EasyChair Preprint

№ 3409

Sentimental Classification Method of Twitter Data for Indian Air Asia Services Analysis

Rajat Yadu and Ragini Shukla

May 15, 2020

SENTIMENTAL CLASSIFICATION METHOD OF TWITTER DATA FOR INDIAN AIR ASIA SERVICES ANALYSIS

Rajat Yadu[1], Dr. Ragini Shukla[2],

[1]Research Scholar, Dr. C.V. Raman University Kota Bilaspur (C.G.).

[2]Assistant Professor, Dr. C.V. Raman University Kota Bilaspur (C.G.).

ABSTRACT

In India there are many airline services which provide various types of services to their customers such as food beverage, important entertainment, Seat comfort, Staff Service and flight punctuality but they cannot express their feedback immediately so twitter provide a sound of data source for them to do customer sentiment analysis. In this research paper we focused on only Air Asia Airline services and collect data from twitter which is a huge SNS (Social Networking Site) with user posting by using WebHarvy tool. In our experiment, this analysis was carried out using 5 different classification strategies: Decision Tree, Random Forest, SVM, Logistic Regression, and Naive Bayes. The outcome of the test set is the tweet sentiment (positive/negative/neutral) with 3 class dataset and calculate the performance in terms of accuracy. We have achieved best accuracy 79.53% in case of Logistic Regression classifier. In this paper we are classifying sentiment of Twitter messages by exhibiting results of a machine learning algorithm using Rapid Miner. The tweets are extracted and pre-processed and then categorizing them in neutral, negative and positive sentiments finally summarizing the results as a whole.

Keywords— Sentiment Analysis, Machine Learning techniques, Lemmatization, Twitter Analysis.

I INTRODUCTION

The sentimental analysis has became most important part of machine learning and In airline services the Customer feedback is very crucial for Airline companies because this helps them in improving the quality of services and facilities provided to the customers. Sentiment Analysis in Airline industry is methodically done using traditional feedback methods that involve customer satisfaction questionnaires and forms. These procedures might seem quite simple on an overview but are very time consuming and require a lot of manpower that comes with a cost in analyzing them. Moreover, the information collected from the questionnaires is often inaccurate and inconsistent. This may be because not all customers take these feedbacks seriously and may fill in irrelevant details which result in noisy data for sentiment analysis (Kumar et al. 2019). Whereas on the other hand, Twitter is a gold mine of data with over 1/60th of the world's population using it which nearly amounts to 100 million people, more than half a billion tweets are tweeted daily and the number keeps growing with every passing day. With the rising demand and advancements of Big Data technologies in the past decade, it has become easier to collect tweets and apply data analysis techniques on them .Twitter is a much more reliable source of data as the users tweet their genuine feelings and feedbacks thus making it more suitable for investigation(Rane et al. 2018). Once the airline tweets are collected, they undergo pre-processing to remove unnecessary details in them. Sentiment classification techniques are then applied to the cleaned tweets.

The main motive of this paper is to provide the airline industry a more comprehensive view about the sentiments of their customers and provide to their needs in all good ways possible (Pugsee et al. 2015). In this paper, we go through several tweet pre-processing techniques followed by the application of five different machine learning classification algorithms that are used to determine the sentiment within the tweets. The classifiers are then compared against each other for their accuracies.

II. RELATED WORK

In the field of airline services there are many researches has been done for sentimental analysis where most of the studies analyze the twitter data extracted with respect to airline industry. Several popular major airline companies are selected across the world based on their followers and number of tweets on the Twitter (Hemakala et al. 2018). Tweets were extracted, preprocessed and converted into feature vectors (numerically suitable form) using n-gram and GloVe dictionary (WE) approach for analysis. Initially, several ANN models were developed and tested along with SVM on pre trained, trained and hybrid word embedding dataset (Rustam et al. 2018). Most existing classification techniques used include ensemble approaches such as AdaBoost (Ensemble) which combine several other classifiers to form one strong classifier and give an accuracy of 84.5% (Prabhakar et al. 2019).The accuracies attained by the classifiers are high enough to be used

by the airline industry to implement customer satisfactory investigation but this is not possible at the present work (Adarsh et al. 2018).

