCUMENTATIONS

3.3.1 USE CASE DIAGRAM

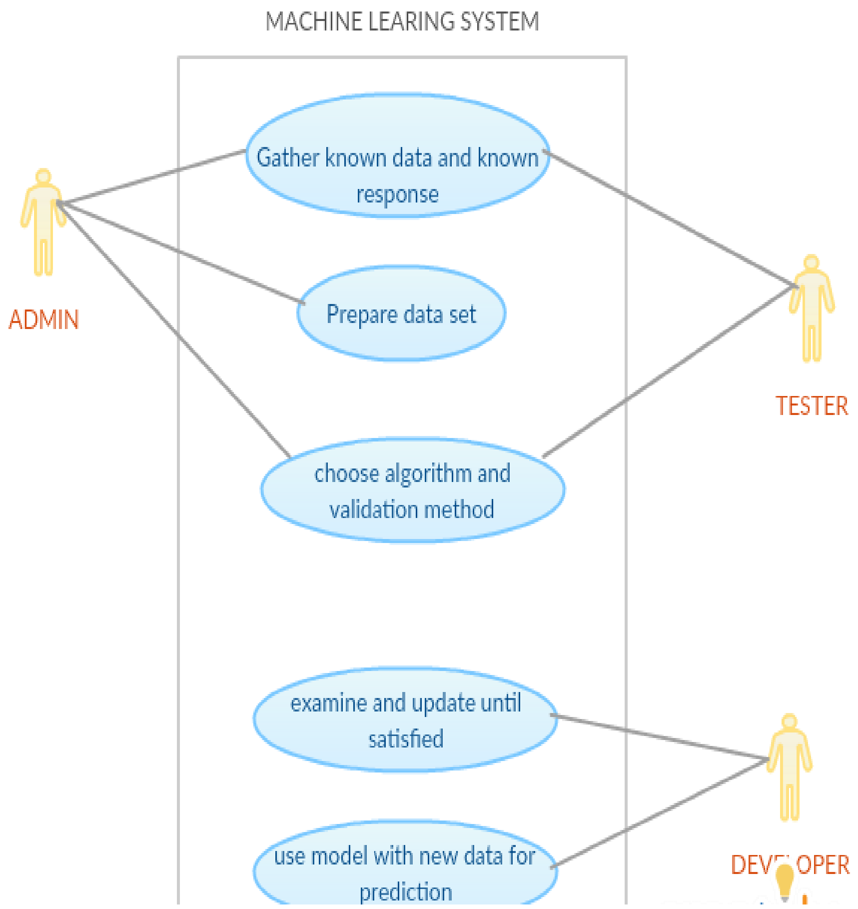

Fig 5 Use Case Diagram

3.3.2 ALGORITHMS

3.3.2.1 LOGISTIC REGRESSION

Logistic regression are one of the famous algorithms in machine learning it's easy to implement and also powerful. This algorithm is similar to linear regression Main difference between linear and logistic is that linear is used to predict values and logistic is used to classify values . Classification can be used on various applications such as classifying whether emails are spam or not, classifying whether website are fraud or not. The algorithm can be used for multi variables and binary variables too.

3.3.2.2 NAÏVE BAYES

Naïve Bayes is one of the most powerful algorithms for classification used for classifying data on basis of probabilities. With millions of records also this algorithms works awesomely. It simply works on Bayes theorem and uses various probabilities to classify data. Below if the formulae for probability.

3.3.2.3 SUPPORT VECTOR MACHINE

SVM (Support vector Machine) is an efficient supervised machine meaning algorithm used for classification and regression purposes. This algorithm draws a hyper-plane to separate classes. This algorithm works extremely well with regression, the effect of SVM increases as we increase dimensional space. SVM also perform good when dimension number is larger than the sample number [12]. There exists a draw back too it does not perform well on huge datasets.

3.3.2.4 DECISION TREE

Decision Tree algorithm is one of the supervised learning algorithms. Unlike other supervised machine learning algorithms, decision tree algorithm can be deployed for solving problem like regression and classification. The core idea of selecting Decision Tree in our experiment is to build a training model that can perform prediction class or value of target variables. Deeper insight would give an idea that this is done by remembering decision rules inferred from prior data (training data). ID3 is the core algorithm for creating decision trees

3.3.2.5 K NEAREST NEIGHBOR

Predictive Problems like classification and regression can find the solution under KNN. However classification problems can have maximum utilization of KNN in the industry. In KNN predictions are given directly using the training data. Predictions are made for a new instance (x) by iterating through the whole training set for the K most likely instances (the neighbors) and concluding the output variable for those K instances

3.3.2.6 ADABOOST CLASSIFIER

A well liked boosting technique which let you unite multiple “weak classifiers” into a single “strong classifier”. A classifier that functions undesirable, but performs better than random estimate or guessing is a nature of weak classifier. AdaBoost initializes “weight” to every training example, which finds the probability of each example appearing in the training set.

