Sentiment Analysis of US Airline Twitter Data using New Adaboost Approach

E. Prabhakar[a*],
[a] Assistant Professor,
Department of Computer Science and Engineering,
Nandha College of Technology, Erode – 638 052,
Tamilnadu, India.

M. Santhosh[b], A. Hari Krishnan[b],
T. Kumar[b], R. Sudhakar[b]
[b] Student,
Department of Computer Science and Engineering,
Nandha College of Technology, Erode – 638 052,
Tamilnadu, India.

***Abstract* -When it comes to decision making, internet is playing a significant role, all around the world. Many people use the blogs, social media and other online platforms to share their thoughts and views via internet. This results in the internet being filled, with full of relevant and irrelevant information. So it creates a great challenge of fetching the desired information over the internet by analyzing each and every document. Sentiment analysis paves the way on handling this problem at ease. This greatly helps customers in decision making on selection of best fit US Airlines on analyzing the other customer's opinion in online review sites like skytrax and other micro-blogging sites like Twitter which provides the Aspect level sentiment analysis. The proposed research methodology introduces new improved Adaboost approach for sentiment analysis. Various machine learning algorithms has been employed for identifying the appropriate algorithm for the system. Performance analysis has been performed based on the confusion matrix and accuracy of the algorithms.**

Keywords:* *Sentiment Analysis, Machine Learning, Adaboost, Twitter, US Airlines.

I. INTRODUCTION

An American online news and social networking service called Twitter is used by users for the purpose of posting of messages and interaction with other people known as "tweets". Twitter was created in March 2006 by Jack Dorsey, Noah Glass, Bis Stone and Evan Williams and launched in July of that year. The process of determining whether a Twitter mention conveys positive or negative attitude is called as Twitter sentiment Analysis. It is the process of analyzing it a piece of online writing expresses positive, negative or neutral attitude.

Airline industry is one of the largest and leading industries in the world which enables services to thousands of customers in a single day. Approximately 2,246,000 passengers adopt flights in the United States of America (USA) per day as per the reports provided by Federal Aviation Administration Air-Traffic [FAA]. This project research is focused on the top ten US based airline carriers namely America Airlines, Alaska Airlines, Delta Airlines, Jetblue Airlines, Hawaiian Airlines, Skywest Airlines, Southwest Airlines, United Airlines, Spirit Airlines and Us Airways are the top ten US based carriers of airline which is taken into account in this paper.

In USA, same geographical area is covered by these airlines during flight which makes it to fall under the primary position for choosing these airlines. Added to this these are the lost cost carriers in USA and also similar flight fare is identified. Furthermore, great competition is going on among them which force the competitors to create a good competitive edge. Not much more results have been found on the research on airline industry based on aspect of sentiment analysis. This research greatly focuses on to bridging the gaps between the customer's views and airlines carriers as a great milestone. Further the implementation of the proposed research occurs in other domains such as entertainment, education, automobiles, etc .

The rest of the paper is organized as follows: Section 2 discusses the literature review. Section 3 overviews the proposed methodology. Experimental results of the proposed scheme are presented in Section 4. Concluding remarks with future work are covered in Section 5.

II. LITERATURE SURVEY

Arockia Xavier Annie R of [1] has plotted the sentiment mining in graph. The sentiment analysis is taken in the form of movie review, product review, social discussion, trip review, etc. It made a difference between computer and human. Airline industry is the growing billion dollar industry and it also have millions of passengers. Sentiment analysis of movie review comments is also taken. It is based on the classification of twitter data into positive, negative and neutral command. The importance of public opinion mining and tools used are discussed in [2[, [3].

The paper [4] discussed about the opinion mining approach for twitter data. [5] given the different ensemble approach to deal with sentimental analysis. In [6], the authors has developed an ensemble sentiment classification system of Twitter data for airline service analysis. They also used lexicon based approach which is done by using lexicon dictionary. Lexicon based approach calculates the sum of the number of positive sentiment words and the negative sentiment words appearing in the text file.

