A Deep Learning Approach To Sentiment Analysis of Movie Reviews

Anusri Katuri[1], V.Varalakshmi[2], Supreethi.K.P.[3]

Department of Computer Science and Engineering JNTUH College of Engineering Hyderabad, India[1,3,] Asst.Prof., SMEC, Hyderabad[2]

Abstract- Deep learning is a machine learning technique that lets the computers learn by examples. Sentiment Analysis is the process of categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive or negative. Various methods for sentiment analysis are lexicon- based, statistical and rule-based approaches. The accuracy has been the major drawback in all these methods. Deep Learning models can help us achieve high accuracy.In this study, a deep learning model will be implemented to predict whether a given sentence conveys a positive or a negative meaning. The model is a combination of convolution and recurrent neural networks which are regularized using dropout layers. IMDB dataset have been used and the model implemented using Keras with Tensorflow backend in Python.

Keywords– deep learning,sentiment analysis, imdb, keras

I. INTRODUCTION

The user generated content on web is being increasing which gave the sentiment analysis a vital role to play.Sentiment analysis has been one of the most researched topics in Machine learning [5]. Sentiment analysis is applied over various fields such as business, politics and others. In the field of business, this analysis can help in predicting the outcome of releasing new products into the market.Sentiment analysis can also be applied in politics to predict the most supported politician in a particular region. All these predictions can result in a better future either for a product or for a person.

Sentiment analysis can be carried out in two major steps. Firstly, data collection and pre-processing, which means collecting the data either from social networking sites or a well defined datasets and the performing data cleaning operations in order to prepare the data in the from suitable for the analysis. Secondly, the selection of algorithm to perform sentiment analysis, which means picking the best algorithm which makes the most accurate analysis.An accurate model will be resulted for performing sentiment analysis, that is, categorizing the data into either positive or negative degree. This process can be done as supervised or unsupervised learning. Unsupervised learning is where the classification is done without any prior knowledge.

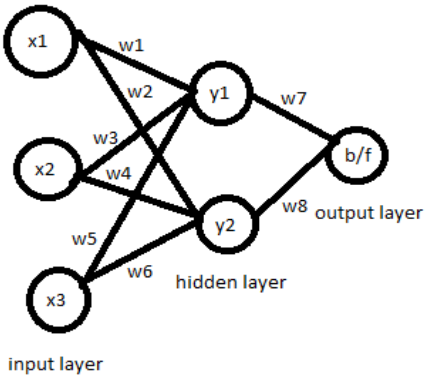
There are basically two main approaches to sentiment analysis which are lexicon-based approach and machine learning approach[2]. Categorizing a sentence into either a positive or a negative sentence so called sentiment analysis is a natural process for a human brain. But this manual procedure is not sufficient to deal with such huge amounts of data over web. In order to get rid of this issue deep learning approaches are developed.

The rest of this paper is organized as follows: Section II presents related work. Section III described the proposed system and Section IVdescribes the experimental results. Section V concludes the paper.

II. RELATED WORK

Deep Learning is an efficient learning method which makes use of neural networks to perform required tasks.

Neural Networks are analogous to the functioning of biological neurons in human brain. These artificial networks consists of three layers namely, input layer, output layer and an optional hidden layer.



Neural networks are fully connected graphs with each node associated with an input value and each edge associated with a weight, which are intially random values and a bias added which is always set. Functioning of a neural network is carried out by calculating the weighted sum.

weighted sum = $\sum w_i x_i + b$

This weighted sum may be applied a special function to optimise the output. Such special functions are called **Activation Functions** which are used to make the output non-linear so that classification is made possible.

Relu is rectified linear unit which is used to obtain only positive values and zeros.

$$\mathbf{relu=max(0,x)}$$

Sigmoid function is a special S shaped curve which limits the values between 0 and 1.

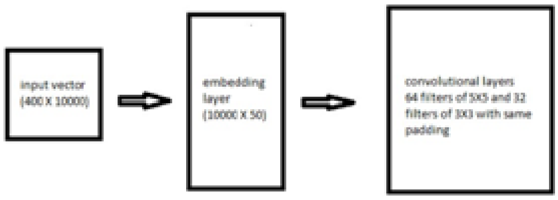
$$\mathbf{sigmoid = 1/(1+e^{-x})}$$

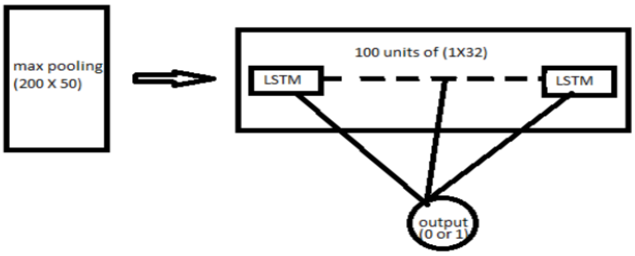
Training dataset is that set of examples which are fed to neural network to make it learn. As the solutions to the input are already known, the neural network learns from these examples so that it could give the expected outputs. The number of correctly classified examples to the total number of examples used as training data gives out the **training accuracy**.