3.3.2.7 EXTRA TREE

Extra Trees stands for extremely randomized trees .This algorithm is another label of ensemble methods particularly created for decision tree classifiers. Original learning sample gives the branch to every sub extra tree created. In our experiment we have use the concept of K random splits at each test node for successfully achieving the best split is determined .In this process each one is found by a randomized selection of an input (without replacement) and a minimum selection called threshold

3.3.2.8 GRADIENT BOOST MACHINE

Regression and classification problems are even solved by Gradient boosting machine learning technique which generates a prediction model in the form of an ensemble of loose prediction models, typically decision trees. The focus of any supervised learning algorithm is to state a loss function and reduce it .Computational formula for Gradient Boosting algorithm.

3.3.2.9 XGB CLASSIFIER

Xgb classifier also knows as gradient boosting, adaptive regression tree or stochastics gradient algorithms. In this algorithm new model is created by using updation and error correction from previous model. Models keep on adding until updation become constant. One advantage of this algorithm is that it can be used both on classification and regression

3.3.2.10 RANDOM FOREST

Random Forest is a tree based classification model which is basically used for supervised classification. In this algorithm generally more trees correspond to better performance and efficiency.

3.3.2.11 HYBRID MODEL OF CNN & Bi-LSTM

We have used a hybrid model of conv1d layer, maxpool layer and bidirectional long-short term memory model with various parameters. Since this model is a deep learning model hence it gives extremely awesome results and also is easy computer on our machines despite being computationally expensive.

Chapter 4: IMPLEMEMENTATION AND RESULTS

4.1 CURRENT NEW RESULTS

For each dataset 2 epochs are used to train data because further if the epochs were increased the it gave better training accuracy but poor testing accuracy hence as epochs were increased beyond 2 the model began to over fit.

DATASET	TRAINING ACCURACY (%)	VALIDATION ACCURACY (%)	TESTING ACCURACY (%)
SS-TWEET Dataset	67.21	67.52	66.38
Health Care Reform (HCR) Dataset	68.3	80.2	71.91
US Airline Tweets Dataset	76.7	76.02	76.37
Whatsapp Chat Dataset	79.9	80.0	80.0

