Twitter Sentiment Analysis using LSTM Algorithm

Aniket Kale, Chetan Bawankule, Payal Singanjude, Ganesh Wattamwar, Dr. Simran Khiani

Department of Information Technology,

G H Raisoni College of Engineering & Management, Pune

ABSTRACT

Sentiment analysis refers to opinion mining and a machine learning task where one would like to work out what theoveral sentiment of a given document is. Using natural language processing and model training we will extract the subjective information of tweets froma particular dataset and checkitout to classify it consistent with its polarity like positive,neutral,or negative. The papers suggest a model created usingtheLSTM algorithmand some additional other layers of machinelearning for better sentiment analysis. Even though sentimentanalysisis really far away from being solved since the language is extremely complex because of objectivity/subjectivity, negation, vocabulary, grammar still this sequential model achieves the accuracy of approximately 81%. Thus , this work shows that there wide applications yet to be done.

Keywords—Twitter, Sentiment Analysis, Natural Language processing, Long Short Term Memory (LSTM)

I. INTRODUCTION

Twitter may be a modern public square where many voices discuss, debate, and share their views. Media personalities, politicians, and therefore the public address socialnetworksfor real-time information and reactions to the day's events. The opinion of individuals matters tons to research on how the propagation of data impacts the lives of a large-scale network like Twitter. Sentiment analysis of the tweets determines the polarity and inclination of the vast population towards aspecific topic, item, or entity. Sentiment Analysis also helps different consumer-centered industries to analyze people's opiniononа particular product or subject.

This project prefers to attempt to classify tweets fromTwitter into "positive" or "negative" sentiment by using Long Short Term Memory (LSTM) algorithm. The majority of people use Twitter including a large number of influencers and celebrities. Twitter is a microblogging website where people cansharetheir feelings quickly and spontaneously by sending a tweetlimited by 280 characters. Even you can directly address a tweet to someone by adding the target sign "@" or participate inatopic by adding a hashtag "#" to your tweet. Because of theusageof Twitter, it's an ideal source of knowledge to work out the present overall opinion about anything.

II. LITERATURE REVIEW

Sentiment analysis and Natural Language Processing have become a trending topic in the field of Artificial Intelligence and Machine Learning. Many researchers and technologistsare investing in progress towards sentiment analysis. Manytools and research are available for twitter sentiment analysis.In paper [1] authors created Bag of words for positive and negative analysis were in form of txt files were taken to compare the tweets and classify theminto positive, negative and neutral tweets. Authors Meylan Wongkar, Apriandy Angdresey in paper [2] did the text processing from data obtained and use Naive Bayes method to predict theclass.Also compared Naive Bayes with other methods such as SVM and KNN. The paper [2] also concluded that Naive Bayes method has a better accuracy level (i.e. 80.90%) compared to using other methods. Paper [3] proposed a sentiment analysis using deep learning techniques by the LSTM network model. They also analyze tweets data for the airline industry in socialmedia which showed better performance in the training dataset.They also suggested that, accuracy can be enhanced using the Bidirectional LSTM network. In paper [4] authors compared different predicting algorithms like LSTM-CNN, CNN, Multilayer CNN, Multilayer CNN decoder, CNN- LSTM, LSTM-MCNN, and LSTM-MCNN Decoder and concluded that Multilayer CNN decoder has highest level of accuracyfor predicting Twitter Sentiment Analysis. Authors used "Sentiment Classification Dataset" from Kaggle.comprovided by University of Michigan. Authors Aliza Sarlan, Chayanit Nadam, Shuib Basri in paper [5] categorized sentiment into positive and negative and represented it in pie chart and html page.

There are also different tools available over the Internet which provides different services regarding sentiment analysis. Bidirectional Encoder Representations from Transformers BERT system created by Google is great example which provides services not only for sentiment analysisbutalsohelps in predicting sentences and so on. In paper [6] authors Mark Cieliebak, Oliver Dürr and Fatih Uzdilli tested most of the available tools for sentiment analysis using 30,000 different short texts like tweets, news headings and reviews etc. Their study concluded that the average accuracy level of thesetools are approximately 60%.

