ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

A COMPARATIVE ANALYSIS OF TEXT CLASSIFICATION USING NAÏVE BAYES AND RNN-LSTM

Madhuri Dadhich[*1], Mausumi Goswami[2] & B.S Purkayastha[3]
[*1,2&3]CHRIST Deemed to be University, Bangalore, India

ABSTRACT

With the growth of Internet , the amount of data pro-duced is also increasing. Sentimental Analysis is the field in which the people's opinions, attitudes, sentiments, emotions and the eval-uations is analyzed from written text. There are various tasks in-cluded in sentiment analysis like extracting the sentiment, classi-fying the sentiments, summarizing of opinions or detecting of opinion spam among others. The main objective is to analyzing user's sentiments,, reviews, emotions etc. for various services of-fered , products, , subjects ,industries etc. The approaches used for the analysis were RNN-LSTM and Naive Bayes . In this paper, it is attempted to analyze the opinions of users using traditional machine learning approach and deep learning approach.

Keywords: *Text Categorization, Text classification, Text Anal-ysis, Sentiment Analysis, Social Network, Social media, Advertis-ing, sentiment, RNN-LSTM, Naive Bayes.*

I. INTRODUCTION

Sentiment means feelings of an individual which includes opinions, attitudes and emotions regarding any product or ser-vices. Sentiment Analysis is a type of text categorization. Text categorization is to categorize text based on some predefined labels or categories. For example, in any business if one can monitor the responses from customers on any product and clas-sify that response either positive, negative or neutral which will actually help the organization to either improve the product or to add extra features. Like ways Text classification can be used in many such applications which includes automating the CRM tasks, classifying panic conversations in social media, classify-ing movie reviews, product reviews, classifying content of any website and many more application areas are there where it can be used. Many single and multi-label text categorization meth-ods have been developed but the problem every method faces is dimensionality problem. To solve this also improvement has been done in order to classify documents or texts correctly. Classification is done to predict the possible results for any of the test instances given . A sample is not been included in the dataset used for training with same set of properties except for the prediction set, the instance should be correctly classified by the algorithm. The accuracy of the prediction determines how good the algorithm is. In this paper various approaches have been compared and gaps have been identified.

1.1 Sentimental analysis

Sentimental Analysis is the field in which the people's opin-ions, attitudes, sentiments, and emotions, evaluations is ana-lyzed from written text . Many surveyors or researchers spend long time in finding the reviews or opinion about particular pol-iticians speech or debate between people about particular topic in order to find if the public happy is with the governing leader or if there is a need to change the existing system people or the laws . Twitter, Google+ are popular as these permit people to express and share the opinions or views about various topics. There has been enough work done in sentiment analysis field using data from twitter. There is a lot of data generation every single day and it becomes difficult to analyze it in a manual way. So automation of things is required.

The dataset considered includes 13871 examples and 21 col-umns. The data set includes tweet data from various accounts.

II. METHODOLOGY

Text Categorization (also called as text classification) is used to classify the documents into categories from predefined sets. The method of assigning unlabeled documents into prede-fined categories is termed as Text categorization. Text catego-rization involves series of steps which includes: data collection, preprocessing the data, removing stop words, calculating weightage of each word, select the words which are having highest weightage, find the category that match to words and return the name of the category.

In the very first step data is collected and it is preprocessed to remove unwanted characters. Once the data is ready, stop words processing is done on the data. The output of this steps helps us to reduce the size of the data . Stop words are not con-sidered as important words as they contribute very less to eval-uate the real meaning of the sentence. The next task is to nor-malize the data, to perform stemming. Stemming process change the word to its root form. E,g, Word eat can have vari-ous forms and all the forms refer to the root word. For every word weightage may be calculated. Highest weighted words are selected . Next task is to categorize the word and returning the category.

The whole process is shown in Fig 1.