Table 2 CNN+Bi-LSTM hybrid model results

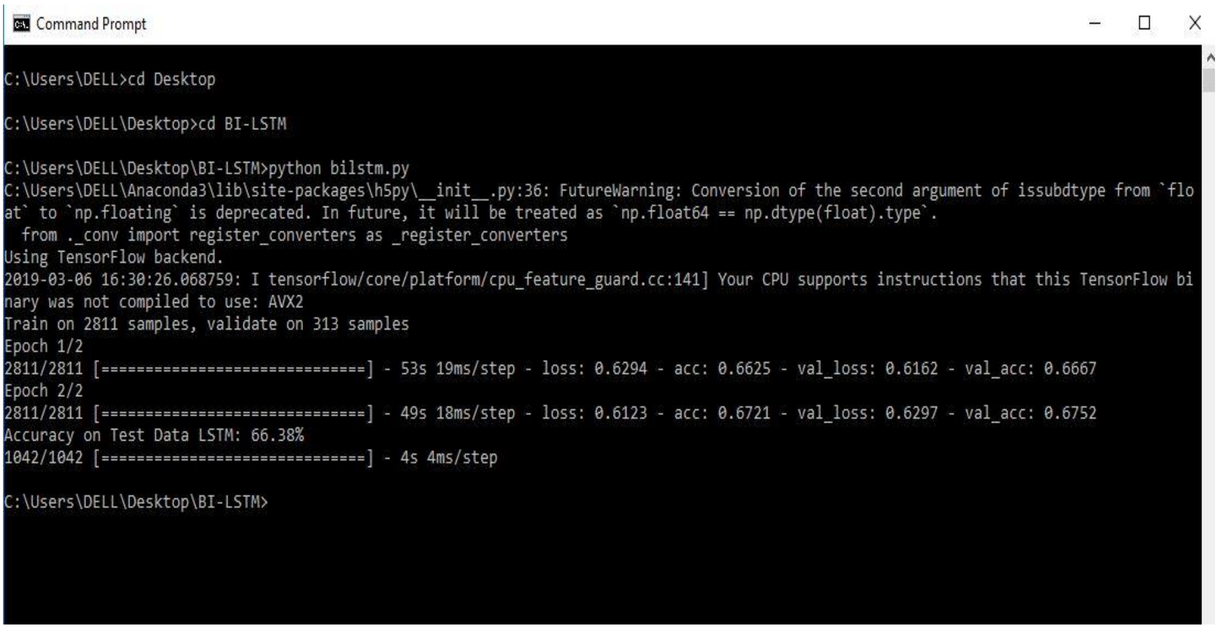

Fig 5: SS-Tweet result

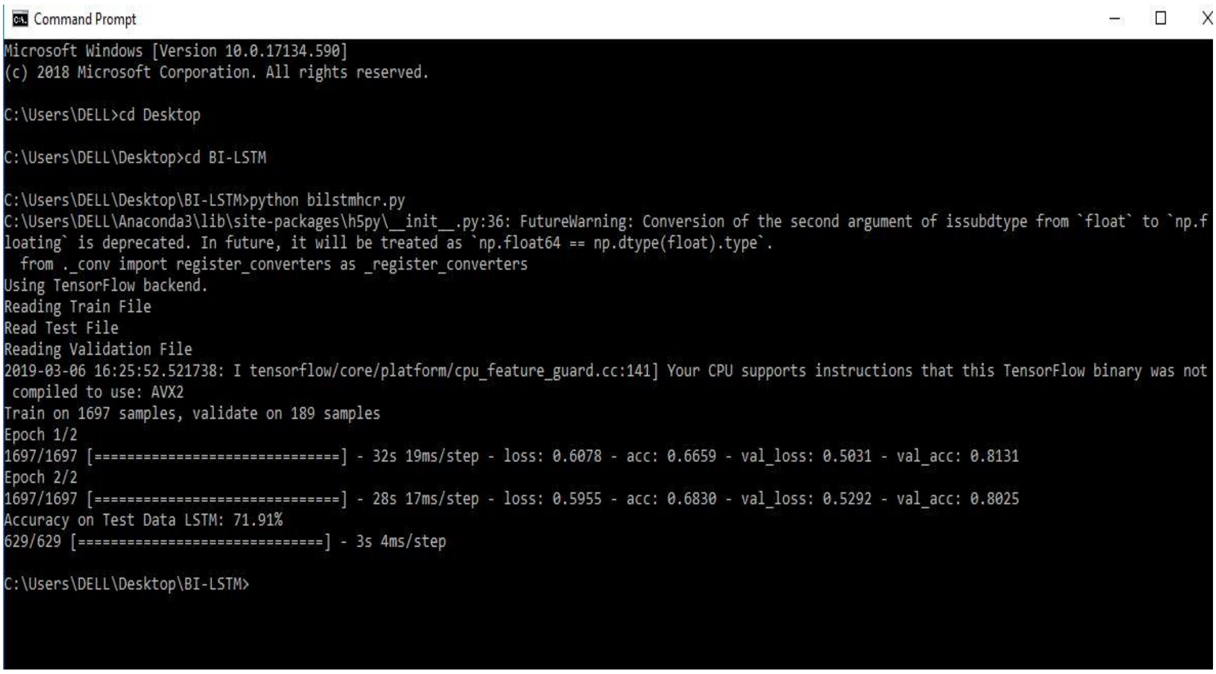

Fig 6: HCR result

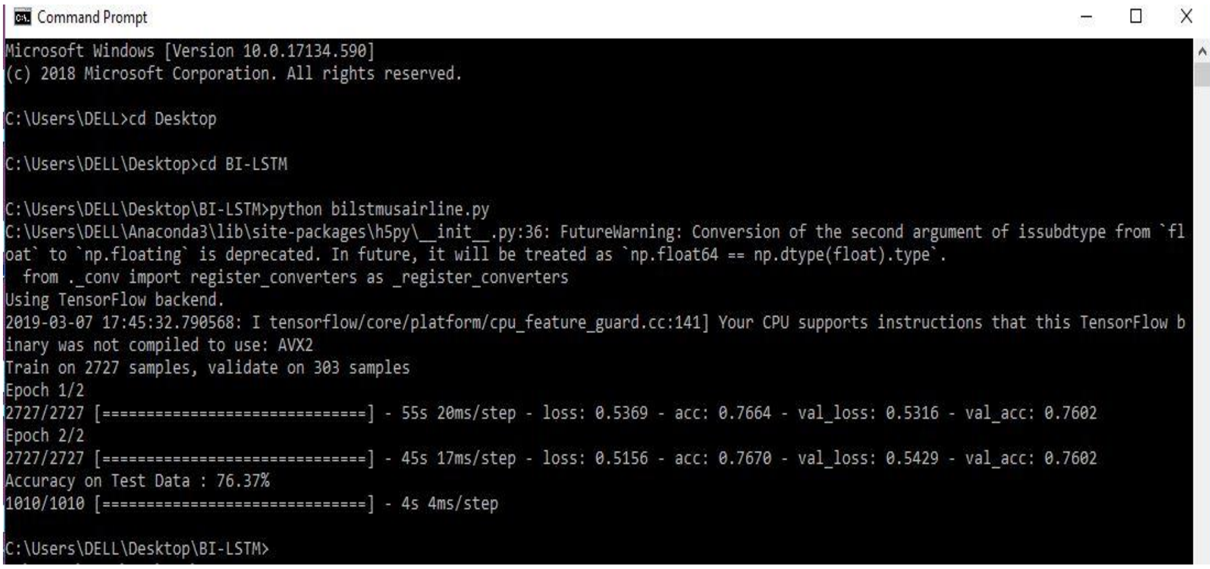

Fig 7: US Airline result

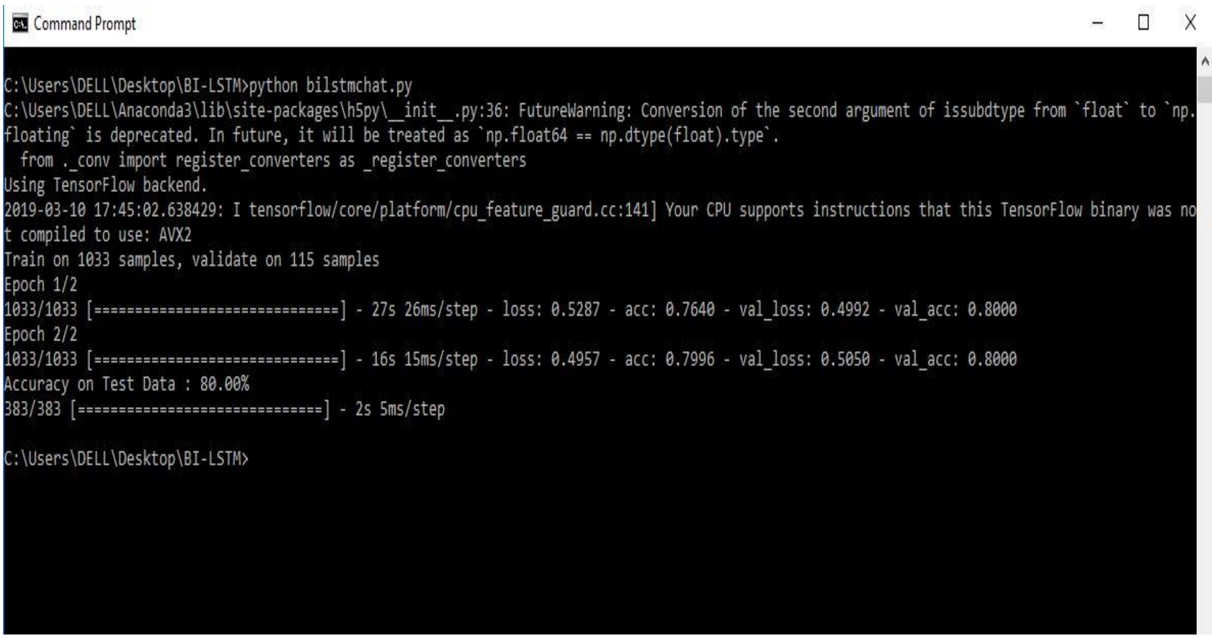

Fig 8: Whatsapp chat result

4.2 PREVIOUS RESULTS

SS-Tweet Dataset - (COUNT VECTORS)				
ML Algorithms	Accuracy (%)	Precision (%)	Recall (%)	Score (%)
KNN	59	43	39	35
Decision Tree	47	43	42	42
Extra Tree	52	50	46	46
GBM	54	55	48	48
XGB	56	57	48	48
SVM	46	15	33	22
AddaBoostClassifier	55	55	48	48
Logistics Regression	58	57	53	54
Naïve Bayes	54	53	54	53
Random Forest	52	50	46	46

