TWITTER SENTIMENT ANALYSIS USING WORD2VEC AND TF-IDF WORD EMBEDDING

Nithya.B, Yeswanth.S- UG Scholar, Sri Venkateswara College of Engineering,
K.Kiruthika Devi Assistant Professor, Sri Venkateswara College of Engineering

Abstract - Sentiment Analysis of Twitter has attracted much attention recently due to its wide applications in both commercial and public sectors and also due to the rapid growth in Twitter's popularity as a platform for people to express their opinions and attitudes towards a great variety of topics. So analysing the twitter tweets accurately is very important. The proposed model makes use of an innovative combination of skip gram model ofword2vec and term frequency inverse document frequency values to generate feature vector which can then be passed to an ordinary dense neural network with only one hidden layer of densely connected neurons. This simple densely connected neural network will show more accuracy when compared to an highly powerful long short term memory RNN with word embedding based upon the index values of the words in the vocabulary. The proposed model will show significantly better results in terms of accuracy and efficiency than the LSTM RNN for analysing the Twitter tweets.

I. INTRODUCTION

Twitter Sentiment Analysis : Twitter, with over 319 million monthly active users, has now become a goldmine for organizations and individuals who have a strong political, social and economic interest in maintaining and enhancing their clout and reputation. Sentiment analysis provides these organizations with the ability to surveying various social media sites in real time. Text Sentiment analysis is an automatic process to determine whether a text segment contains objective or opinionated content and it can furthermore determine the text's sentiment polarity. The goal of Twitter sentiment classification is to automatically determine whether a tweet's sentiment polarity is negative or positive.
Most of the existing methods of Twitter sentiment classification apply machine learning algorithms to build a classifier from tweets with manually annotated sentiment polarity label. In recent years, there has been a growing interest in using deep learning techniques which can enhance classification accuracy.

Neural Network: Deep neural networks do the input-to-target mapping via a deep sequence of simple data transformations (layers) and these data transformations are learned by exposure to examples.

Recurrent neural network: The idea behind RNNs is to make use of sequential information. In a traditional neural network we assume that all inputs (and outputs) are independent of each other. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous computations. RNNs is that they have a "memory" which captures information about what has been calculated so far.

LSTMRNN : Long Short Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter and Schmidhuber (1997), and were refined and popularized by many people. They work tremendously on a large variety of problems and are now widely used. LSTMs are explicitly designed to avoid thelong-term dependency problem.

II. RELATED WORK

Zhao Jianqiang,GuiXiaolin1[1] uses "Deep Convolution Neural Networks for Twitter Sentiment Analysis". This paper introduced word embeddings obtained by unsupervised learning on large twitter corpora that uses latent contextual semantic relationships and co-occurrence statistical characteristics between words in tweets. These word embeddings are combined with n-grams features and word sentiment polarity score features to form a sentiment feature set of tweets. The feature set is integrated into a deep convolution neural network for training and predicting sentiment classification labels.

HassanSaifa, YulanHeb, MiriamFernandeza, HarithAlania [2] presented "Contextual semantics for sentiment analysis of Twitter". SentiCircles, a lexicon based approach for sentiment analysis on Twitter is being proposed. SentiCircle aims to learn the sentiment orientation of words from their contextual semantics . This paper allows for the detection of sentiment at both entity-level and tweet-level. This Senticircle approach is being evaluated in three different Twitter datasets using three different sentiment lexicons to derive word prior sentiments.

Nazan Ozturk, Serkan Ayvaz [3] presented "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis". This paper investigated the public opinions and sentiment towards the Syrian refugee crisis which has affected millions of people and has become a widely discussed topic in social media around the world. Based upon the sentiment scores of both the English and Turkish tweets, this paper came up with an analysis result stating that the ratio of positive sentiments in Turkish tweets is 35% whereas the English Tweets were only 12%, thereby the English tweets contained remarkably less positive sentiments towards Syrians and refugees.

Feng Li, Timon C. Du presented [4]"Maximizing micro-blog influence in online promotion". In this paper, a framework is being proposed for identifying the opinion leaders and maximizing the dissemination of messages by analyzing existing micro-blogs. This study has proposed a three-step process of data-envelopment analysis(DEA) to identify opinion leaders capable of maximizing the influence of a message.

JunXu, Rouhollaah Rahmatizadeh, Ladislau Boloni, Senior Member, IEEE, and Damla Turgut, Member, IEEE [5]presented "Real-Time Prediction of Taxi Demand Using Recurrent Neural Networks". This paper proposed a sequence learning model that can predict future taxi requests in each area of a city based on the recent demand and other relevant information. Learning to predict taxi demand in small areas is difficult. So this paper used Geohash library that can divide a geographical area into smaller subareas with arbitrary precision. With the use of this library, it becomes easy to predict the demand as well as sufficiently accurate for the drivers.

III. PROPOSED WORK

The proposed system makes use of an innovative combination of skip gram model of word2vec and term frequency inverse frequency document values to generate feature vector which can then be passed to an ordinary dense neural network with only one hidden layer of densely connected neurons.

Proposed architecture: The proposed system consists of the following modules:
A. Pre-processing
B. Training the Dense NN(skip-gram model of word2Vec)
C. Training the LSTMRNN (vocabulary index based)
D. Comparing the efficiency of Dense RNN with LSTMRNN.

These are the modules that are being used to predict the sentiment of a tweet either as positive or negative and thereby comparing the accuracy and efficiency of one RNN with another RNN.

