International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-06, Issue-01, Apr 2020

Usense: A User Sentiment Analysis Model for Movie Reviews by Applying LSTM

Undru Taran Rishith, Student, Gokaraju Rangaraju Institute of Engg and Technology, Hyderabad, India, taranrishit1234@gmail.com

Abstract - Sentiment analysis is a field which deals with assessing the sentiments or emotions of the users on products and services. It takes user comments as input and applies natural language processing techniques to identify the mood of the user. Usually a sentiment is deemed to be positive, negative or neutral depending upon the mood that he expresses in the comments or feedbacks. It is largely used by businesses to improve products and services and also to present its customers with a set of products and services based on their likes and dislikes. State-of-the-art indicates many techniques have been applied in past such as, linear regression and SVM models. Recurrent Neural Networks (RNNs) have improved the way in which sentiment analysis could be done with greater accuracy, but they suffer from major drawback when applied to longer sentences. This paper proposes a sentiment analysis model using Long Short-Term Memory (LSTM) based approach , which is a variant of RNNs. LSTMs are good in handling long sentence data. The model is applied to reviews collected from IMDB dataset. It is large dataset that contains 50K reviews. Out of the available reviews 50 % are used for training purpose and 50% are used for testing purpose. The model gives a training accuracy of 92% and validation accuracy of 85% which is neither an over fit nor an under fit. The overall accuracy here is 85%, which seems to be better than some of the existing techniques such as SVM with linear kernel.

Keywords — IMDB, Lemmatize, LSTM, Movies, NLP, Sentiment Analysis, Tokenize

I. INTRODUCTION

In this competitive age businesses not just have to focus on selling the products and services to their customers but they have to focus on the likes and dislikes of the customers. Gathering information about customers sentiments about products and services have become easier with the addition of customer feedback options made available to customers by the businesses. Customer feedbacks not only help in understanding the customers' expectations but also help in analyzing the value of the products and services offered to them.

In recent years a field of study called Sentiment Analysis has come up. It involves interpretation and classification of user emotions into positive, negative or neutral within textual data by applying text analysis tools and techniques.

Sentiment analysis is used by businesses to know user sentiments towards their products and services through online feedback.

Sentiment analysis models detect user opinions (e.g. *positive* or *negative* opinion), from a whole document, paragraph, sentence, or clause. Nowadays it has become easier than before for the customers to express their thoughts and emotions. Businesses are able to gather customer feedback from surveys and social media communications and tailor make their products and services as per customers' needs.

Applications of sentiment analysis are endless; it could be applied to social media channels to identify possible product supporter or influencers. It can also be applied to corporate networks to identify the "tone" used in e-mails. It could also be used to track the negative threads about your products and services, thereby allowing businesses to be proactive in dealing with such feedbacks.

Sentiment analysis is a text classification problem as it is applied to text to extract the emotions. We discuss two main ways of implementing sentiment analysis in the field of computers.

(i) Supervised Learning and (ii) Unsupervised Learning

Supervised learning involves a classifier which classifies the input text as positive, negative or neutral. Three main classification techniques are (i) Naïve Bayes (ii) Maximum entropy and (iii) Support vector machines (SVM).

Unsupervised Learning has three steps.

1. Implement Part of Speech (POS) tagging , then, two consecutive words are extracted to identify if their tags conform to given patterns
2. Estimate the sentiment orientation (SO) of the extracted phrase
3. Compute the average SO of all phrases that were extracted in terms of positive or negative

 DOI : 10.35291/2454-9150.2020.0315

In this paper we are applying unsupervised sentiment analysis model called long short-term memory (LSTM) to analyze the user reviews on movies by classifying them into three polarities: positive, negative or neutral.

Section 2 deals with the literature review on the problem; section 3 describes the proposed method and section 4 winds up the paper by analyzing the results.