Table 3: SS-Tweet results through count vectorization

SS-Tweet Dataset - (WORDLEVEL TF-IDF)				
ML Algorithms	Accuracy (%)	Precision (%)	Recall (%)	Score (%)
KNN	46	32	33	21
Decision Tree	46	43	42	42
Extra Tree	51	48	46	46
GBM	54	53	48	49
XGB	55	56	47	47
SVM	46	15	33	21
AddaBoostClassifier	55	55	49	49
Logistics Regression	57	57	50	50
Naïve Bayes	53	56	44	42
Random Forest	51	47	44	44

Table 4: SS-Tweet results through wordlevel tf-idf

SS-Tweet Dataset - (N-GRAMS VECTORS)				
ML Algorithms	Accuracy (%)	Precision (%)	Recall (%)	Score (%)
KNN	46	41	34	26
Decision Tree	48	44	42	42
Extra Tree	53	51	46	46
GBM	49	46	41	39
XGB	49	53	38	42
SVM	46	15	33	22
AddaBoostClassifier	50	52	38	33
Logistics Regression	49	51	39	54
Naïve Bayes	50	52	41	38
Random Forest	51	49	43	42

Table 5: SS-Tweet results through n-gram tf-idf

SS-Tweet Dataset - (CHARLEVEL VECTORS)				
ML Algorithms	Accuracy (%)	Precision (%)	Recall (%)	Score (%)
KNN	50	47	47	46
Decision Tree	45	42	42	42
Extra Tree	52	48	46	46
GBM	52	50	48	48
XGB	55	53	49	49
SVM	46	15	33	22
AddaBoostClassifier	55	53	50	50
Logistics Regression	59	57	53	53
Naïve Bayes	56	56	48	48
Random Forest	51	47	45	44

Table 6: SS-Tweet results through char-level tf-idf

HCR Dataset – (COUNT VECTORS)				
ML Algorithms	Accuracy	Precision	Recall	Score
KNN	60	56	57	56
Decision Tree	49	41	39	40
Extra Tree	25	25	34	34
GBM	65	55	57	54
XGB	59	60	57	56
SVM	62	34	51	48
AddaBoostClassifier	63	60	56	57
Logistics Regression	62	62	58	50
Naïve Bayes	35	37	30	31
Random Forest	63	59	51	52

Table 7: HCR results through count vectorization

HCR Dataset – (WORDLEVEL TF-IDF VECTORS)				
ML Algorithms	Accuracy	Precision	Recall	Score
KNN	36	35	35	35
Decision Tree	39	21	29	20
Extra Tree	57	52	47	49
GBM	52	50	47	49
XGB	48	46	47	45
SVM	61	43	50	41
AddaBoostClassifier	61	20	26	27
Logistics Regression	60	61	55	49
Naïve Bayes	25	33	29	20
Random Forest	61	62	50	51

Table 8: HCR results through wordlevel tf-idf

HCR Dataset – (N-GRAM VECTORS)				
ML Algorithms	Accuracy	Precision	Recall	Score
KNN	35	25	25	25
Decision Tree	49	41	39	40
Extra Tree	57	52	47	49
GBM	35	45	54	54
XGB	58	60	57	56
SVM	55	34	51	48
AddaBoostClassifier	63	60	56	57
Logistics Regression	23	21	29	25
Naïve Bayes	22	33	29	20
Random Forest	60	50	50	51

Table 9: HCR results through n-gram tf-idf

HCR Dataset – (CHARACTERLEVEL VECTORS)				
ML Algorithms	Accuracy	Precision	Recall	Score
KNN	60	56	57	56
Decision Tree	19	11	19	14
Extra Tree	56	52	47	49
GBM	52	50	47	49
XGB	54	54	54	54
SVM	38	34	31	38
AddaBoostClassifier	23	56	56	57
Logistics Regression	60	61	55	49
Naïve Bayes	21	33	29	20
Random Forest	63	55	51	52

Table 10: HCR results through char level tf-idf

CHAPTER 5: CONCLUSION AND FUTURE SCOPE

5.1 FINDINGS

PREVIOUS RESULTS		CURRENT RESULTS			
SS-TWEET (Highest Accuracy)	HCR (Highest Accuracy)	SS-TWEET	HCR	US AIRLINE	WHATSAPP CHAT
59% by KNN & count vectorization	65% by GBM & count vectorization	66.3%	71.91%	76.3%	80.0%

Table 11 Comparison from previous results.