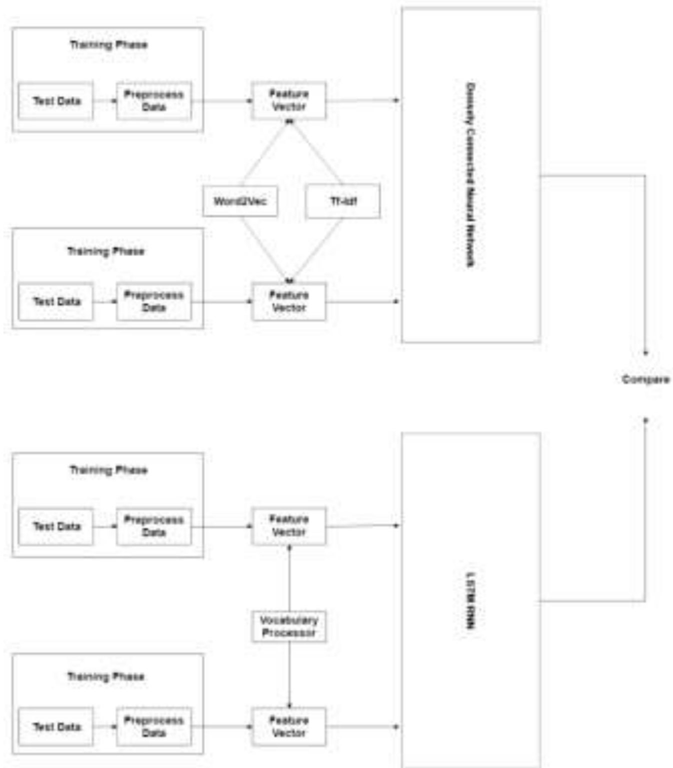

Fig:1 Skip-Gram and TF-IDF Word Embedding Architecture

MODULE DESCRIPTION

A. Preprocessing : Dataset is being loaded and necessary columns required for sentiment analysis are retained and other columns are dropped from the csv file. The necessary columns required for processing the tweet includes, Sentiment and Sentiment Text. It is also necessary to remove the unnecessary information contained in each tweet, that includes user mentions, hashtags, url's etc.These kind of unnecessary information contained in tweets may pose problems during tokenization of tweets, as these url's, hashtags, user's mentions may not effectively help in analyzing the sentiment of tweets.

B.Training The Dense NN(skip-gram model of word2Vec) : The main notion of this module is to train the dense neural network with Twitter dataset. The feature vector being used for training this neural network is the dot product of word2vec value of each token(words of a tweet are processed and tokenized) and the tf-idf value of each token. The proposed system proves that the sentiment analysis done using word2vec and tf-idf word embedding would lead to a better result with just one layer of densely connected hidden layer of neurons. Thereby this system outperforms the more commonly used and an efficient and advanced Lstm RNN trained using one-hot encoded word-embedding.

C.Training The LSTMRNN (vocabulary index based word embedding) : This main notion of this module is to train the Long Short Term Memory of Recurrent Neural Network with Twitter dataset. Each Tweet is tokenized and a vocabulary of words is being created, and then feature vector is being extracted based upon the index location of the words in the vocabulary.

The word embeddings obtained through vocabulary index based is then fed into the Long Short Term Memory Recurrent Neural Network and analysis of sentiment is being done. The optimizer being used to adjust the weights via back propagation is Adam Optimizer.

D.Comparing The Efficiency Of Dense Rnn With Lstmrnn : The main notion of this module is to compare the densely connected neural network with one hidden layer to the recurrent neural network with long short term memory.

IV. PERFORMANCE EVALUATION

Accuracy is a performance metrics that is essential for analysing the performance of the classification test.

$$Accuracy = \frac{\text{no. of TP} + \text{no. of TN}}{\text{No. of TP} + \text{no. of TN} + \text{no. of FP} + \text{no. of FN}}$$

Test Data	True Positive	True Negative	False Positive	False Negative	Accuracy
500000	352500	100201	40796	10003	86.08

V. CONCLUSION AND FUTURE WORK

Social Media is growing to be the most powerful forum for product promotions and advertisements. So analyzing the sentiment of the tweets is very important, thus by making use of a simple densely connected neural network with a more robust word embedding that provides contextual semantics , tweets can be classified better than the more powerful long short term memory recurrent neural network.This approach makes it very easy to classify the tweets accurately and with less effort and time.

The future work would be to include several word embedding scores such as score from sentiment polarity from senti-wordnet and also to make use of n-grams in phase of feature extraction and then to train the more advanced neural network models such as CNN,LSTM RNN,etc. There by the accuracy of sentiment prediction may increase much more.

REFERENCES

[1] Zhao Jianqiang ,Gui Xiaolin ,” Deep Convolution Neural Networks for Twitter Sentiment Analysis”,2017 IEEE Translations and content mining.

[2] Hassan Saifa, Miriam Fernandeza,Harith Alania, “Contextual semantics for sentiment analysis of Twitter”,Information Processing and Management 52(2016)5-19.

[3] Nazan Öztürk, Serkan Ayvaz,” Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis”, Telematics and Informatics 35 (2016) 136-147.

[4] Feng Li, Timon C. Du, ” Maximizing micro-blog influence in online promotion”,Expert Systems With Applications 70 (2017) 52-66.

[5] Jun Xu , Rouhollah Rahmatizadeh, Ladislau Bölöni, " Real-Time Prediction of Taxi Demand Using Recurrent Neural Networks”, 2017 IEEE Transactions on Intelligent Transportation Systems.

[6] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,”in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2015.

[7] A.Krizhevsky,I.Sutskever,andG.E.Hinton,“Imagenetclassifica- tion with deep convolutional neural networks,” in Proc. Advances in Neural Inf. Process. Syst., 2012.