ISSN: 2581-3404 (Online)
International Journal of Innovative Research in Technology and Management, Vol-4, Issue-2, 2020.

Sentiment Analysis of Movie Reviews

[1]Shubha Mishra, [2]Dr. Piyush Shukla, [3]Dr. Ratish Agarwal
[1]Department of Information Technology, UIT RGPV Bhopal, (M.P.), India.
[2]Department of Computer Science Engineering, UIT RGPV Bhopal, (M.P.), India.
[3]Department of Information Technology, UIT RGPV Bhopal, (M.P.), India.
[1]mis.shubha@gmail.com, [2]piyush@rgtu.net, [3]ratish@rgtu.net

ABSTRACT
In today's era of internet, everyone is connected on the go. We all have access to huge chunks of information available which can be structured, semi-structured or unstructured data. The unstructured data is available in huge amounts but this unstructured data can be used through statistical analysis to derive useful information from it. The multiplexes based on the user ratings determine whether a movie should be continued in cinema or not. So, the movie sentimental analysis holds huge importance for these multiplexes because if the movies have good user reviews then more viewers are expected to come to theatres which will generate higher revenues to the multiplexes. This paper introduces an approach to sentiment analysis which uses the support of Keras and Sequential model.

Keywords:- Sentimental Analysis ,Opinion Mining ,Recurrent Neural Networks, Sequential Model.

INRODUCTION
Sentimental Analysis is a new subject in Research and is useful in many other fields .In Modern World, A huge amount of textual data is collected using surveys, comments and reviews all over the web[5].It was difficult to predict the rating on the basis of text ,further operations had to applied on the text in order to make those reviews helpful for predicting the review sentiments. It is a sub-domain of opinion mining where the analysis is focused on the extraction of information, emotion and opinions which are converted into features that can be used to convert into vectors by different vectorizers which are used to generate the output features this is known as feature extraction [1]. Now the sentiments are ranged from 0 to 5 in the inclusive range, from not preferred to highly prefer.
The sentiment labels are:
0 - negative
1 - somewhat negative.
2 - neutral
3 - somewhat positive.
4 - positive

I A. Word Embeddings
Prior to training this model, convert every word into a word embedding. We can understand, by thinking of word embeddings as numerical vector representation of words to allow our model to learn .Machine Learning models can only be trained on the numerical values therefore the text values needs to be converted into an integral value these integral values are used to train the machine learning model.

Word Embeddings are vector representations of the words that store the meaning of the underlying words in relation to different words present in the sentence. This modification results in words which have similar meaning being clustered closer in the hyper-plane and different words being positioned at greater distances in the hyperplane.

This way the model we input the whole paragraphs or reviews into the LSTM. In this model we just loop up for each word, the integer value in the word to index map, create the appropriate one-hot-encoded vector and performed a dot product with

the matrix. The review is then fed word by word(vector by vector) into the LSTM network.
Each vector that represents a word in the dataset is obtained from a large matrix, called embedding-matrix. The number of rows of this matrix represents the dimensionality of the word embedding, the number of columns represents the vocabulary size or number of unique words in the dataset. Thus each column of this matrix represents an embedding vector for a unique word in the dataset.

How do we know which column represents which word? This is where we use the word-to-index map. Consider you want to get the embedding vector for the word "although", according to the word-to-index map this word is represented by the number 2511. In the next step, it is necessary to create a one-hot-encoded vector of size 18339 (number of words in the dataset), where each entry is 0 except for the 2511th entry which has the value of 1. By doing a dot-product between the embedding matrix and the one-hot-encoded vector we obtain the 2511th column of the matrix, which is the embedding vector for the word "although"

I B. Sentimental Analysis
The most important aspect of business is to know what the customer desires and what are their likes and dislikes. It is vital for them to know what exactly does the customer think of the new products or services,or recent tech or non-tech initiatives and customer service offerings like the movies that are released. Sentimental analysis is one way to achieve this vital task. Sentimental Analysis is a sub-domain of Natural Language Processing(NLP) that builds models which try to understand the emotion or the opinion expressed in a text based form for eg:
Polarity : whether the person says something negative or positive with respect to some topic.
Subject: The topic, which is talked about or analyzed .
Opinion holder: the reviewer or the speaker who provides his review or opinion.
In today's world , we generate about 2.5 quintillion bytes of data everyday .It has become a necessity to understand what is being expressed in this huge amount of unstructured data available on the web.Sentiment analysis plays a key role as it analyses the given texts and determines the sentiment of the data , this sentiment can be used by different organizations to improve upon their service or products. Even sentiment analysis can play a vital role in the election campaigns to determine the public opinion and can also be used to change the outcome of the elections by making certain amendments based on the likes of the people like the technique used by Donald Trump.