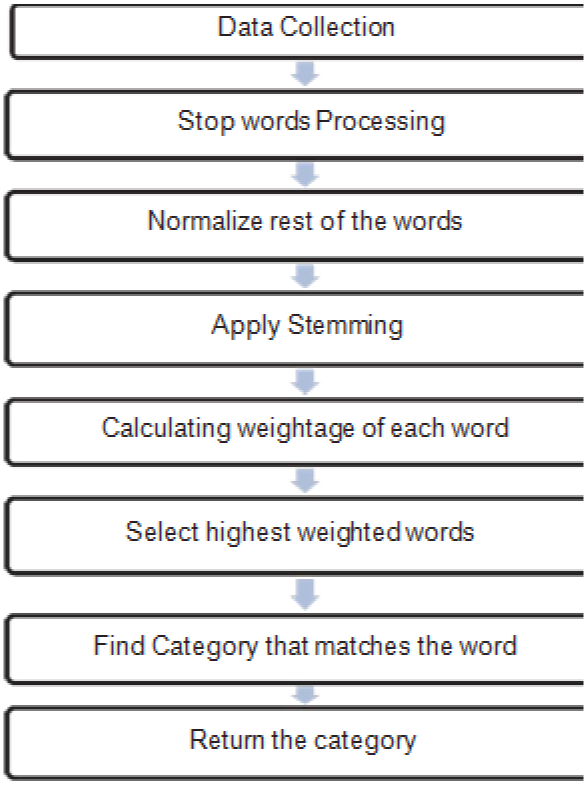
Data Collection
Stop words Processing
Normalize rest of the words
Apply Stemming
Calculating weightage of each word
Select highest weighted words
Find Category that matches the word
Return the category

Fig 1: Text Categorization process

2.1 Sentimental analysis methodology

Sentimental Analysis is the field which focuses on analyzingthe people's feelings, opinions, attitudes, sentiments, and emo-tions, evaluations from written text . Study says 92 percent of consumers or online buyers trust online reviews and consider such recommendations are as important as personal recommen-dations. Many surveyers or researchers spend long time in find-ing the reviews or opinion about particular politicians speech or debate between people about particular topic in order to find if the public happy is with the governing leader or if there is a need to change the existing system people or the laws .Online Sentiments can save researcher's time and public opinions about anything new introduced or changes in existing system can reach the leader which could actually help in developing the nation.

In this paper, implementation of a technique to analyze the First Republic Party debate in 2016 data which is a collection of tweets from people on First Republic Party debate in 2016.

2.1.1 Approaches used for Sentimental Analysis: The approaches used for Sentimental Analysis are: Machine learning is a type of artificial intelligence (AI) that can enable computer to learn without being programmed by humans. It is considered that there are mainly two categories of machine learning techniques : Supervised and Unsupervised Machine Learning Techniques.

Unsupervised learning: In general unsupervised learning ,there is no labelling of data inferences are drawn from data itself .It doesn't provides correct targets and rely on clustering.

Supervised learning: In this type labelling for the data is pro-vided before the data is being processed and then the data is trained to make decisions or predictions. The success of these two learning methods is essentially focused on selecting and taking out the specific features set which is used to detect sen-timent.

Common machine learning approaches are being used. Few such approaches are Naive Bayes(NB) , maximum entropy (ME), and support vector machines(SVM) , neural networks have been found successful in sentiment analysis.

It starts with collection of data and classifier is trained by mak-ing use of training data. After selecting supervised classifica-tion technique, feature is selected which tells how the docu-ment are represented. The different techniques that can used for sentimental analysis includes: Decision Tree, k-Nearest Neigh-bor, SVM Classifiers, Bayesian Classifiers.

Decision Tree : It has tree kind of structure where root node represents the question and sub nodes represented the answer. The next step after the tree creation, the document that needs to be classified is kept on root and traversed untill it reaches it leaf.The main disadvantage when using such technique is error rate which is high for a smaller training set.

k-Nearest Neighbor: The classification is performed by comparison of the category frequencies to the k nearest docu-ments(neighbors).The main disadvantage of the KNN algo-rithm is that training data itself is used for prediction or classi-fication which makes no sense.

SVM Classifiers: SVM Classifiers partitions the dataspace by using linear or non-linear delineations between different set of classes. speed and size, for both train dat and test data is the main disadvantage.

Bayesian Classifiers: The naïve bayes classifiers has inde-pendency , meaning one feature is independent of the other.Na-ive bayes works on conditional probablity .It can only process binary vectors which is the main disadvantage.

Artificial Neural Networks: Neural Networks (NN), also known as Artificial Neural Network . Neural network need not to tell how the problem to be solved,and it makes use of obser-vation data to solve the problem.

2.2 Preprocessing the data:

A lot of opinions are put in the tweets about the data which are from different users. The data set consist of a column named 'text' which contains the tweet from user which is la-beled into two classes called negative and positive polarity. These are defined under 'sentiment' column. This actually makes the process easy to observe. It helps to observe the ef-fect of various features.

