International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

A Novel Random Forest Implementation of Sentiment Analysis

Arnav Munshi[1], Sanchit Sapra[2], M.Arvindhan[3]

[1]Student, School of Computer Science and Engineering, Galgotias University, Greater Noida, India
[2]Assistant Professor, School of Computer Science and Engineering, Galgotias University, Greater Noida, India
[3]Assistant Professor, School of Computer Science and Engineering, Galgotias University, Greater Noida, India

---

Abstract - *In today's world, we humans have been communicating with each other through calls, social media applications like WhatsApp, Facebook, Twitter, etc. From the social media apps, we get social media data from those applications and check what sentences are positive and negative sentiment using sentiment analysis and using deep learning methods like deep neural networks for classifying them under positive or negative sentiment polarity from twitter accounts. The data that we get from these social sites are being used for many social problems and used in government to analyze the opinion about social media users. This technique is called sentiment analysis. The main purpose of this sentiment analysis for this project will be to comparatively determine the writings made by user and check if they are going towards positive or negative. Only one technique will be used her – Machine Learning Algorithms – Random Forest. This paper uses the machine learning algorithm-Random Forest. The scope of this project will be widely used in classifying the text in terms of positive and negative polarity and help the government to handle social and threats due to the text classification.*

***Key Words*: Sentiment Analysis, Random Forest, NLTK.**

1. INTRODUCTION

There are about more than 100 micro blogging sites in today's world being used my millions of users across the globes through short messages and post. Like other micro-blogging sites, Twitter is also one of them where users daily post and share messages which is about millions of posts daily on an average and also comment about positive or negative views on a post made by someone. This kind of nature of humans making positive and negative comments are being supervised by many of the manufacturing companies where they see this as a way of earning profit regarding to the products and services and how the customer interacts with their products and services and this is how positive and negative comments and sentences or words got differentiated and studied leading to a new concept called SA(Sentiment Analysis).Sentiment Analysis , being a study of the positive and negative comments , sentences or even words and these techniques are now being used by many of the Companies to know about their product well like Amazon , Google etc and so the list goes on.For this particular paper , I have used sentiment analysis on the Hindi tweets dataset and checked that how are the sentences being classified into positive , negative or neutral sentiment . This can be achieved by using method – Machine Learning. ML - The best Machine Learning Algorithms for text classification will be Random Forest. Random forest or Random Decision forest are the algorithms used for Classification , regression consist of a large number of individual decision trees and each individual tree comes out with a class prediction.

2. LITERATURE SURVEY

"Sentiment analysis of Twitter Data" (Aggrawaal, 2014), used two combinations of models like: Unigram model, Tree Kernel Model and 100 Sentiment features model etc. The author used Support Vector Machine and report averaged 5 fold cross validation test results and also used a binary classification task with positive and negative polarity and chance of baseline occurred to be 50 percent. He also investigated 2 kinds of models: tree kernel and feature based models and founded that these two model outperform the unigram model. Sentiment Analysis using Deep Learning System, "Coooll" (Tang, Wei, 2015) which is a deep learning model that builds sentiment classifier from Tweets and manually annotated sentiment polarity. The author used two kinds of features which are: SSWE feature and STATE features. As proposed by the author , the SSWE feature learned from 10 Million Tweets consisting of positive and negative emoticons and been verified in positive and negative classification of Tweets.

" Sentiment analysis on Twitter Data" (Sahayak , Shete , 2015) , where the author uses Maximum Entropy method , Naïve Bayes Classifier and Support Vector Machine for Sentiment Analysis and the author concluded that three Machine Learning Algorithms used by him outperforms the model namely , unigram , feature based model , Kernel model using WEKA. He also concluded the difficulty increases with nuance and complexity of opinion expressed. "A Survey on : Classification of Twitter Data using Sentiment Analysis" (Deshpande , Joshi , Madekar , 2020) proposed a survey on Sentiment Analysis and surveyed on methods like Feature Selection , use of Twitter API for the extraction of tweets and using of Machine Learning algorithms like SVM and Naïve Bayes and also some data cleaning process like tokenization ,Stemming , Stop-words and so on.

