Online Social Media-based Sentiment Analysis for US Airline companies

Heba Hakh, Ibrahim Aljarah, Bashar Al-Shboul
Department of Business Information Technology
The University of Jordan
Amman, 11942, Jordan
heba.hakh@gmail.com, i.aljarah@ju.edu.jo, bashar.shboul@gmail.com

Abstract

Social media has shown powerful impact on various life choices considering that most people refer to their contacts among other people on various social media platforms for opinion, advice, or reviews. Air travel is not an exception; therefore, people expression their opinions using social media applications while we focus on analyzing a dataset of tweets, specifically, for US airline companies, and then classify these tweets according to their sentiment. We apply SMOTE method to solve the imbalanced challenge of the datasets. Furthermore, we apply different levels of feature selection to speed the sentiment analysis process. Finally, our proposed methodology is evaluated against the dataset relevance judgment, yielding promising results on all utilized evaluation metrics.

Keywords: Feature Selection; Twitter; SMOTE; Sentiment Analysis, Airline companies

1. INTRODUCTION

In the recent years, online social media websites have been growing widely, and their users are increasing rapidly. Noticeably, people have a heavy reliance on the social media where they can freely express their feedbacks and reviews. It can also be noticed, that researchers are showing an increasing amount of interest in the social media websites such as Facebook, and Twitter. These websites have huge number of customer engagements that considered very important source of information. The purpose of studying these websites is to extract these reviews and feedbacks to determine whether the customers like or dislike a specific product or service. But due to the strict privacy policy of Facebook and Instagram, it is difficult to extract the information needed. Therefore, researchers turn to Twitter to extract and analyze the customers' tweets, hashtags, and reviews related to their field of study.

Sentiment analysis is one of the popular methods that is used to automatically extract people's reactions, opinions and feedbacks towards a specific product or service. It is based on analyzing text by applying several text mining and natural language processing techniques to classify the sentiment polarity of the text as Positive, Negative or Neutral. Sentiment analysis plays an important role in many applications such as e-commerce, transportation, automotive industry, cloud platforms, marketing and promoting sales, and many others [14], [15].

Sentiment analysis can be categorized into two main classes: lexicon-based sentiment and machine learning sentiment. In the lexicon-based sentiment, each sentence is observed independently, split into tokens (words), and then analyzed using a special language dictionary called lexicon/dictionary to determine its polarity. The lexicon contains a huge set of standard words that categorized based on the polarity score. However, due to humans' tendencies to use abbreviations and slang words when expressing opinion, and because lexicons do not contain such words, the researchers have found a need to apply an alternative technique for detecting sentiment in text. Therefore, machine learning techniques were used to aid in solving this problem. On the other hand, machine learning based sentiment states the fact that the machine learns when given enough training instances to finally predict a certain outcome in the future. When using machine learning-based methods, the problem is formulated as a classification problem such as each document is represented by a set of features. After that, these documents are labeled based on the polarity (i.e. Positive, Negative, or Neutral), and finally converted to a matrix such as the matrix's rows represent the documents and the matrix's columns represent the features. Machine-learning based methods have showed significant improvements in detecting sentiment.

Nowadays, air travel is considered one of the most used transportation methods. Therefore, airline companies try to gain competitive advantage by constantly improving their services. Some of airline companies focus solely on reducing their prices and some others company focus on the quality of service and the user experience of their services. Social media is an important source for keeping track of the

users' interactions, receiving their feedbacks, and analyzing their sentiments. This can be through analyzing the contents of the social media websites to understand whether customers like or dislike a specific product or service.

In this research, we analyze a collection of tweets about six airline companies found in the United States. The analysis is conducted using traditional machine learning techniques. In addition, we test the effect of feature selection and over-sampling techniques on airline datasets that obtained from the literature. Finally, we evaluate in terms accuracy and F1 measure.

The remainder of this paper is organized as follows: related works is discussed in Section II, and description of the proposed methodology can be found in section III. Section IV presents the experiments and results. Finally, we present our conclusion and future directions in Section V.

2. RELATED WORKS

Many researchers have shown interest in studying the overall process of sentiment analysis by detecting emotions found in text [2, 3]. Others have proposed evaluating sentiments by observing human behavior responding to a certain experience [4].