Preprocessing of data set includes the following points,

- Removing URLs (e.g. www.pqr.com) from data, hash tags from data (e.g. #topic), targets (@username) etc.
- Removing Stop words
- Removing all symbols, numbers, and punctuations.

2.2.1 Model used for Sentiment Analysis:

The steps followed for sentimental analysis for First Repub- lic Party debate in 2016 are
1. Collecting the data
2. Extracting necessary column i.e. Text and sentiment col- umn
3. Dropping the neutral sentiments as main focus is on pos-sitive and negative sentiments
4. Removing the stop words and Urls , hash tags etc.
5. Applying Naive Bayes and RNN-LSTM (Long short term memory networks) .
6. Training and testing the data for both approaches.
7. Finding score and accuracy for both algorithms.

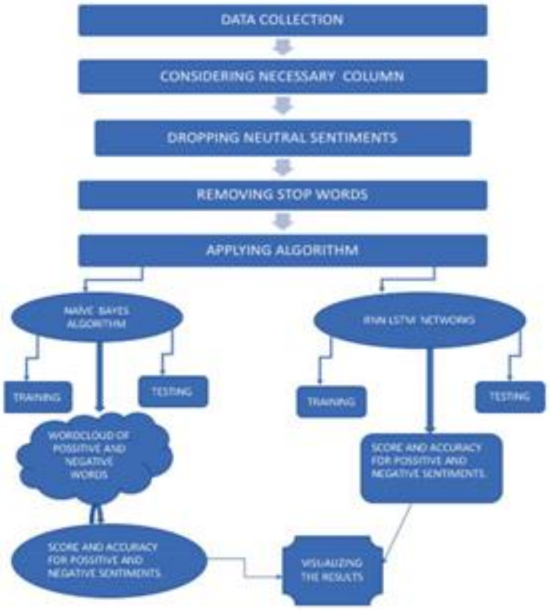
DATA COLLECTION
CONSIDERING NECESSARY COLUMN
DROPPING NEUTRAL SENTIMENTS
REMOVING STOP WORDS
APPLYING ALGORITHM
NAIVE BAYES ALGORITHM
RNN LSTM NETWORKS
TRAINING
TESTING
TRAINING
TESTING
WORDCLOUD OF POSSITIVE AND NEGATIVE WORDS
SCORE AND ACCURACY FOR POSSITIVE AND NEGATIVE SENTIMENTS
SCORE AND ACCURACY FOR POSSITIVE AND NEGATIVE SENTIMENTS
VISUALIZING THE RESULTS

Fig 2.1 Workflow Model

2.3 Naïve Bayes Approach:

Naive Bayes is a supervised learning method which can be used for statistical method as well as classification method .The name is inspired by the name of the scientist called Thomas Bayes (1702-1761). Thomas Bayes proposed the Bayes Theo-rem. It is a classification technique based on Bayes' Theorem . The assumption made is of independence among predictors. A Naive Bayes classifier assumes that the occurrence of a partic-ular feature in a class is unrelated to the occurrence of any other feature. For example, a fruit may be considered to be an orange if it is orange in color, about 1.5 inches in radius and round in shape. Even if these features are interdependent or may depend upon the existence of the other features, all of those features or properties independently pay or contribute to the probability that this fruit is an orange. This is why it is known as 'Naive'. Naive Bayes model is very useful for very large data sets. It is also considered to be a sim ple model.

Along with simplicity, Naive Bayes is known to outdo even highly classy and sophisticated classification method. The following were the steps used for Sentimental analyzing using Naive Bayes Approach :

1. First of all, dividing the dataset into a training and a testing set. The test set contains 10 percent of the original dataset. For this particular analysis neutral tweets were dropped, as goal
was to differentiate only positive and negative tweet.
2. As a next step separating the Positive and Negative tweets of the training set so that it can be easily visualized including their contained words. After that cleaning the text from mentions and links ,hashtags. Now Its ready for a Word Cloud visualization which will only show the most emphatic word from the Positive and Negative tweet Creating word cloud for both possitive and negative words.
3. After the visualization, hashtags, mentions, links and stop words were removed from the training set. Stop Words are the terms which are non-significant or non important. Because of this reason these words are filtered out from search queries. Stop words return vast amount of unnecessary information. (the, for, this etc.)
4. As a next step features were extracted , first by measuring a frequent distribution and by selecting the resulting keys
5. Creating word cloud for the most frequently distributed words.
6. Using the nltk NaiveBayes Classifier classifying the ex-tracted tweet word features
7. Measuring how the classifier algorithm scored and calculat-ing accuracies for both possitive and negative words. NaiveBayes Machine Learning algorithm for Sentiment Anal-ysis works well for negative comments. The problems arise when the tweets are ironic, sarcastic has reference or own dif- ficult context.