Referring these works we concluded that we canuseLongShort Term Memory (LSTM) algorithm. It is heavy and efficient algorithmto analyze sentences because of its good memoryto remember previous parameters.

III. OVERVIEW OF THE DATA

The data for training the model is used from the site kaggle.com. The dataset is named as Sentimet140 dataset [7]. This dataset comprises of 1.6 million tweets extractedusingthe twitter API. The dataset is created by Marios Michailidis.This dataset consists of 6 fields i.e. target, ids, date, flag, user, text. The field target describes the polarity of the text as 0 –positive,

4- negative. The field ids is theid for a particular tweet. Field date shows the date on which tweet is generated. Field flag defines the query generated over the particular tweet.Userfield describes the user name of the person who tweeted. Last text field gives the total text in the tweet.

The Fig. 2 gives the sentiment data distribution as positiveand negative given in the dataset.

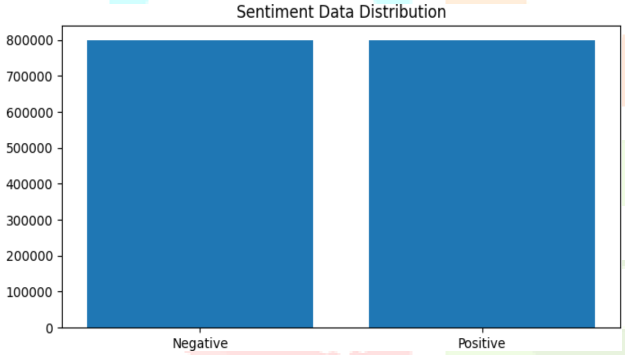


Fig 1. Sentiment Distribution in available dataset.

IV. PROPOSED SYSTEM

The main idea is to create sequential model which can predict sentiment of larger population using their tweets. This sequential model is created using LSTM algorithm to have accuracy more than the models that are already available.There are certain processes to be followed to analyze these tweets because tweets often consist of hyperlinks, user mentions, punctuation and emoji's. We shallnot let those text fortraining model. The processes that needs to be followedare:

- Stemming/Lemmatization: Stemming refers to a process that cut offthe ends of words to accomplish objective effectively. Lemmatization try to do things properly using vocabulary and morphologicalanalysis of words, it normally remove unnecessary endingsof word and to provide the actual formof a word as per the dictionary.

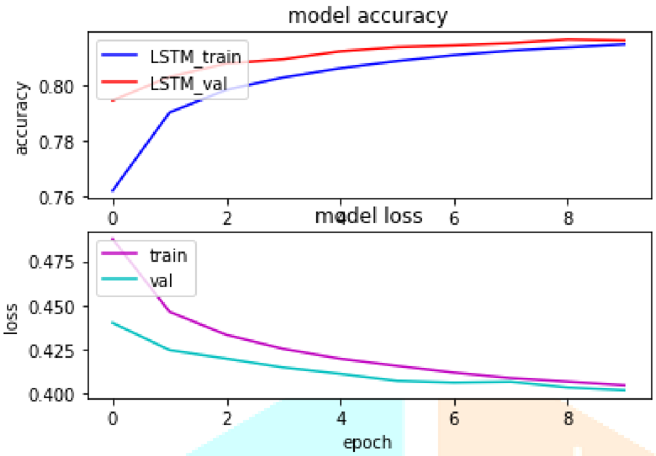
- Hyperlinks and Mentions: Tweets usually contain heaps of Hyperlinks and mentions of twitter handles, as they contain no sentiment we can remove them.

- Stopwords: Stopwords are regularly utilized wordsin English which have no logical importance in a sentence. So hence we eliminate them before classification.

- Tokenization: In tokenization sentence is choppedof in to pieces of word called tokens and at the sametime remove some character in particular punctuation.