Few studies related to the first time experience where translating passenger data from "Airline" timeline to a timeline relative to each passenger first flight has been considered with yields high accuracies. Furthermore, taking advantage of recently made available data mining libraries we out performed simple extrapolation models and previous works.

III. PROPOSED WORK

A. DATA EXTRACTION

In this research paper, the dataset contains various tweets that were taken from the standard Web harvy tool where firstly we have created multiple fields related to the customer's feedback but we only distribute the tweet section and polarity class or sentimental class labels such as the given feedback was how positive or negative or neutral. Indian Air Asia released there customer's tweets in own website. A total of 500 tweets were extracted which formed the experimental dataset. The tweets were a mix of positive, negative and neutral sentiment. The tweets are pre-labeled with the type of sentiment which led us to follow the approach of supervised machine learning. The implementation is done by using Rapid Miner where dataset has been imported and five major classification techniques have applied for correct classifying sentiments with best accuracies. The following table gives the tweets sentiment distribution.

B. DATA PREPROCESSING

Data preprocessing is a data mining technique that transforms real world data into understandable format. Twitter data is often inconsistent and lacks certain features (missing values) which need to be dealt with before performing any kind of analysis (Samonte et al. 2017). The tweets undergo various stages of preprocessing to get the cleaned tweets which can be used for further analysis. In our research we have found various tweets where no objective or no stop word has included specifying the sentimnt.The tweets are tokenized which transforms the tweets into a list where each word in the tweet stages of preprocessing to get the cleaned tweets which can be used for further analysis. The tweets are tokenized which transforms the tweets into a list.

Table 1. Sentiment Distribution of Tweets

Sentiment Tweet	Counts
Positive	479
Negative	09
Neutral	11

The collected has been transformed to the three class labels such as positive, negative or neutral by removing or eliminating the stop words by using lemmatization which is the process of reducing the word to its base form with the use of vocabulary. The sentimental analysis is done through the classification technique Rapid Miner.

IV METHODOLOGY

In this paper, we use main five classifiers to for text classification that can be also used for Air Asia Services tweets sentimental analysis. In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the input data and then uses this learning to classify new observations (Khan et al. 2018). The techniques are as follows-

A. Linear Regression- It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. This is better than other binary classification like nearest neighbor since it also explains quantitatively the factors that lead to classification.

B. Random Forest- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees (Rane et al. 2018). Random decision forests correct for

decision trees' habit of over fitting to their training set.

C. **Naïve Bayes-** It is a classification technique based on Bays Theorem with the assumption of independence among predictors. In other words, Naive Bayes classifiers assume that the presence of a particular feature in a class is unrelated to the presence of any other feature or that all of these properties have independent contribution to the probability (Maxson et al. 2018). This family of classifiers is relatively easy to build and particularly useful for very large data sets as it is highly scalable.

D. **Decision Tree-**
Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes (Tiwari et al. 2019). A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

E. **Support Vector Machine(SVM)-**
This algorithm works on a simple strategy of separating hyper planes. Given training data, the algorithm categorizes the test data into an optimal hyper plane. The data points are plotted in a n-dimension vector space (n depends upon the features of the data points). SVM algorithm is used for binary classification and regression tasks but in our case, we have a 3-class sentiment analysis making it multiclass SVM classification. We adopt the pair wise classification technique where each pair of classes will have one SVM classifier trained to separate the classes (Shirbhate et al. 2016). The overall accuracy of this classifier will be accuracies of every SVM classification included. Then on performing classification we find a hyper plane that differentiates the 3 classes very well. Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible.

