RESEARCH ARTICLE OPEN ACCESS

# Analysis of Airline Tweets by Using Machine Learning Methods

Ufuk Bezek*, Parvaneh SHAMS**
**Istanbul Aydin University, Computer Engineering Department, Istanbul / Turkey*
***Istanbul Aydin University, Computer Engineering Department, Istanbul / Turkey*

**ABSTRACT**
As technology advances, the airline industry has also grown rapidly in recent years. During this growth, the feedback from customers had a great importance. Airline companies use the feedback forms they prepare to get feedback from customers. However, due to the long and difficult process of collecting, reviewing these forms they use and taking actions for these forms, the use of social networking areas by customers through technology has increased. Twitter comes to the fore in terms of getting quick feedback from these social sharing areas. Twitter is a good resource for collecting customer feedback tweets and making a sentiment analysis. In this study, a dataset containing tweets for 7 different airlines was studied. As dataset, the dataset with open access in UCI is used. There are tweets about 15000 airlines in this data set. This study was made by using 6 different classification and optimization algorithms. In the algorithms used, 80% of the data is reserved for training and the remaining 20% is used for the testing stage. As a result of the analyzes and comparisons, the highest result was achieved with 92.6% nadam optimization.In addition, the highest success rate was achieved among the algorithms used in other studies.

**Keywords:** Airlines, Twitter, Machine Learning, Artificial Neural Networks

Date of Submission: 20-06-2020 Date of Acceptance: 07-07-2020

## I. INTRODUCTION

Customer feedback is very important for airline companies to further improve the quality of their services rendered and facilities to their customers. Sentiment analysis in the airline industry is made by using traditional feedback methods including customer satisfaction surveys and forms.These procedures may seem quite simple first, but they are very time consuming and costly to analyze them, and they also require intense manpower. In addition to that, the information collected from surveys is often incorrect and inconsistent. This may be because not all customers take these feedback seriously and fill in irrelevant details that result in noisy data for sensitivity analysis. Twitter is a social platform used by more than 1/60 of the world population and reaches about 100 million people [1]. It continues to grow day by day. As a result of increasing demands and developments of big data technologies in the last decade, it has become easier to tweet and apply data analysis techniques to them. After processing and editing the collected data, the opportunity to make healthier and faster feedback for companies was provided by using appropriate classification algorithms for analysis. With the diffusion of the Internet and, accordingly, social media entering our lives in all areas, people have become more free to find an addressee for every situation they encounter. Along with the technology, the complexity disappeared, resulting in a more regular and faster communication situation. The studies that we encounter in the literature search on this subject are; in the study of Ankine Rane and Anand Kumar on the same dataset, they achieved the result of 85.6% with Random Forest metho [2]. In the study made by Joshua Acosta et al., sentiment analysis via twitter messages was found to be 72% as a result [3]. Shyamasundar and Jhansi Rani achieved a result of 89.61% in their sentiment analysis studies [4]. Twitter is a much more reliable data source because users tweet their true feelings and feedback, thus it becomes more suitable for research. After airline tweets are collected, they are subjected to pre-process to remove unnecessary details within them. Sensitivity classification techniques are applied to these cleaned tweets. This provides data scientists and airline companies with a broad scanning on their customers' feelings and thoughts. The main objective of this study is to provide the airline industry with a more comprehensive opnion about their customers' feelings and to meet their needs in the best possible manner. In this study, a few tweets were subjected to the pre-processing technique. Then 6 different machine learning classification algorithms and optimization algorithms used to determine the emotions in tweets were applied.

## II. OBJECTIVE

For the study, the Tweets UCI dataset with 15000 data was used. It is planned by using the data to identify the negative rate of the tweets about the airline company by using machine learning methods. In this data set, 9106 people tweeted negatively. The dataset contains different attributes such as tweet id used for each tweet, emotion of tweet, emotional correctness of negative tweet, correctness of tweet, reason for negative tweet, correctness of reason of negative tweet, airline company, name of the person who tweeted, number of retweets of the tweet, content of tweet, coordinate of tweet, date of tweet, location of tweet, time period used by the person who tweeted, etc.

A total of 13 attributes were used to find the sentiment analysis status of the tweets posted about airline companies. A total of 6 machine learning methods were applied to these attributes. It was intended to estimate the negative percentage of the tweets posted about the airline companies with the algorithm that gives the best results by finding the results of the applied methods.