II BACKGROUND & MOTIVATION

Sentimental Analysis is extremely important for the enterprises in order to know the views of their customers. The sentimental analysis of the movie reviews play a vital role in understanding whether the audience is liking the movie in the theatre or not . It determines whether the movie should be continued in the theatres or not , which is of extreme significance to generate huge profits. Sentimental analysis involves understanding the opinion of the user which can also be termed as opinion mining. The Sentimental Analysis involves natural language understanding as well as natural language processing which are sub-domains of Deep Learning.

III LITERATURE REVIEW

Gurshobit Singh Brar in [1] Sentimental Analysis of movie reviews using Supervised machine learning techniques. In this model is able to do binary classification that it is generates whether the movie reviews is positive or negative. In the POS tagging is used which generally tags the nouns, pronouns and adjectives in the sentences and the feature extraction is done and the sentiment is generated on the basis of the positive or negative polarity of a sentence. Joscha et.al in [2] in their paper compared different approaches like bag of words or n-grams techniques in which the model is fed with unigram ,bigram, trigram i.e. n-gram which avoid a semantic relationship between the different words of review. Monu Kumar et.al in [4] made the use of hadoop and cloud service for handling the huge amounts of

data which is generally termed as the big data .Here the twitter's data sentimental analysis is done by making the use of the cloud platform. Baid, Apoorva et.al in [3] Sentimental Analysis of movie reviews using Movie Reviews using Machine Learning Techniques . In this paper the authors used 3 different algorithms which are Naïve Bayes Classfier, Random Forest ,K-Nearest Neighbours which had accuracies 81.45%,78.65% and 55.30% respectively. The polarity method is used here as well in which the review is classified only has either positive or negative. Kamal Ahmad in [11] has designed a framework of opinion analysis that facilitates objective analysis, feature extraction and review summarization etc. Supervised machine learning algorithms like Naïve Bayes , Decision Trees and Bagging to improve the performance, in his paper, he improved mining techniques to achieve greater performance. A. Mudinas and Zhang in [13] gave a hybrid technique which generates a higher performance as compared to the lexicion and nearly performs like learning based technique. Hybrid Techniques are almost efficient as lexicon technique .The proposed model has an accuracy of 82.3%. Aurangzeb Khan in [14] The model proposed has a document level accuracy of 91% and sentence based accuracy of 86%. Humeriya Shaziya et.al in [12] performed sentiment analysis for the movie reviews by making the use of WEKA tool. In their paper the output is in the form negative or positive sentiment. In the paper , they even considered the reviews which might have a mixed sentiment positive or negative. They used WEKA and made a conclusion that Naïve Bayes performs more efficiently than SVM with an accuracy of 85.1%.

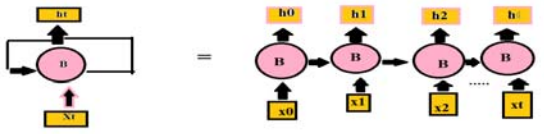
IV PROPOSED WORK

IVA Problem Definition

To computationally find out the sentiment of a movie review whether it is positive or negative based on the past data available as the resources of the internet which usually contains unstructured data. The old models generated ,a lower accuracy by the use of different classification algorithms.

IV B. Model Architecture

For us to train the model we will need a type of Recurrent Neural Network known as LSTM (Long Short Term Memory). The most important reason for selecting the LSTM type of neural network is that it is able to remember the sequence of the past data in our case that can be the words ,the useful words for us to make the decision on the sentiment of the word.



As seen in the above picture it is basically a sequence of copies of the cells, where output of each cell is forwarded as input to the next. LSTM network are essentially the same but each cell architecture is a bit more complex. This complexity as seen below allows the each cells to decide which of the past information to remember and the ones to forget.