Consider the sample tweet:

*"Muhaha, So sad to know that the Trump couldn't destroyed by Liberals . forward Marching."

As it is thought that, the words **sad** and **destroy** influ-ences the evaluation, but this tweet should be positive when observed its meaning and context.

2.4 RNN LSTM Network:

Recurrent Neural Networks (RNN) are also one of the kind of neural networks which creates cycle by adding extra weights to the network. The core idea for RNN is to use the sequential information. When we talk about traditional or normal neural networks ,it is assumed that the layer inputs and output are in-dependent of each other But this won't work fine every task we are trying to compute using neural networks.

Let's take an example where one wants to predict next word of sentence or letter of a word ,in this case the previous words or letters should be known in order to make prediction for next word or letter, there where RNN comes into picture as they can the task with the results being depending on the previous out-puts .In general terms RNN have a memory which has data or information of whatever has been done till that state. In theory RNN can use the information arbitrarily for long sequences but in practical sense ,RNN can only go back for few steps.

The below figure shows RNN network unfolded to a full net-work, unfolding means writing the network for full sequence. For instance if the sequence has 10 words, the network could be unrolled to ten different layers i.e. one for each word.

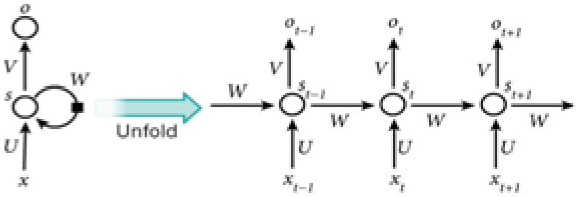

Fig 2.2 RNN Network

2.4.1 LSTM Networks:

Long Short Term Memory networks -generally called as "LSTMS" are special type of RNNs which are capable of learn-ing long term dependencies.They are specially designed for large problems and long term dependencies. LSTM also have chain like structure like RNNs but the modules that are repeat-ing has different structure, the layers interact in a special way.LS?TM use different function to compute the hidden state ,it can combine previous, current state and memory for better results. Conventiona LSTM contains units known as "memory" blocks in the hidden layer.The input gate in the model controls input activation into memory cell.The output flow is controlled by output gate.The forget gate controls what information needs to be kept or deleted.
LSTM has a special architecture which lets it to keep or forget the unnecessary information.

Steps used:

- The sigmoid layer takes input X(t) and h(t-1) and decides which parts from old output should be kept or removed.

2)The next step is to decide and store the next sate or input X(t) in cell state.
Sigmoid layer decides upon new information to be updated or ignored.

- Finally output is decided. Sigmoid layer decides which part of cell will go for output. Then cell state is put through a tan h which generates all possible states and then it is multiplied with output from sigmoid function

So LSTM can decide which information to be kept for long The following were the steps used for Sentimental analysing using LSTM Approach :

1.Taking only text and sentiment column from dataset.
2.As a next step dropping neural sentiments as our main focus is possitive and negative separation.
3.Filtering the texts with valid texts,t hen defing maximum fea-tures as 2000 and using tokenizer for vectorizing
4. Converting text into sequences so that network can deal with it as input
5. Composing LSTM Network using softmax function as the network is using categorical crossentropy and softmax function is right activation for that.
6.Defining training and testing data
7.Calculating accuracy and score
8.Visualizing the results.

III. RESULTS AND ANALYSIS

In this section the results of implemented techniques are tech-niques. The results showed that the LSTM(Long Short Term memory networks) worked best for the data set discussed and naive Bayes also worked fine with possitive tweets but the problem arise when the tweet contains negative words and the meaning of the tweet is possitive.