II. RELATED WORK

One area where sentiment analysis is being applied nowadays is the movies. Movie industry is one of the largest businesses in the world making a whooping amount of millions of dollars every week. Movie reviews helps in measuring the performance of a movie. A numerical/stars rating given by a viewer for a movie helps us quantitatively to know about the success or failure of a movie, whereas collection of movie reviews is provides us with a deeper qualitative insight on different aspects of the movie. A review gives an idea about the strength and weakness of a movie and analysis of movie reviews can help the movie makers understand in general if the movie met the expectations of the viewers.

In this section we review the state-of-the-art on the problem domain and try to understand the research gaps and areas that ought to be researched and come out with more improved solution to the problem.

In the paper [**1]** the authors describe the implementation of two machine learning algorithms for analyzing movie reviews - Naive Bayes and Support Vector Machines (SVM) can be used to detect sentiments from the text. Two datasets of size 2,000 and 5,000 were applied for training the algorithms each having subfolders labeled as positive and negative. On 2000 dataset size Naïve Bayes showed 78.59% accuracy and SVM depicted 84.37 percent accuracy. Similarly, on the dataset of size 5000Naive Bayes showed 79.19% accuracy and SVM gave 80.85% accuracy.

In the paper [**2**] the authors analyzed the user sentiments on movie reviews by applying three different analyzers namely Textblob, SentiWordNet and Word Sense Disambiguation (WSD). Three thousand tweets about a hindi movie named "Race 3" were gathered from public accounts. Among the three sentiment analyzers that were compared in the research, they found that TextBlob resulted in the highest rate of tweets with neutral sentiment, 1729 in number and 57.6 % in percentage. SentiWordNet gave 377 & 197 the highest negative & highly negative sentiment rate, 12.5% & 6.6 % respectively.

Authors in the paper [**3]** applied Latent-Semantic Analysis (LSA) algorithm to analyze the movie reviews and recommends users based on the reviews. Sentiment analysis is done on the comments given by the users. The analysis of comments given by various users was done and a common review was generated. The generated review was a simple English statement and will help user to take a correct decision while selecting any movie and also allows the user to recommend the movies to their friends.

In the paper [**4]** the authors proposes a system that is able to classify sentiments from review documents into positive sentiment and negative sentiment. Naive Bayes Classifier was used to classify the documents. Movienthusiast, a movie reviews in Bahasa Indonesia website was used as the source of our review documents. A total of 1201 movie reviews were collected: 783 positive reviews and 418 negative reviews that were used as the dataset for the learning classifier. The average accuracy achieved by the proposed system was 88.74% out of five time testing processes.

In the paper **[5]** the authors have explored different natural language processing (NLP) methods to perform sentiment analysis. Two different datasets were used, one with binary labels, and one with multi-class labels. For the binary classification they applied the bag of words, and skip-gram word2vec models followed by various classifiers, including random forest, SVM, and logistic regression. For the multi-class case, they implemented the recursive neural tensor networks (RNTN). In order to overcome the high computational cost of training the standard RNTN they introduce the low rank RNTN, in which the matrices involved in the quadratic term of RNTN are substituted by symmetric low-rank matrices. We show that the low-rank RNTN leads to significant saving in computational cost, while having similar a accuracy as that of RNTN.

Dan Li and Jiang Qian [6] in their work on text sentiment analysis using LSTM have compared the LSTM model with conventional RNNs. They found that LSTM is better in analyzing emotions in long sentences it also produced better accuracy levels as compared to RNNs.

A paper by Baid et.al, [7] describes a comparative study of three algorithms namely, Naïve Bayes, K-Nearest Neighbour , and Random Forest for sentiment analysis on movie reviews. Experimental results indicate that Naïve Bayes outperformed the other two algorithms. Accuracy reported were 81.45%, 55.30%, and 78.65% respectively.

In the paper [8] the authors performed text categorization of movie reviews by applying Naïve Bayes and SVM classifiers on 2000 movie reviews taken from Cornell University dataset repository. The results indicated that Niave Bayes classifier slightly did well than the SVM.