- Word Embedding: This sequential modelintend to learn contextual meaning behind word to do that we have to do Word Embedding, it is capable to capture the meaning of word in particular statement. It also helps in semantic and syntactic similarity and relation with other words etc. It is nothing but the vector representation of words which we use for NLP.Weare using pre-defined word embedding library GloVe which gives additional insights for the word used for classification.

- Model Training: There are some words which are categorized in both positive and negative sentiment. This problem could not be addressed using SVD, Naïve Bayes etc. efficiently. To overcome this we are using Sequence Model using LSTMalgorithm.



In Sequence model architecture, we are implementing following layers.

1) Embedding Layer – This layer helps in creating vectors using the predefined library.

2) Conv1D Layer –This layer helps in validating data using vectors that are created in previous layer.

3) LSTM – LSTM stands for Long Short TermMemory, it's a type of RNN which hold memory state cell to learn the contextual meaning of words which further get transferredwith neighboring word unlike RNN where it only transfers neighboring word. LSTM work well than RNN when it comes to short term memory. LSTM cell is group of a state cell, an input gate, an output gate and a forget gate. The state cell remembers necessary values and forward it to the nextunitand the three gates are responsible for flow of the information into and out of the state cell.

4) Dense –This layer is a deeply connected neuralnetworkhelps in training model.

After the training is done, the model can be used for predicting sentiments on new tweets. This project is inspired from the works of Arun Pandian R. The Fig. 2 describes the sequential model architecture.

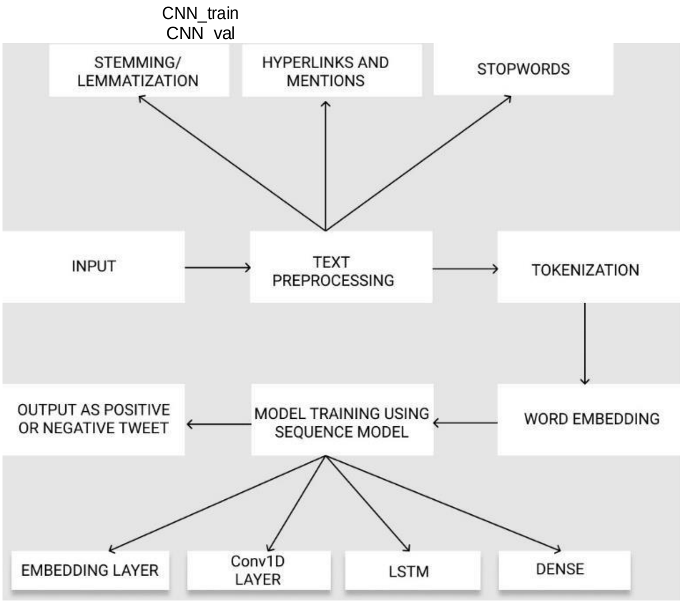


Fig. 2 System Architecture Diagram

V. RESULTS

As of now, the entire model is trained on only 1280000 tweets and being tested on 320000 tweets, using cross validationwhile training, accuracy is around 80.99%. The accuracy and loss in the sequential model can be shown in the given Fig3.

Fig 3. Sequential LSTM Model Accuracy – 80.99 %

Using the same data parameters with different models likeCNN and RNN prediction accuracy is much lower. AccuracyofCNN model is 58.69 % and accuracy of RNN model is 77.73%. Following Fig. 4 and Fig. 5 shows the accuracy for CNN and RNN model respectively.

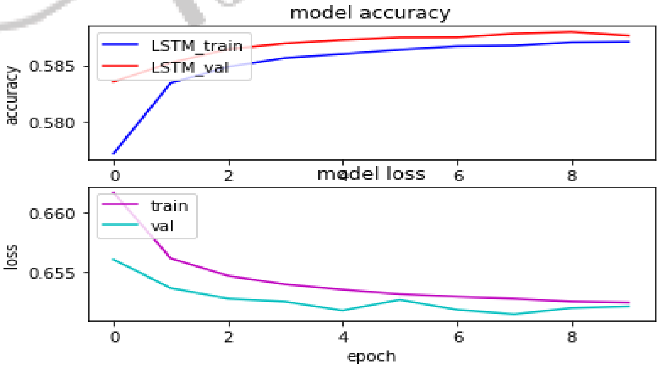


Fig. 4 CNN Model Accuracy - 58.69 %.