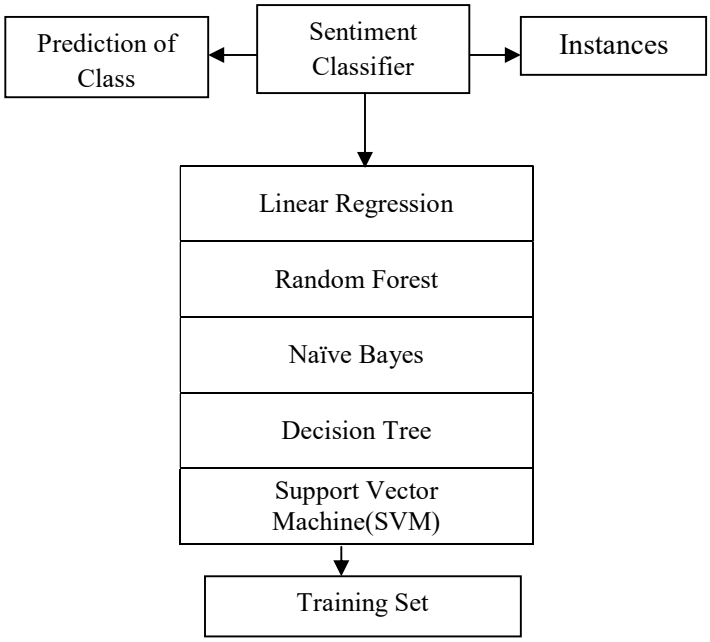

Figure 1. Sentiment Classification System

V. EXPERIMENT AND EVALUATION

The dataset consists of 500 tweets on which we perform a train-test and remove non sense vocabulary. The overall sentiment count which accounts for the total number of tweets in each sentiment category i.e. positive, negative or neutral for all Indian Air Asia Airline Services was visualized in Figure.2

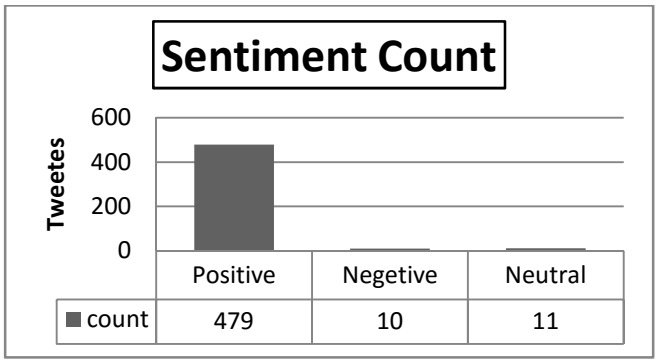

Figure 1. Sentiment Count

using Webharvy dataset and manual Lemmatization process, majority of the tweets expressed positive sentiment, this maybe because people generally use the social media platform to convey their satisfactory remarks.

The five classifiers which are shown in figure 1where trained using the training data and tested on the test set for their accuracies. In accuracy evaluation,

consider precision, recall evaluate the overall accuracy of the classifier. Here, precision is the fraction of correctly classified instances for one class of the overall instances which are classified to this class and recall is the fraction of correctly classified instances for one class of the overall instances in the dataset. The Table II shows the accuracies of each classifier. The reasons for the positive feedback from the customers as mentioned in the dataset were also plotted and presented in the form of a graph in Figure.4.

Table 2. Accuracy of Classifiers

Classifiers	Accuracy
Linear Regression	79.53%
Random Forest	78.84%
Naïve Bayes	76.53%
Decision Tree	76.54%
Support Vector Machine(SVM)	50.00%

There are many reasons of positive tweets such as Good Flight, Tell Late Flight Schedule before time, solving Customer Service Issue Flight Booking Problems etc.

VI. CONCLUSION

In this work , we have used various Classification techniques include five techniques in which Linear Regression gives high accuracy of 79.53%.The accuracies attained by the classifiers are high enough to be used by the air Asia Company to implement customer satisfactory investigation. This paper makes an effective contribution to the air line services research area by comparing the performance of different classifiers, which future improves the sentiment classification approach. In this paper, we compare various traditional classification techniques and compare their accuracies. The past work that has been done does a Country level analysis of tweets without preserving the word order. However, in this research have done multi-level analyses of tweets using Sentimental Classifier System. We use Rapid Miner software to classify the sentiments. In the future we are planning to further expand our research and analysis by gather a huge number of data and expanding the process of text mining involved in this analytical approach.