Many research have been proposed a model to evaluate the usefulness of lexical resources as well as feature collect

information in formal and creative language used in blogging. Sentimental analysis for Airline Twitter data is done with the help of feedback forms or online questionaries' in their respective websites. The data was obtained from tweets based on United Airlines controversy and then it was classified. Sentiment analysis is a latest trend to understand the needs of mass public. It also uses Naive Bayes theorem. There is a more scope for improvement in sentiment analysis since it is new and yet not been tested.

In [8], the authors has proposed the sentiment mining which is used to address the problems. It is found by comparing the tweets against a predefined corpus of subjective words. Bayes theorem is used here. Bayes theorem is used to evaluate the posterior probability. They have conducted the case study for three airlines namely Air train, Frontier and Sky west. In [10], the importance of feedback taken into consideration.

In [5], the authors has determined that Online media is growing very fast in these recent years which created a need for sentiment analysis. Purpose of studying website is used to extract reviews and feedbacks from the customer and to determine the customer likes and dislikes on that specific product. It can automatically extract people's reactions, opinions and feed back towards a product. It is based on text mining. A natural language processing techniques. It is employed in e-commerce, transportation and automobile industry. In [10], the new ensemble based sentiment classification system of twitter data for airline services analysis has been proposed. There are two types of sentiment. One is lexicon based sentiment and another one is learning sentiment. Lexicon based sentiment is done by using lexicon dictionary. It collects the feedback of travelers regarding airline companies.

In [10], [11] the authors has presented new ensemble approach to improve the accuracy. The authors also introduced new type of improved version of boosting approach to obtain the better accuracy and to handle imbalanced data.

III. SYSTEM DESIGN

The proposed methodology consists of following steps:

- ✓ Dataset Collection
- ✓ Data Pre-processing
- ✓ Data Mining Techniques
- ✓ Performance Analysis
- ✓ Identify the Best Model
- ✓ Apply the Model

A. DATA COLLECTION

Data collection is the foremost step in this methodology. The data has been derived from different sources like Twitter tweets and online reviews from skytrax and tweets of Twitter for the top 10 US based airline carriers (2014-2017).

B. DATA PRE-PROCESSING

Next to data collection, data pre-processing occurs in which the data has been refrained by eliminating unwanted data. This process holds a very important possession as this process helps to construction of an efficient, stable and robust model. Each and every tweet will be undergone a conversion to a sentence after this process of data pre-processing.

C. SUPPORT VECTOR MACHINE

Support vector machine briefed as SVM is one of the supervised learning methods in the field of computer technology involving science and statistics. Support Vector Machine aims to analyze the data and to recognize the patterns. Classification and the regression analysis are also dealt by it. Data could be separated linearly which allow researchers to identify the two hyperplanes in margin. SVM aids in splitting the data with hyperplane and also extend the non-linear boundaries with the help of the kernel trick. Thus SVM would perform classification method properly for classifying the data present. It can be described mathematically as follows,

$$x_i.w + b \geq +1 \quad for\ y_i = +1$$

$$x_i.w + b \leq -1 \quad for\ y_i = -1 \qquad \text{----------1}$$

Above two equations are combines to form one set of differences as expressed below,

$$y_i(x_i.w + b) - 1 \geq 0 \qquad \forall i \qquad \text{-------2}$$

Thus,
x represents the vector point
w Represents the weight parameter could be a vector

Data splitting take place based on value obtained from above equation. It is acceptable to split the input data, if the result is higher than 0. SVM would select the point one having distance as longer from hyperplane among all. Available margin could be maximized by the hyperplane and it bisects the lines in closest points. The above process would occur only if hyperplane which has been chosen is possibly at the longest point from data. Additionally, this approach has been applied on points at other side. The Summed Distance could be defined on splitting the hyperplane to the nearest points. Finally summed distance is obtained on means of subtracting and solving the presently available 2 distance.

D. DATA MINING

Statistical operations likely correlation and linear regressions and correlation were performed on the aspects and polarity score for the airlines. Thus the better insight of the model could be obtained by these statistical data and also aids in understanding the relationships among different factors in the data set.

E. INTERPRETATION, EVALUATIONS AND VISUALISATION

The evaluation of the research has been performed based on values obtained from precision, recall and f-score. Various machine learning algorithms were being performed in order for identifying the best fit algorithms for the system. A case

study has been proposed at last to provide answer to all the research questions in addition to that of visualizing the results.