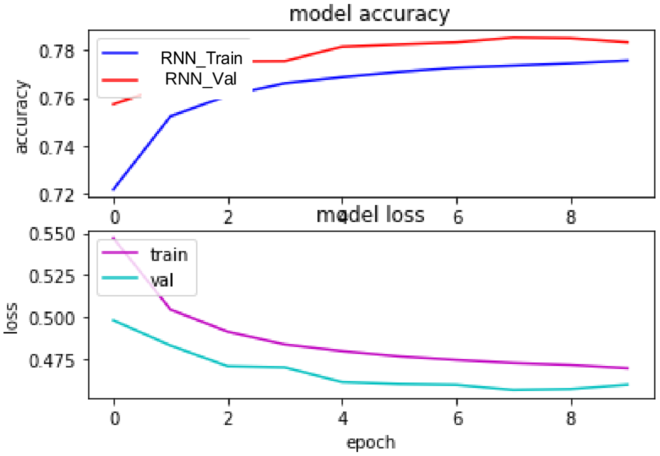


Fig. 5 RNN Model Accuracy – 77.73 %.

VI. ACKNOWLEDGEMENTS

This paper and the research behind it would not have been possible without the exceptional support of Prof. Simran Khiani. We are highly indebted to Prof. Simran Khiani for their guidance and constant supervision as well as for providing necessary information regarding the paper & also for their support in completing the paper.

We are likewise grateful to our institute for giving motivation to compose this paper.

References

[1] Prakruthi V, Sindhu D, Dr. S. Anupama Kumar, "Real Time Sentiment Analysis Of Twitter Posts", 3rd IEEEInternational Conference on Computational Systems and Information Technology for Sustainable Solutions 2018.

[2] Meylan Wongkar , Apriandy Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter", 2019 Fourth International Conference on Informatics and Computing (ICIC)**.**

[3] Ms.R.Monika, Dr.S.Deivalakshmi, Dr.B.Janet,

[6] Mark Cieliebak, Oliver Dürr, Fatih Uzdilli "Potential and Limitations of Commercial Sentiment Detection Tools",Zurich University of Applied Sciences.

[7] Sentiment140dataset from Kaggle https://www.kaggle.com/kazanova/sentiment140

VII. CONCLUSION AND FUTURE SCOPES

In this project, there can successful analysis of multipletweets showing their sentiments as positive or negative.Technologies such as HTML, CSS can be used for creating the user interface of the main project, MySQL/SQ-lite can be used for database management, and Natural Language processing is used for determining texts used in tweets. Efficient handling of tweets for the user, using a user-friendly interface can be done. Still, there is certain limitations regarding this project such as it works only in one language i.e English. Also, new data is generating rapidly and thus no one can rely on old data to predict today's people sentiment.

There are vast applications regarding this project like one can create API for this project. Also, these projects can be made available to predict regional languages in different countries, etc. As of now, this model is trained on Google GPU with more hardcore processors more accuracy can be achieved.

"Sentiment Analysis of US Airlines Tweets using LSTM/RNN", IEEE 9th International Conference on Advanced Computing (IACC) 2019**.**

[4] Nan Chen, Peikang Wang, "Advanced Combined LSTM- CNN Model for Twitter Sentiment Analysis", 5th IEEE International Conference on Cloud Computing and Inteligence Systems (CCIS) 2018.

[5] Aliza Sarlan, Chayanit Nadam, Shuib Basri, "Twitter Sentiment Analysis", International Conference on Information Technology and Multimedia (ICIMU) 2014.