III. PROPOSED MODEL AND IMPLEMENTATION

The objective of the proposed model is to analyze the sentiments hidden in the reviews given by the users in the form of comments or feedback. The proposed sequential model for sentiment analysis based on LSTM is depicted in figure 1 below.

 DOI : 10.35291/2454-9150.2020.0315

Figure 1: Proposed Model

The movie reviews are collected from IMDB database which is retrieved from imdb.com. First 50000 sample reviews are taken from IMDB along with a vocabulary of 15000 words. The data is split into train/split data in the ratio 50:50.

We convert the target variable data which is positive /negative into categorical data if it isn't already. Here positive and negative data is converted into 1 and 0.

If we have sentence we split it into individual words using tokenizer .

Lemmatizer is used to give the root word instead of verb forms of word

Then every sentence (array of words) is converted into integers representing an index.

For example, if we have a sentence of the form shown in figure 2.

```
['fawn', 'tsukino', 'nunnery', 'sonja', 'vani', 'woods',
'spiders', 'hanging', 'woody', 'trawling', "hold's", 'comicall
y', 'localized', 'disobeying', "'royale", "harpo's", 'canet',
'aileen', 'acurately', "diplomat's"]
```

Figure 2: Sample sentence

Then this sentence is converted into array of indices as shown in figure 3.

```
[('fawn', 34701), ('tsukino', 52006), ('nunnery', 52007),
('sonja', 16816), ('vani', 63951), ('woods', 1408), ('spiders
', 16115), ('hanging', 2345), ('woody', 2289), ('trawling',
52008), ("hold's", 52009), ('comically', 11307), ('localize
d', 40830), ('disobeying', 30568), ("'royale", 52010),
("harpo's", 40831), ('canet', 52011), ('aileen', 19313),
('acura
tely', 52012), ("diplomat's", 52013)]
```

Figure 3: Sample sentence converted to array of indices

Because each sentence is not as equal length as others, padding is done to sequences with zeros to a limit "N" which might be max length of sentence in the dataset or it can be a value greater than the largest sentence in the dataset. At this stage a model is build.

First layer of the model creates an embedding which is passed as input to the LSTM. The embedding will be of the shape (250*128) since we pass sentences of length 250 and d each unit (or word) is in between (0, 15001). See figure 4.

Figure 4: Embedding

Then it is passed to LSTM containing 100 units . we train 100 units at a time. Next we use use "dense" function which will take the output from LSTM and condense it to two classes. The two classes being positive and negative which is categorically mapped to 1 and 0.

Sigmoid activation is used to get values between 1 and 0. Those values above 0.5 are considered 1 and below 0.5 are 0. We use the 50% data of training on this model and validate it with the 50% test set as shown in model.fit() in figure 4 above. We now save the model as shown in figure 5 below.

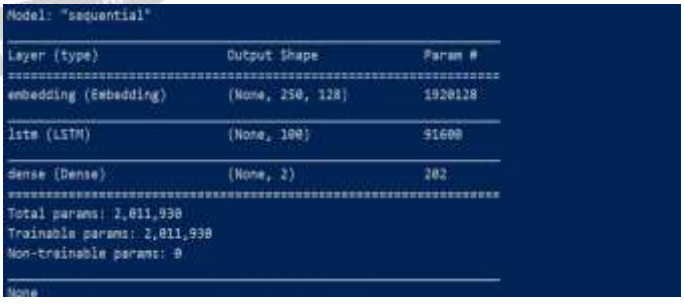

Figure 5: Saved Model

IV. RESULTS AND CONCLUSION

The proposed model is applied to reviews collected from IMDB. It is large dataset that contains 50K reviews. Out of the available reviews 50 % are used for training purpose and 50% are used for testing purpose. After applying all the required steps the loss and accuracy plots of the proposed approach is obtained, as shown in figures 6 and 7 respectively.