## III. SCOPE

This study can be used primarily in air transport and also in many sectors by developing. Considering that most people use airline today; it can be showed as an activity to quickly and easily solve the problems that may arise. It was thought by the machine learning methods used in this study that it would be helpful to quickly evaluate the situations that customers deem negative and to reach a solution. According to the result of this study, it can be used to assist customers in evaluating the problems, not solving their problems.

## IV. METHOD

The application is written in Pyhton language. The study was made on a laptop with 16 GB Ram, 4 GB Graphics Card, i7 10th generation processor. It was run 100 times in total for the program to work properly. Numpy, Pandas, Matplotlib, libraries were used for the graphics used in the study and Scikit-Learn library was used for the algorithms used in the study. The normalized dataset was processed separately with machine learning algorithms. 4/5 (80%) of 14873 data entered training, and the remaining 1/5 (20%) of the data were used for testing. It was intended in this study to further increase the value found in previous studies and increase the accuracy rate by using different algorithms.

### 4-1-Support Vector Machines

The purpose of the support vector machines is to play a role in defining the hyper-plane, which ensures the best separation of the two classes available [5]. It is also appropriate to be used in big data. This algorithm essentially uses the statistical learning approach.

### 4-2- Decision Tree

Decision tree is one of the most used classification algorithms. The decision tree divides every incoming data into two by splitting it as yes-no. It takes the variables in the dataset as a node. Branching takes place by analyzing whether the variables in the node are realized and not. This algorithm does not use the assumption of basic data distribution in data distribution, it enters training with data that it has not consider as true before. After all the data is classified, branching ends and the leaves of the tree form the class labels [6].

### 4-3- Random Forest

Random forest creates a lot of decision trees. It is a simple learning method used for the mode of classes or for other tasks that handle the class, which is the average estimate of individual trees [7]. After the parameters are received from the user, the data set is taken to training and used for learning. It is one of the most widely used algorithms. It is used for classification and regression. It is an algorithm with a high success rate, because the number of trees can be determined as desired [8].

### 4-4- Gradient Boosting

Gradient Boosting is a gradient boost algorithm. It is one of the machine learning methods used to solve gradient boosting, classification and regression problems. This method is a model created by combining multiple weak estimation models [9]. It is used as a machine learning model that develops a prediction model in the form of a set of low prediction models, usually decision trees, to solve problems such as regression and classification [10].

### 4-5-Naive Bayes

The Naive Bayes classifier greatly facilitates learning, especially assuming that the features are independent [11]. It calculates every possibility related to the data within the data set and gives a result in relation to them. It's able to work with unstable and irregular data. The higher the number of training data, the higher the probability of successful results. It is suitable for use with big data. This classification algorithm is probabilistic algorithm. This algorithm has no training phase. A high rate of success can be achieved with the high number of training data [12].

**4-6- Keras**

Artificial neural network is a machine learning model of biological learning processes having a mimic systematic network structure. It has a repetitive variant based on previous estimates of these networks. Its structure mimics the working principle of the human brain [13]. Nadam optimization was used in this study as Keras. The result had at the end in this optimization is 92.6%. As Nadam is RMSprop with Adam's momentum, Nadam is also RMSprop with Nesterov momentum. The default parameters follow the parameters given on the paper. It is recommended that the parameters of this optimizer are left at their default values [14].

## V. FINDINGS AND EVALUATION

A total of 6 machine learning algorithms were applied to the data set. Considering these algorithms, the best result was obtained with the Nadam optimization used with Keras. The accuracy was found to be 92.6%. The number of training for each algorithm was determined as 100. Batch size was also found as 4. In this way, it is intended to update an equal amount of parameters and the number of sub-samples to be equal. The Nadam optimization was left as default. In other machine learning algorithms, the results were 64.7% for SVC, 73.6% for Decision Tree, 79.6% for Random Forest, and 75.9% for Gradient Boosting. The lowest result was found in Naive Bayes algorithm with 62.8%. This rate has been the highest rate of success in this field and among studies conducted with this data set.

| Accuracy on training set (Small NN) | Accuracy on validation set (Small NN) | Accuracy on training set (Always Predict Negative) | Accuracy on validation set (Always Predict Negative) | Accuracy for Neural Network |
|---|---|---|---|---|
| 97.2% | 75.6% | 62.8% | 62.2% | 92.6% |