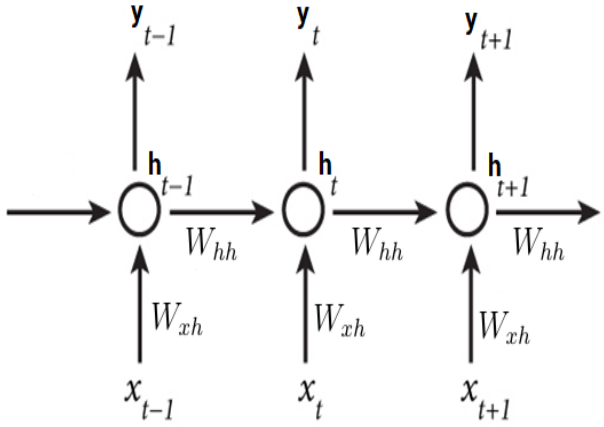
I created the network using Keras. Keras is built on tensorflow and can be used to build most types of deep learning models. In order to estimate the parameters such as dropout, no of cells etc I have performed a grid search with different parameter values and chose the parameters with best performance.

A. Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are popular models that have shown great promise in many NLP tasks. RNN's make use of sequential information such as text. In a "traditional" feedforward neural network we assume that all inputs are independent of each other. But for many tasks that's a very bad idea. A sentence, for example, has a clear grammatical structure and order, where each word depends on the previous word. If you want your neural network to learn the meaning (or sentiment in our case) the network must know which words came in which order.RNNs are called recurrent because they

ISSN: 2581-3404 (Online)
International Journal of Innovative Research in Technology and Management, Vol-4, Issue-2, 2020.

perform the same task for every element of a sequence, with the output being dependent on the previous computations. Another way to think about RNNs is that they have a "memory" which captures information about what has been calculated so far.

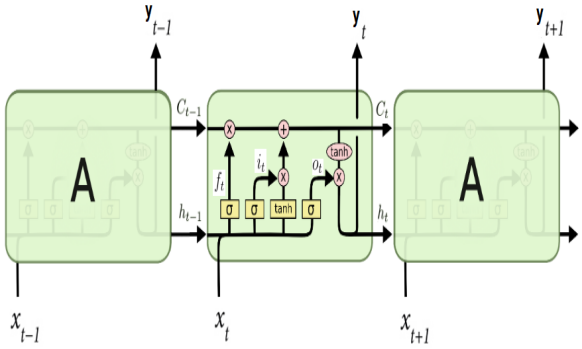


x(t-1), x(t), x(t+1) are sequential inputs that depend on each other (such as words in a sentence). y(t_1), y(t), y(t+1) are the outputs. Unique for RNN is is the fact that the calculation of the current hidden state h(t) of the neurons for the input x(t) depends on the previous hidden state h(t-1) for the previous input x(t-1). Wxh and Whh are weight matrices that connect the input x(t) with the hidden layer h(t), and h(t) with h(t-1) respectively. This way we introduce a recurrence to the neural network which can be considered as a memory on the previous inputs. In theory, this way "vanilla" RNNs can make use of information in arbitrarily long sequences, but in practice, they are limited to looking back only a few steps.
This is where LSTMs come in handy.

B. LSTMs

Long Short-Term Memory networks -- usually just called "LSTMs are a special kind of RNN, capable of learning long-term dependencies. LSTMs don't have a fundamentally different architecture from RNNs, but they incorporate additional components.

The key to LSTMs is the cell state C(t), the horizontal line running through the top of the diagram. A cell state is an additional way to store memory, besides just only using the hidden state h(t). However, C(t) makes it possible that LSTMs can work with much longer sequences in opposite to vanilla RNNs. Furthermore, LSTMs have the ability to remove or add information to the cell state, carefully regulated by structures called gates. Gates are a way to optionally let information through. An LSTM has three of these gates, to protect and control the cell state



- Forget Gate: After getting the hidden state h(t-1) of the previous input x(t-1), Forget gate helps us to make decisions about what must be removed from h(t-1) state and thus keeping only relevant stuff.
- Input Gate: In the input gate, we decide to add new stuff from the present input x(t) to our present cell state C(t).
- Output Gate: The output gate as the name suggests, decides what to output from the current cell state C(t) to the next C(t+1). For the language model example, since it just saw a subject, it might want to output information relevant to a verb, in case that's what is coming next. For example, it might output whether the subject is singular or plural so that we know what form a verb should be conjugated into if that's what follows next.

Behind each of these states are separate neural networks.