F. New AdaBoost Approach

This proposed methodology integrates the boosting and bagging ensemble. It considers the advantages of both the algorithms.

Step 1: Import the dataset
Step 2: Give 75% data for training and remaining data for testing.
Step 3: Apply "n" different base learners for Bagging
Step 4: Assign train dataset to the models
Step 5: Update the weight based on the misclassification rate
Step 6: Choose the best learners among "n" different base learners
Step 7: Apply the selected learners for Bagging
Step 8: Now create the best combination for Bagging
Step 9: Create the model using new approach
Step 10: Make predictions for test dataset and calculate accuracy

IV. RESULT

In this model the sentiment analysis has been performed by means of comparison with various ensemble algorithm in which the precision, recall and F-score has been obtained.

TABLE I: Performance Analysis

Algorithm	Precision	Recall	F-score
SVM	0.74	0.67	0.65
Decision tree	0.72	0.68	0.63
Random Forest	0.71	0.66	0.60
Bagging	0.68	0.67	0.63
Boosting	0.68	0.66	0.58
New Approach	0.78	0.65	0.68

From the above table it is clear that new approach is the best fit algorithm as the F-score is found to be high for this algorithm, it is known to provide the appropriate result in sentiment analysis on airlines based on twitter data. Thus the accuracy of outcome makes the customers to understand and choose the best airline carriers based on the model.

V. CONCLUSION

An Aspect- based sentiment analysis has been implemented in the research by employing the open NLP library. However, the Random Forest has provided the best F-score of over to percent. There are two main segments in this project. The main categories of the reviews were identified first. Next, the sentiment analysis has been performed for the aspects or categories detected from reviews.

Research performed on Airline industry particularly by employing sentiment analysis based on aspect is minimal in the part. Significant contribution has been provided by this research to the customers in deciding the airlines and also helps in bridging the gap between the carriers and the customers. This research greatly aids the US airlines to look for the areas of improvement and could easily make comparison of their performances with their competitors for obtaining a better competitive edge in the market. In this research, there are short comings too. Only for English language, the model used for this research can be applied. Since there will be different grammatical structure for different languages. Thus for other language this model will not work for other languages. urther, it requires user's input for adding or changing the existing features.

REFERENCES:

[1] Arockia Xavier Annie ,Vignesh Mohan, Sree Harissh Venu"Sentiment Analysis Applied to Airline Feedback to Boost Customer's Endearment", MSIG, 2015.
[2] E.Prabhakar, R.Parkavi, N.Sandhiya, M.Ambika, "Public Opinion Mining For Government Scheme Advertisement", International Journal of Information Research and Review, Volume 3, Issue 4, Page No.2112-2114, April 2016.
[3] E.Prabhakar, G.Pavithra, R.Sangeetha, G.Revathy, "Mining Better Advertisement Tool for Government Schemes", International Journal for Technological Research in Engineering, ISSN (Online): 2347 – 4718, Volume 3, Issue 5, Page No.1023-1026, January 2016.
[4] Pranika Jindala Varun Jaiswala and M. Umac, "Opinion Mining of Twitter Data for Recommending Airlines Services", International Journal of Control Theory and Applications, 2016.
[5] HebaHakh, Ibrahim Aljarah, Bashar Al-Shboul "Online Social Media-based Sentiment Analysis for US Airline companies", New Trends in Information Technology, 2017.
[6] Nadia F.F. da Silva, Eduardo R. Hruschka, Estevam R. HruschkaJr, "Tweet sentiment analysis with classifier ensembles" Science direct, 2014.
[7] Yun Wan, Dr. Qigang Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis" IEEE, 2015.
[8] Yun Wan, Dr.Qigang Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis", IEEE, 2015.
[9] EsiAdeborna, KengSiau, "An approach to sentiment analysis the rating case of airline Quality rating", PACIS, 2014.
[10] E. Prabhakar and K. Sugashini, "New Ensemble Approach to Analyze User Sentiments from Social Media Twitter Data", The SIJ Transactions on Industrial, Financial & Business Management (IFBM), Vol. 6, No. 1, 2018.
[11] E. Prabhakar, "Enhanced AdaBoost Algorithm with Modified Weighting Scheme for Imbalanced Problems", The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 6, No. 4, 2018.