5.2 CONCLUSION

Previously we applied 10 ML algorithms on two datasets SS-tweet and HCR to classify tweets into positive, negative or neutral classes. Using few-short learning technique and different feature extraction techniques we obtained highest accuracy of 59% for SS-tweet and 65% for HCR. In order to improve our results we now applied an hybrid model of convolutional neural network (CNN) and Bidirectional long-short term memory (Bi-LSTM) and saw an significant increase in our results by around 10%. Hence by applying deep learning model we obtained good results. Main uniqueness in our results is that we have used few short learning technique and are applying deep learning models which require heavy computational power on our regular machines.

5.3 FUTURE WORK

Our future scope is to further bring our accuracy above 80 % as most of the paper shows by using different hybrid models and feature extraction techniques such as SURF/SIFT which brings invariance. Our final work would be unique in the sense that we have used very few data to classify sentiments and that to with great accuracy

APPENDIX

REFERENCES

[1]: Multimedia Tools and Applications, 2018, Volume 77, Number 22, Page 29799
Leiming Yan, Yuhui Zheng, Jie Cao

[2]: W. Zhao *et al.*, "Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 1, pp. 185-197, 1 Jan. 2018.
Doi: 10.1109/TKDE.2017.2756658

[3]: R. G. Guimarães, R. L. Rosa, D. De Gaetano, D. Z. Rodríguez and G. Bressan, "Age Groups Classification in Social Network Using Deep Learning," in *IEEE Access*, vol. 5, pp. 10805-10816, 2017.
Doi: 10.1109/ACCESS.2017.2706674

[4]: A. Hassan and A. Mahmood, "Convolutional Recurrent Deep Learning Model for Sentence Classification," in *IEEE Access*, vol. 6, pp. 13949-13957, 2018.
Doi: 10.1109/ACCESS.2018.2814818

[5]: X. Zhang, S. Huang, J. Zhao, X. Du and F. He, "Exploring Deep Recurrent Convolution Neural Networks for Subjectivity Classification," in *IEEE Access*, vol. 7, pp. 347-357, 2019.
Doi: 10.1109/ACCESS.2018.2885362

[6]: Y. Woldemariam, "Sentiment analysis in a cross-media analysis framework," *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, Hangzhou, 2016, pp.1-5.doi: 10.1109/ICBDA.2016.7509790

[7]: X. Fan, X. Li, F. Du, X. Li, and M. Wei, "Apply word vectors for sentiment analysis of APP reviews," *2016 3rd International Conference on Systems and Informatics(ICSAI)*, Shanghai,2016,pp.1062-1066.doi: 10.1109/ICSAI.2016.7811108

[8]: S. Chen, C. Peng, L. Cai, and L. Guo, "A Deep Neural Network Model for Target-based Sentiment Analysis," *2018 International Joint Conference on Neural Networks (IJCNN)*, RiodeJaneiro, 2018, pp.1-7.doi: 10.1109/IJCNN.2018.8489180

[9]: K. S. Sabra, R. N. Zantout, M. A. E. Abed and L. Hamandi, "Sentiment analysis: Arabic sentiment lexicons," *2017 Sensors Networks Smart and Emerging Technologies (SENSET)*, Beirut, 2017, pp.1-4.doi: 10.1109/SENSET.2017.8125054

[10]: J. Li and L. Qiu, "A Sentiment Analysis Method of Short Texts in Microblog," *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, Guangzhou,2017,pp.776-779.doi: 10.1109/CSE-EUC.2017.153

[11]: A. Alfarrarjeh, S. Agrawal, S. H. Kim, and C. Shahabi, "Geo-Spatial Multimedia Sentiment Analysis in Disasters," *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Tokyo, 2017, pp. 193-202.
Doi: 10.1109/DSAA.2017.77

[12] : A. S. Zharmagambetov and A. A. Pak, "Sentiment analysis of a document using deep learning approach and decision trees," *2015 Twelve International Conference on Electronics Computer and Computation (ICECCO)*, Almaty, 2015, pp. 1-4.
Doi: 10.1109/ICECCO.2015.7416902