"Sentiment Analysis using Deep Learning Techniques" (Kalaivani A , Thenmozhi D , 2019) , only described that sentiment analysis has 2 categories of techniques which are Machine Learning and Deep Learning The feature of Machine Learning approach namely unigram , bigram , trigram and used algorithm like Support Vector Machine and Naïve Bayes. The Deep Learning approach on sentiment analysis is very much efficient because they deliver impressive performance in NLP application and they don't need to be handpicked by doing the work themselves. Every single unit in Neural Network is simple and by stacking layer of NN units of one competent to learn highly sophisticated decision boundaries and Algorithms like RNN are also competent in Sentiment Analysis. "Twitter Sentiment Analysis using Deep learning method " (Ramdhani , 2017) , where the author used Deep Neural Network with including the specification of ReLU activation function , 3 Hidden Layers , Feed-forward Neural Network and using Mean Square Unit and Stochastic Gradient and preprocessing techniques like I mentioned earlier. The author concluded that Deep Neural Network achieved result of 75.03% and 77.45% and for MLP accuracy got 67.45% for train data and 52.05% on test data.

3. PROPOSED MODEL

In the previous topic, we looked at various methods used by authors ranging from Random Forest to Neural Network and now we have seen the existing models used by existing authors , Support Vector Machine was the Machine Learning algorithm mostly used and for this we will be using Random Forest so that we can compare our result among other existing results.

Now below are the steps for the proposed system that we are using in this project and will also some explanation about each step.

Step 1: Downloading Hindi tweets and translating to English Tweets dataset:

In this step we are using Github link to download the dataset in Hindi tweet dataset. Later on, we convert add another column of Translated English tweet from Hindi Tweets and then import using Pandas library.

Step 2: Data visualization and Analysis:

Data visualization will be done on the dataset using matplotlib library in Python and analyze the data by cleaning by removing noisiness in them.

Step 3: Removing Usernames, stop-words, punctuations, symbols:

In this step , after we have visualized the data in graph using matplotlib and seaborn library in python , we will be now doing the preprocessing of the dataset by removing user name , stop-words , Punctuations ,symbols etc. to use the dataset in proper format and also using methods like stemming , lemmatization and tokenization which are techniques used in NLTK.

Step 4: Using Features extraction methods – BOW, TF-IDF:

These methods – Bag of Words, TF-IDF will be used for creating features from text and will help in splitting the dataset into train and test dataset.

Step 5: Model Building –Random Forest Classifier:

In this step, after we have preprocessed the data in the proper format, we will be building the model using Random Forest Classifier. Random Forest Classifier is being defined as a large number of decision trees acting as a multiple machine learning algorithm to get better predictive performance as compared to using one decision tree algorithm acting as a class prediction on a individual level and class prediction's O/P individually make up votes and most votes with output make the model prediction O/P for Random Forest. The best part of Random Forest is that it can be used for Regression and Classification problem.

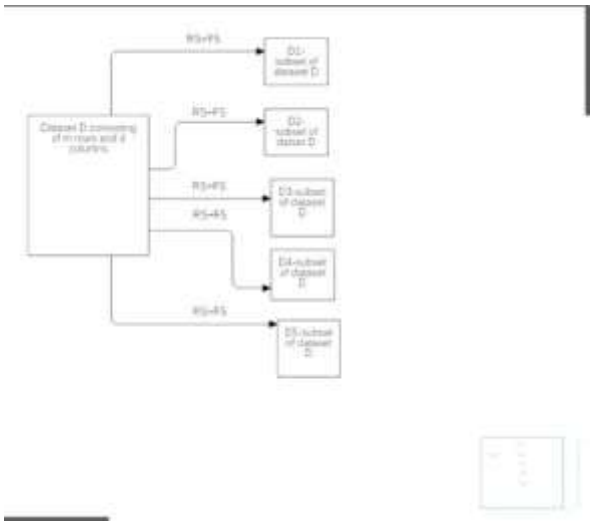

Figure 1: Architecture of Random Forest.