As you can imagine this makes LSTMs quite complex. At this point, I won't go much more into the detail about LSTMs

IV C. Preprocessing

To use the reviews from the dataset to train the model first it is required to preprocess the reviews data so as to make it fit for the model utilization. It is required to do some preprocessing on the data. Main objective here is to reduce the dataset to only to the required set free of null and unnecessary content which can be done by the process of normalization. Consider some words such as “Different” and “different”. For us these have words have similar meanings, the only difference between them is that the first word is capitalized in one word and the other has all the characters as the lower case. But for the neural network, these words will have(before training) a different meaning because of their different spelling. Only during training or after the training , the neural network i.e. the LSTM may learn these words have same meaning or may not learn. We want to avoid such misconceptions. Because of this the first step is to lowercase all the letters in all the reviews then we remove the stopwords from the sentences ,then we perform the tokenizaion to break the reviews into single tokens and then lemmatization is performed to scrap out the lemma and remove the unnecessary tenses and only the important lemmas of the words are left over in the data after performing this preprocessing which is termed as normalization.

IV D. Model Parameters

Activate Function: Activation function is one, which helps us find complex relationships in the data to be captured by the model

Optimiser: In this model Adam optimizer is used, which Is an adaptive learning rate optimizer.

Loss function: The proposed model trains a network to output a probability over the 2 classes using Binary_cross entropy loss. It is very useful for binary classification.

IV E. Datasets

The dataset used in this paper contains tab-separated files with phrases which are present on the Rotten tomatoes. The dataset is divided into two files one is train.csv and the second being test.csv. The test set is used for the validation purposes. There are multiple parameters present in the train.csv and test.csv which includes the reviews text as well as their sentiment label which ranges from 0 to 5.

IV F. Initial Approach

The unstructured data is available in huge amounts but this unstructured data can be used through statistical analysis to derive useful information from it. Initial model tried to understand the user opinion and emotions through statistical approaches termed as machine learning principles used through Natural Language Processing. All the machine learning models Natural Language Processing is used to understand the text under observation, it tries to convert the textual data into a n-dimensional vector which are represented in the hyper plane. Models operate on the numerical data internally , therefore we need to understand the convert the string type or the textual type values into the numerical values which can be understood by machine learning models.

The initial accuracy generated by the Xgboost algorithm is quite low , which is 70.21%, it cannot be used to create a real time model . Initially I used the N-gram approach to generate the unigrams to create the features which were used to train the model but the accuracy was severely low so I had to change the approach.

IV G. Approach

The sequential model of keras is used in the proposed paper. Model aims at achieving high validation accuracy, because accuracy is with respect to training data i.e. how well the model predicts the training data values. The most important thing that is focused while creating this model is to reduce the problem of overfitting. Overfitting can be termed as the model being training data specific i.e. model can predict the output for the training data accurately but it fails to provide accurate output for the data it hasn’t seen before. Here the features are created using the recurrent neural network type LSTM where the normalized data after performing word embedding is converted into vector and the vector

representation of each word is input which in turn generates a y(n) for each input vector n and for all the words in the review such outputs are generated and these are used to generate features and the mean of these features is used to generate the final sentiment suggested by the user review.

The features generated from the RNN i.e. LSTM network is further used provide to the sequential model which is chosen for training .The specific parameters are set determined using the grid search cv technique which is generally used in machine learning to determine the model hyperparameters as there can be many combination of hyperparameters ,therefore to choose the best set of parameters for the dataset we can have for the model.The hyperparameters of the proposed model in this paper are determined using grid search cv and further the model is trained.

This approach is of utmost importance as the reviews collected in real time like this and the useful analysis of such data will lead to a better understanding of the online generated content and will also create a significant impact on the effective decision making in the current context. To do so, we require rich dataset from across not only a single site, but from different social sites in bulk and this data is made useful for performing this vital analysis possible in an efficient manner. In the following figure, a set of steps in the prescribed manner are explained to demonstrate the process of text classification and on the basis of it categorize the data to decide the context and semantic of the dataset analyzed. The features that are generated from word vectors is feed to the Recurrent Neural Networks and the mean features from LSTM fit into the sequential model and trained to give an estimate of the validation accuracy.