 DOI : 10.35291/2454-9150.2020.0315

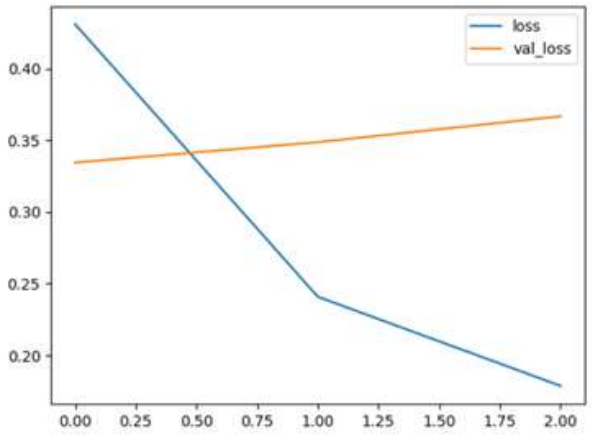

Figure 6: The loss plot

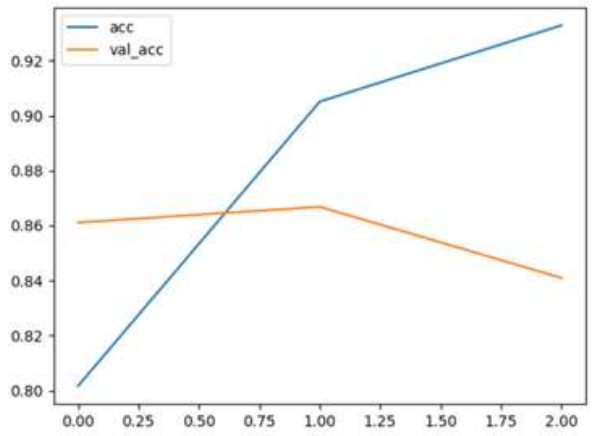

Figure 7: The accuracy plot

We can see that built model training accuracy obtained is more than 92% and validation accuracy is 85% which is neither an over fit nor an under fit. The overall accuracy here is 85%, which seems to be better than some of the existing techniques such as SVM with linear kernel.

Although the proposed model performs well for classifying sentiments from the reviews, it could be further improved if it can be applied with other embedding techniques on various data sets.

Although this model performs fairly well with English language , to further the scope of this work the model can be improvised by adding multilingual component to it and a classifier that segregates the reviews based on the language and check for the user sentiments.

REFERENCES

[1] P. Sanghvi, D. Shah , H. N.Bharathi , *"Movie Review System Using Sentiment Analysis,"* *IJRASET*, vol.7,no.5, pp.1193-1196, May 2019.

[2] M.Gupta and P.Sharma, *"Sentimental Analysis of Movies Tweets with Different Analyzer,"* *IJCST*, vol.3, no.3, pp.200-204. May-June 2018.

[3] R. M. Sharma, S. S. Barkul, P.K. Sawane, R.M. Jeughale, *"Online Movie Review System,"* *MJRET*, pp.547-552, 2015.

[4] Y.Nurdiansyah, S.Bukhori, R. Hidayat, *"Sentiment Analysis System for Movie Review in Bahasa Indonesia Using Naive Bayes Classifier Method,"* *ICCGANT IOP Conf. Series: Journal of Physics: Conf. Series 1008*, 2018.

[5] H.Pouransari, and S. Ghili *"Deep learning for sentiment analysis of movie reviews,"* https://cs224d.stanford.edu/reports/PouransariHadi.pdf

[6] D. Li and J. Qian, *"Text Sentiment Analysis Based on Long Short-Term Memory,"* *First IEEE International Conference on Computer Communication and the Internet*, 2016.

[7] B. Palak, G. Apoorva and C. Neelam, *"Sentiment Analysis of Movie Reviews using Machine Learning Techniques,"* *International Journal of Computer Applications*, vol.179, no.7, pp.45-49, Dec 2017.

[8] S.Humera. G.Kavitha , R,Zaheer , "*Text Categorization of Movie Reviews for Sentiment Analysis,"* *IJIRSET*, vol.4, no.11, pp. 11255-11262, Nov 2015.

 DOI : 10.35291/2454-9150.2020.0315