Likewise, some authors have made an overview about the use of machine learning techniques for sentiment classification, in which they have illustrated three classifiers; (Naïve Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM)) [7]. The authors in [9] applied the previously mentioned classifiers (NB, ME, and SVM) on the Internet Movie Database (IMDB) consisting of movie reviews either expressed with stars or as numerical values. Their approach was evaluated using accuracy and recall measures. The accuracy results ranged from 77% to 81% whereas, the recall ranged from 50% to 69%.

In [11], authors tried to investigate the factors driving the travelers' loyalty towards an airline company in China. The authors have determined travelers' loyalty based on ten airline attributes namely; operational factors (punctuality, aircraft, and safety), attractive factors (food and beverages, and the staff service), competitive factors (schedule, tickets prices, reputation, flyer program). The authors have concluded that the factors driving the travelers' loyalty towards a specific airline company are reputation, staff service, frequent flyer program, aircraft, and punctuality.

The authors in [8] used statistical methods (Mean, Standard Deviation, and Chi-Square) to determine the customers' engagement on social networking sites on the airline companies in India. Moreover, they have studied the post interactions level as well as the customers' involvement in supporting the functions of the airline company. The authors finally proposed that the airline companies should add more content on other social media websites in order to compete with foreign airline companies.

In [6], the authors rated an airline company according to "Rating Features" where users were asked to rate the airport, airline, lounge, and their assigned seats on a scale from 1-10. To that end, they combined Rating Features with Textual Features to measure the passengers' satisfaction. Whereas in [1], [12], the authors gathered data from the U.S. Department of Transportation's monthly Air Travel Consumer Report to measure what they called "Airline Quality Rating (AQR)", to assess the airline's overall quality based on a given set of weights and their impacts. In the case of sentiment classification, the authors used Naïve Bayes.

3. PROPOSED APPROACH

In this work, the dataset contains various tweets on different airline companies at the US is used. The "Twitter Airline Sentiment" dataset was obtained from Kaggle contains tweets covering six U.S. airline companies with a total number of (14,640) tweets, each of which is labelled according to sentiment polarity as: positive, negative, and neutral. The dataset was first split into six smaller datasets each of which corresponds to tweets mentioning a certain company. Summary of the datasets are shown in Table 1.

Table 1: Summary of datasets

Dataset (Airline)	***Number of Tweets***	***Percentage of***		
		Positive	*Negative*	*Neutral*
Virgin America	504	30%	35%	33%
US Airways	2913	13%	77%	10%
United	2434	12%	68%	18%
Southwest	4841	23%	48%	27%
Delta	2222	24%	43%	32%
American	2760	12%	71%	16%

All datasets share seven features in addition to the class label. The original features are described in Table 2. Furthermore, additional features were added, such as tweet word counts in addition to the features (i.e. terms) resulting after the tokenization process.

All tweets will cleaned by removing stopwords, lowercasing all terms, and then applying word

stemming. After that, terms were tokenized then Term frequency – Inverse Document Frequency (TF-IDF) was used to measure importance weight of terms. Further, a tweet-term weight matrix is generated where the terms represent the features, and weights are the TF-IDF scores calculated earlier. Table 2 shows the number of features/terms for each dataset.

Table 2: Airline features list

Original Feature	Description
Airline Sentiment Confidence	A numeric feature representing the confidence level of classifying the tweet to one of the (3) classes.
Negative Reason	The reason behind considering this tweet as negative (i.e. bad flight). Positive and neutral tweets had no negative reasons.
Negative Reason Confidence	The level of confidence in determining the negative reason behind a negative tweet.
Airline	The name of the airline Company.
Retweet Count	The number of retweets of a tweet.
Text	The original tweet posted by the user.
Airline Sentiment	A feature containing the class labels for tweets (positive, negative, neutral).