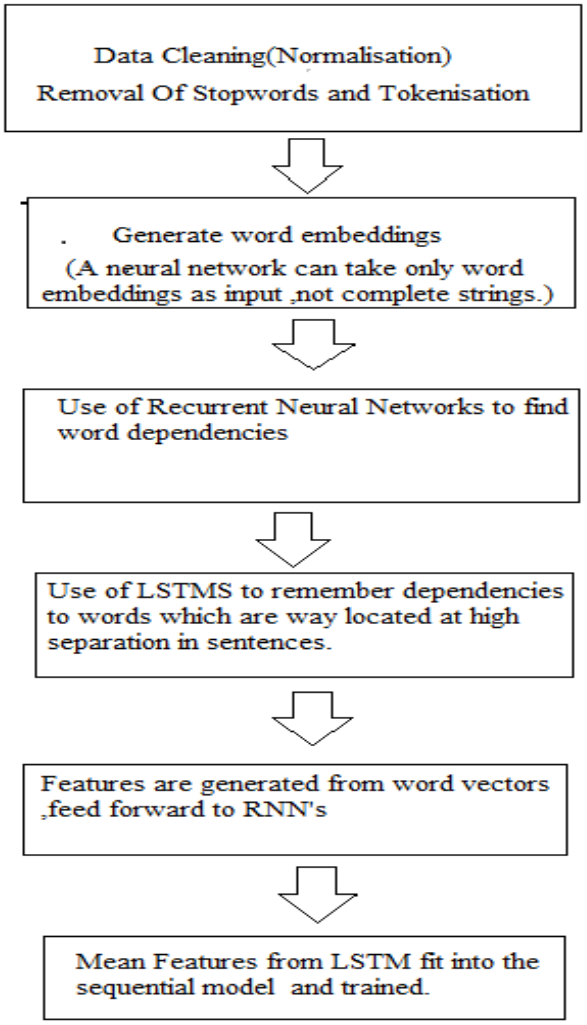


Figure 1: Process Flow Diagram.

V RESULT ANALYSIS

The final validation accuracy achieved is 92.33% whereas the accuracy on training data is 94.99%, therefore the real world implementation of the proposed model will perform with an accuracy of 92.33%.The training dataset provides with the review rating from 0 to 5 because of which the output is extremely refined and accurate. Unlike the previous proposed models like by Aurangzeb in his paper has an accuracy of 86%,Baid an

accuracy of 81.75%, Mudinas and Zhang an accuracy of 82.3% , Humera Shaziya an accuracy of 85.1%(using the WEKA tool), but the proposed model has an accuracy of 92.33%.

VI CONCLUSION & FUTURE WORK

The proposed model has an accuracy of 92.33%. An API can be created using Django or Flask, which can be used by the multiplex owners to feed in the user reviews and check the sentiment of the customers. The model can be extended by using a dataset which also includes emojis in it so that we can even generate the sentiment on the basis of reviews which only contain the emojis or the reviews which contains both emojis as well as the textual review of the user.

REFERENCES:

[1] Gurshobhit Singh Brar, "Sentimental Analysis of Movie reviews using Supervised Machine learning techniques", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 16(2018) pp. 12788-12791.

[2] Joscha Markle-HuB, Stefan Feuerriegel Helmut Prendinger, "Improving Sentiment Analysis with Document-Level Semantic Relationships from Rhetioric Discourse Structures",2017 Proceedings of the 50th Hawaii International Conference on System Sciences.

[3] Palak Baid, Apoorva Gupta, Neelam Chaplot, "Sentimental Analysisof Movie reviews using Machine Learning Techniques", International journal of Computer Applications- December 2017 .

[4] Monu Kumar Thapar University, Patiala, " Analyzing Twitter sentiments through big data", IEEE 2016.

[5]Kigyon Lyu Korea University, Korea "Sentimental Analysis Using Word Polarity of Social Media", Springer, 2016.

[6] Minhoe Hur Seoul National University "Box-Office forecasting based on sentiments of movie reviews and Independent subspace method", Information Sciences ,2016.

[7] Jorge A Balazs University of Chile, "Opinion Mining and Information Fusion – A survey", 2015.

[8] Min Chen Huazhong University of Science ,China "BigData: A Survey", Springer 2014.

[9] Chirag Sagnani Standford University, USA "Sentiment Analysis of App Store Reviews", 2013.

[10] Martin Wollmer Technical University of Munich, Germany "Youtube movie reviews-Sentiment analysis in an audio-visual ", IEEE Computer Society,2013.

[11] Kamal A., "Review mining for feature based opinion Summaritzation and Visualisation,2015.

[12] Humera Shaiya, G Kavitha, Raniah Zaheer, " Text Categorization of Movie Reviews for Senitment Analysis", International Journal of Innovative Research in Science, Engineering and Technology, Vol.4 Issue 11,2015.

[13] A. Mudinas,D. Zhang,M. Levene, "Combining Lexicon and learning based approaches for concept-level sentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York, NY,USA, Article 5,2012.

[14] A. Khan,B. Baharudin, K. Khan, "Sentiment Classification from Online Customer Reviews using lexical contextual sentence structure" ICSECS 2nd International Conference on Software Engineering and Computer Systems,Springer,pp-317-331, 2011.