In this Architecture, there is a large dataset D consisting of m rows and d columns and since we are using Random Forest means, that there will multiple models of individual decision trees using the concept of Bagging, which means of using multiple machine learning algorithms.

To use the new records in Decision Trees from D1 to D5, Row Sampling and Column Sampling will be done for each individual trees.

Though with the Row Sampling and Column Sampling, there will be some rows and columns which will be same among (RS+FS) taken into other Decision Trees models.

Decision Tree D, will give prediction now and now the Test Data will be used in Decision Trees.

For example – binary classifier and similarly, the same test data will be used for other Decision Trees. These Decision Trees will produce output for Binary Classifier as 1,0,1,1,1 and from which majority vote will be taken as 1.

In case of Regression Situation, each individual tree will produce a continuous data, and the result of Random forest will be considered by taking the Mean of the Individual trees output or using the Median method.

Another thing to note about Random Forest, is that, the decision trees consist of two properties – Low Bias and High Variance.

Low Bias is defined as a concept where if the decision tree created, is being completed to the end, the result will be more accurate and the training error will be very less.

As compared to Low Bias, High Variance is defined when we get a new test data and apply on the decision tree , the decision tree has more tendency to give more error.

Now, the Random Forest works well by deciding that each Decision Tree will have a high variance , as having new test data but when each of these Trees will be combined and vote will be for majority , the high variance for Decision Trees will be converted to Low Variance.

If we have 1000 records, in a dataset and if we change 200 records, then also with the help of combined Decision Trees , the result will be in Low Variance resulting in good accuracy.

Step 6: Result with accuracy score and f1 score:

After the model is being fit using Random Forest Classifier and testing for prediction, we will use the accuracy score and f1 score.

With this we have explained the Proposed System and explained the some concepts.

4. CONCLUSION

The implementation of the project concludes that the f1 score is achieved by 0.66354 by using Random Forest Algorithm for building and testing the model. This proves that the Hindi Tweets dataset implemented by Random Forest has a f1 score of 0.66354 and accuracy score of 90.24. This results explains that:

This score of getting 0.66354 is somewhat similar to the results scored by the existing authors whom we specified in the Literature Survey. If we observe, then we see that the authors mostly used SVM (Support Vector Machine) and mostly Neural Networks which motivated us to change our method to use only Random Forest Classifier method and then use Hindi Language Dataset and after the we got the score we compared to previous authors and our score was mostly near to them as compared to most of them got results from 70 percent to 80 percent.

This sentiment Analysis being done of Hindi Language from Twitter Data has been gone more advanced with automation sentiment analysis and making the more use of the Machine Learning Algorithms .

REFERENCES

1. Kalaivani A , Thenmozhi D did a survey on Deep Learning where they presented paper on Sentiment Analysis using Deep Learning Techniques doing their survey published at IJRTE in April 2019.
2. Apoorv Agarwal , Boyi Xie , Ilia Vovsha , Owen Rambow , Rebecca Passonneau ,"Sentiment Analysis on Twitter Data.
3. Duyu Tang, Furu Wei(2014) developed a Deep Learning System , called "Coooll" which is a deep learning model that builds sentiment classifier from Tweets and manually annotated sentiment polarity.
4. Varsha Sahayak , Vijaya Shete (Jan 2015) , "sentiment analysis on Twitter Data."
5. Dimitris Effrosynidis (21st Conference on Theory and Practice of Digital Libraries).
6. Adyan Marendra Ramdhani(2017) , "Twitter Sentiment Analysis using Deep Learning."
7. Aanusha Ghosh and Indranil Dutta , "Real Time Sentiment Analysis using Hindi Tweets" .
8. Prateek Garg (2016), "Sentiment Analysis on Twitter Data using NLTK ".