**Table 1:** The Highest Result

## VI. CONCLUSION

Various classification techniques and their accuracy are compared in this study. Not many studies have been performed in the field of sentiment analysis for airline services. A sentence-level analysis of tweets about airline services was made with this study. Considering the machine learning algorithms used, it was seen that the highest result of Keras's Nadam optimization was taken as 92.6%. Considering the studies on the subject we have examined, we can see that 92.6% result is the highest result for Keras. The accuracy achieved by the classifiers is high enough to be used by the airline industry to investigate customer satisfaction. Since the biggest problem is the limited number of tweets used in the training of the model, it's still possible to make improvement in this analysis. A higher accuracy can be obtained by increasing the number of tweets.

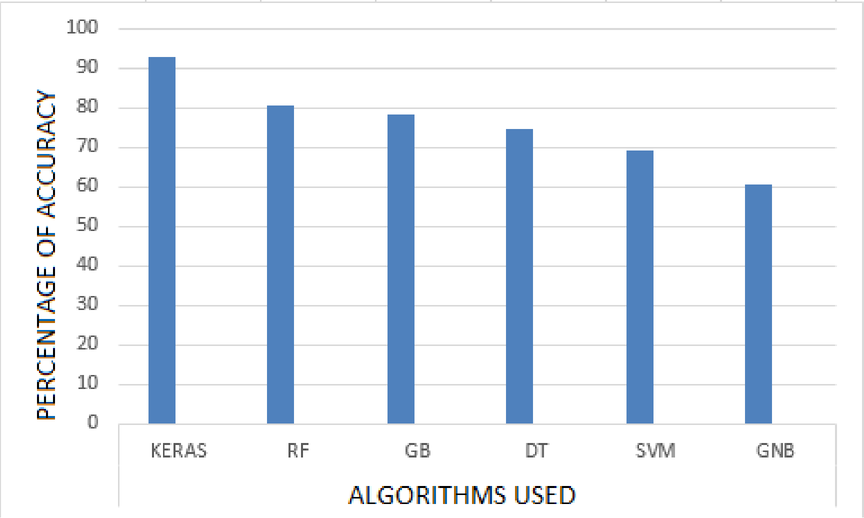

**Figure 1:** Comparison of Algorithms

## REFERENCES

[1]. E. Cambria, H. Wang, and B. White, “Guest editorial: Big social data analysis,” Knowledge-Based Systems, vol. 69, 2014, pp. 1–2.

[2]. RANE, Ankita; KUMAR, Anand. Sentiment classification system of twitter data for US airline service analysis. In: 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC). IEEE, 2018. p. 769-773.

[3]. ACOSTA, Joshua, et al. Sentiment analysis of twitter messages using word2vec. Proceedings of Student-Faculty Research Day, CSIS, Pace University, 2017, 7.

[4]. SHYAMASUNDAR, L. B.; RANI, P. Jhansi. Twitter sentiment analysis with different feature extractors and dimensionality reduction using supervised learning algorithms. In: 2016 IEEE Annual India Conference (INDICON). IEEE, 2016. p. 1-6.

[5]. Vapnik, Vladimir N., and Aleksei Yakovlevich Chervonenkis. "The uniform convergence of frequencies of the appearance of events to their probabilities." Doklady Akademii Nauk. Vol. 181. No. 4. Russian Academy of Sciences, 1968.

[6]. Edwards, Ward, and Detloff von Winterfeldt. "Decision analysis and behavioral research." Cambridge University Press 604 (1986): 6-8.

[7]. Ho, Tin Kam. "Random decision forests." Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. IEEE, 1995.

[8]. BREIMAN, Leo. Random Forests. Statistics Department. University of California, Berkeley, CA, 2001, 4720.

[9]. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.

[10]. CHEN, Tianqi, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 2015, 1-4.

[11]. McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." AAAI-98 workshop on learning for text categorization. Vol. 752. No. 1. 1998.

[12]. BAYES, Thomas. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. Philosophical transactions of the Royal Society of London, 1763, 53: 370-418.

[13]. C. C. Aggarwal, «Neural Networks and Deep Learning,» 2018

[14]. GAZEL, S. E. R.; BATİ, Cafer Tayyar. Derin Sinir Ağları ile En İyi Modelin Belirlenmesi: Mantar Verileri Üzerine Keras Uygulaması. Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi, 29.3: 406-417.