Testing dataset is that set of examples which are fed to neural network to test how well the neural network learned to classify from the training dataset. As the solutions to these inputs are already known, the neural network is tested on these examples to check if it's giving out the desired outputs when it is tested on new data which is different from the training dataset. The number of correctly classified examples to the total number of examples used as testing data gives out the **testing accuracy**.In **supervised learning** data feed to the algorithm includes desired solutions called **labels** [1].In **unsupervised learning**, training data is unlabeled. The system tries to learn without any supervision [1].

III. PROPOSED SYSTEM

The flow of the system can be seen as in several blocks. Firstly, the inputs are to be fed to the neural networks which are in english language. As neural networks doesn't understand English language, the sample statements undergo word-to-vector representation where each word is represented by its rank given according to the number of the frequently repeated word among the other words in the dataset. The vector finally obtained is term as input vector. This input vector is trained to two consecutive types of neural networks namely, convolution neural network and long short term memory network.Convolution neural network is a layered neural network. The first layer is embedding layer which embeds the words into low-dimensional vectors. The second layer is convolutional layer. The easiest way to understand a convolution is by thinking of it as a simple multiplication of the input vector and the weight vector to finally calculate the weighted sum. The third layer is max pooling layer. Here a filter is used to reduce the dimension of input to this layer by replacing values with the maximum among considered values.





The next type of neural network is long short term memory(LSTM). This consists of three gates namely, input gate to read in the input, output gate to write out the output to next layers and forget gate which decided what to remember and what to forget. All these gates are sigmoid associated gates. LSTM gives several features to a fine-grained control over memory; this aspects control how much the present input matters for forming the new memory, also how much the prior memories matters in designing the new memory, and what parts of the memory are essential is producing the output[3].Finally a single node dense layer is used to obtain the output of the classification. Meanwhile, the neural network adjusts its network to predict the outcome with at most accuracy. As a part of it, it uses special algorithm called back propagation to calculate the loss and propagate the error back in the network and update the weights accordingly.

IV. EXPERIMENTAL RESULTS

The IMDB dataset used for training the neural network which consists of 50,000 labelled samples which are either positive or negative. Out of which 25,000 samples are used for training and rest for testing the neural network.The model built so can successfully obtain a training accuracy of 96.5 % and a testing accuracy of 88.7%.These results are influenced by various parameters.Like the epoch, which is a measure of number of times all of the training inputs are used once to update the weights and the batch size, the number of training inputs in one forward or backward pass. It is observed that the performance degraded upon increasing the epoch and batch sizes. This suggests that optimization of these two parameters is crucial to good performance of both CNNs and RNNs[4].

V. CONCLUSION

Classification of text has become a challenging task because of the sarcasms and informal conversations used. Deep neural networks are implemented to try to capture such type of conversations which are in the form of text. Text has been classified into either positive or negative sentence this technique. Consideration of the third label as neutral may improve the classification process which can be seen as a future scope of this project

VI. REFERENCES

[1]. Aurelien Geron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, Second Indian Reprint, Shroff Publishers and distributors.

[2]. Addlight Mukwazvure, K.P Supreethi, A Hybrid Approach to Sentiment Analysis of News Comments, 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 2-4 Sept. 2015, pp 472-478.

[3]. Abdalraouf Hassan and Ausif Mahmood, Deep Learning for Sentence Classification, IEEE conferences, 2017.

[4]. Wenpeng Yin , Katharina Kann , Mo Yu and Hinrich Schutze , Comparative Study of CNN and RNN for Natural Language Processing, February 2017.

[5]. Anmol Chachra, Pulkit Mehndiratta, and Mohit Gupta, Sentiment Analysis of text using deep convolutional neural networks Tenth International Conference on Contemporary Computing (IC3), 10-12 August 2017.

[6]. .Qiongxia Huang; Riqing Chen; Xianghan Zheng; Zhenxing Dong, Deep Sentiment Representation Based on CNN and LSTM International Conference on Green Informatics (ICGI) 2017.