Two approaches results are given below in the proceeding sec-tions:

- The positive tweet or sentiment accuracy when using Naive Bayes approach is 44.07
- The negative tweet or sentiment accuracy when us-ing Naive Bayes approach is 94.04
- The positive tweet or sentiment accuracy when using LSTM approach is 49.31
- The negative tweet or sentiment accuracy when us-ing LSTM approach is 92.71

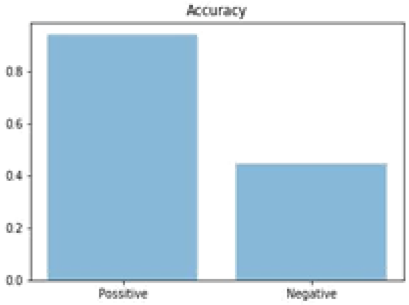

Fig 4.1: Naïve Bayes Accuracy

IV. CONCLUSIONS AND FUTURE WORK

Text categorization techniques have been compared and gaps have been identified. Text classification, categorization has many applications and results could be further improved if im-provised version of the existing algorithm or new technique is introduced. Sentimental analysis for GOP 2016 data was done and LSTM performed with a better accuracy. In this paper an attemapt is made to compare Traditional machine learning with Deep learning approach.

REFERENCES

1. *Guanyu Gao and Shengxiao Guan, "Text Categorization Based on Im-proved Rocchio Algorithm," International Conference on Systems and Infor-matics,pp.2247-2250,2012.*
2. *DashenXue and Fengxin Li, "Research of Text Categorization Model based on Random Forests," IEEE International Conference on Computational Intel-ligence & Communication Technology, pp. 173- 176, 2015. ̂*
3. *Shie-Jue and Jung –Yi Jiang, "Multi-label Text Categorization Based on Fuzzy Relevance Clustering," IEEE Transactions on Fuzzy Systems, vol. 22, no. 6, pp. 1457-1471, 2014.*
4. *R. Gayathri Devî and Sumanjani .P̂̂, "Improved classification techniques by combining KNN and Random Forest with Naive Bayesian Classifier ," 2015 IEEE International Conference on Engineering and Technology (ICETECH), 20th March 2015, Coimbatore, TN, India.*
5. *Sotiris Kotsiantis, "Increaing the accuracy of incremental naïve bayes clas-sifier using instance based learning", International Journal of Control, Auto-mation and systems, 2013.*
6. *M. Krendzelak and F. Jakab "Text Categorization with Machine Learning and Hierarchical Structures"*
7. *Suresh Yaram" Machine Learning Algorithms for Document Clustering and Fraud Detection , 2016 IEEE International Conference on Data Science and Engineering (ICDSE).*
8. *Tahira Mahboob, Sadaf Irfan, Aysha Karamat "A machine learning ap-proach for Student Assessment in E-Learning Using Quinlan's C4.5, Naïve Bayes and Random Forest Algorithms", 978-1-5090-4300-2/16/$31.00 ©2016 IEEE .*
9. *Thamarai Selvi. S, Karthikeyan. P, Vincent. A, Abinaya. V, Neeraja. G, Deepika. R," Text Categorization using Rocchio Algorithm and Random For-est Algorithm " 2016 IEEE Eighth International Conference on Advanced Computing (ICoAC) .*
10. *Thanabhat Koomsubha, Peerapon Vateekul "A Character-level Convolu-tional Neural Network with Dynamic Input Length for Thai Text Categoriza-tion ", 978-1-4673-9077-4/17/$31.00 ©2017 IEEE .*
11. *Luis Pinto and Andrés Melgar"A Classification Model for Portuguese Documents in the Juridical Domain*
12. *.Nidheesh Melethadathil ,Priya Chellaiah , Bipin Nair,Shyam Diwakar," Classification and Clustering for Neuroinformatics: Assessing the Efficacy on Reverse-Mapped NeuroNLP Data using Standard ML Technique", 2015 In-ternational Conference on Advances in Computing, Communications and In-formatics (ICACC*
13. *I. Budiseliü , G. Delaþ and K. Vladimir ,̃"Developing a Text Classifier With Constrained Development and Execution Time "*

14. *Mansur Alp Tocoglu, Adil Alpkocak "Emotion Extraction from Turkish Text ", 2014 European Network Intelligence Conference.*
15. *Madhuri Dadhich, Mausumi Goswami , "A Review of Text Categoriza*-tion *Techniques" , in the proceedings of National Conference on Challenges and Opportunities in Computer Engineering, 2018.*