REFERENCES

[1] Ahuja, V., & Shakeel, M. (2017). Twitter Presence of Jet Airways-Deriving Customer Insights Using Netnography and Wordclouds. *Procedia Computer Science*, *122*, 17–24.

[2] Adarsh, M. J., & Ravikumar, P. (2018). An Effective Method of Predicting the Polarity of Airline Tweets using sentimental Analysis. *Proceedings of the 4th International Conference on Electrical Energy Systems, ICEES 2018*, 676–679.

[3] Berengueres, J., & Efimov, D. (2014). Airline new customer tier level forecasting for real-time resource allocation of a miles program. *Journal of Big Data*, *1*(1), 1–13.

[4] Breen, J. O. (2012). *Mining Twitter for Airline Consumer Sentiment*. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*.

[5] Ceccato, V., & Masci, S. (2017). Airport Environment and Passengers' Satisfaction with Safety. *Journal of Applied Security Research*, *12*(3), 356–373.

[6] Hemakala, T., & Santhoshkumar, S. (2018). Advanced Classification Method of Twitter Data using Sentiment Analysis for Airline Service. *International Journal of Computer Sciences and Engineering*, *6*(7), 331–335.

[7] Khan, R., & Urolagin, S. (2018). Airline sentiment visualization, consumer loyalty measurement and prediction using twitter data. *International Journal of Advanced Computer Science and Applications*, *9*(6), 380–388 .

[8] Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. *Journal of Big Data*, *6*(1), 1–16.

[9] Langley, P., & Carbonell, J. G. (1984). *Approaches to machine learning*. *Journal of the American*

Society for Information Science (Vol. 35).

[10] Liu, B. (2012). *Sentiment Analysis: A Fascinating Problem*. *Sentiment Analysis and Opinion Mining*.

[11] Maxson, R. W. (2018). Prediction of Airport Arrival Rates Using Data Mining Methods.

[12] Pérezgonzález, J. D., & Gilbey, A. (2011). Predicting Skytrax's Official World Airline Star ratings from customer reviews, (October), 48–50.

[13] Prabhakar, E., Santhosh, M., Krishnan, A. H., Kumar, T., & Sudhakar B B Student, R. (2019). Sentiment Analysis of US Airline Twitter Data using New Adaboost Approach, *7*(01), 1–3.

[14] Pugsee, P., Chongvisuit, T., & Nakorn, K. N. (2015). Opinion mining on Twitter data for airline services. *2015 5th International Workshop on Computer Science and Engineering: Information Processing and Control Engineering, WCSE 2015-IPCE*, 639–644.

[15] Rane, A., & Kumar, A. (2018). Sentiment Classification System of Twitter Data for US Airline Service Analysis. *Proceedings - International Computer Software and Applications Conference*, *1*, 769–773.

[16] Rustam, F., Ashraf, I., Mehmood, A., Ullah, S., & Choi, G. S. (2019). Tweets classification on the base of sentiments for US airline companies. *Entropy*, *21*(11), 1–22.

[17] Samonte, M. J. C., Garcia, J. M. R., Lucero, V. J. L., & Santos, S. C. B. (2017). Sentiment and opinion analysis on twitter about local airlines. *ACM International Conference Proceeding Series*, 415–422.

[18] Shirbhate, A. G., & Deshmukh, S. N. (2016). Feature Extraction for Sentiment Classification on Twitter Data. *International Journal of Science and Research (IJSR)*, *5*(2), 2183–2189.

[19] Singh, S., Pareek, A., & Sharma, A. (2019). Twitter Sentiment Analysis using Rapid Miner Tool. *International Journal of Computer Applications*, *177*(16), 44–50.

[20] Tiwari, P., Pandey, H. M., Khamparia, A., & Kumar, S. (2019). Twitter-based opinion mining for flight service utilizing machine learning. *Informatica (Slovenia)*, *43*(3), 381–386.

[21] Yakut, I., Turkoglu, T., & Yakut, F. (2015). Understanding Customer's Evaluations Through Mining Airline Reviews. *International Journal of Data Mining & Knowledge Management Process*, *5*(6), 01–11.