Table 3: Summary of dataset features

Dataset (Airline)	Number of Features/Terms
Virgin America	1487
US Airways	3714
United	4727
Southwest	3685
Delta	3929
American	3518

Tokenization caused generating a large number of terms, and therefore term/feature selection is used in two levels to select small number of terms representing the best subset of features. At the first level of feature selection, we used genetic search [17] as a filter to select the best subset of features based on the features' correlation with the class label. After that, and the second level, a univariate feature selection was applied which is aiming at selecting the best subset of features based on statistical tests. The best K features returned are selected for further experiments. K is selected empirically as a percentage of the number of features resulting from the first level of feature selection, after experimenting different percentages ranging between 10% to 70%. The percentage that effectively boosted our evaluation metrics was selected and then reported in Table 4.

It can be noticed that sentiment polarity of all the datasets, except for Virgin America, are not evenly distributed; therefore, the training process of any classifier considering the imbalanced datasets will result in a bias due to the larger number of instances, sampled for training, and belonging to one class.

Table 4: Summary of dataset features after feature selections

Dataset (Airline)	Before Feature Selection	Level 1 Genetic Search		Level 2 Univariate Selection	
		Number of features	*Percentage of dataset*	*Number of features*	*Percentage of Level 1 features*
Virgin America	1487	266	18%	187	70 %
US Airways	3714	540	15%	270	50 %
United	4727	238	5%	119	50 %
Southwest	3685	1140	31%	798	50 %
Delta	3929	1224	31%	612	50 %
American	3518	321	9%	192	60 %

To solve the imbalance problem, we used Synthetic Minority Over-Sampling Technique (SMOTE) [13] to increase the number of instances used for the training process from the minority class. SMOTE method is considered one of the popular methods used to increase the number of instances sampled for training from the minority classes, which is adopted to decrease the skewed class distribution. After selecting samples from the minority class, the nearest minority neighbors are identified to generate more samples in between the samples chosen and the nearest minority neighbors. Table 5 illustrates the number of instances before and after SMOTE was applied. It is important to notice that SMOTE was not applied on the Virgin America dataset as it is almost evenly distributed.

Finally, on each of the partial datasets various classification techniques were applied (i.e. AdaBoost, Decision Tree, Linear SVM, Naïve Bayes, Random Forest, K-NN, and Kernel SVM) which is aiming at predicting the class of each tweet provided the subset of features obtained after the two levels of the feature selections and data balancing.

Table 5: Datasets after applying the SMOTE method

Dataset (Airline)	Number of Instances		Percentage		
	before SMOTE	After SMOTE	Positive	Negative	Neutral
US Airways	2913	6787	33%	34%	33%
United	2434	5176	33%	34%	33%
Southwest	4841	7114	33%	34%	33%
Delta	2222	2865	33%	34%	33%
American	2760	5880	33%	34%	33%

4. EXPERIMENT SETUP & RESULTS

As datasets now seem more balanced, classifiers were applied to evaluate the sentiment classification (i.e. AdaBoost, Decision Tree, Linear SVM, Naïve Bayes, Random Forest, K-NN, and Kernel SVM). Python language are WEKA [16] tool are used to implement all the used methods.

Classification settings per algorithm were set empirically after performing experiments with different settings. For example, number of trees in the random forest classifier was set to 4 as it showed the best results. On the other hand, the kernel chosen for the Kernel SVM was the Radial Basis Function (RBF). Additionally, for the kNN classifier, the value of number of neighbors (k) are selected between 1, 3, and 5.

To compare classification effectiveness: Accuracy and F1 measures were used. Classification results of each dataset is shown in Figures 1 through 6, where the dataset name is shown below each classification results group, while accuracy and F1 results are shown in different figures, separated for each feature selection level. In figures 1 and 2, results reported for classifying the datasets considering all features, no feature selection of any level is applied. Further, in figures 3 and 4, results are reported for classification after feature selection of level 1. Finally, in figures 5 and 6, classification results reported after level 2 feature selection.

A general conclusion is that classifiers' results vary according to the dataset, however, it is noticeable that kNN, RF, and DT were almost better than other classifiers on the majority of the experiments. Another major conclusion is that feature selection (levels 1 and 2) almost enhanced every performed experiment on any dataset, except for rare cases where using the whole feature set was better (i.e. Naïve Bayes, and Adaboost on the "American" dataset). Another observation that can be made is that no significant enhancement in classification accuracy can be observed comparing level-1 term selection to level-2.

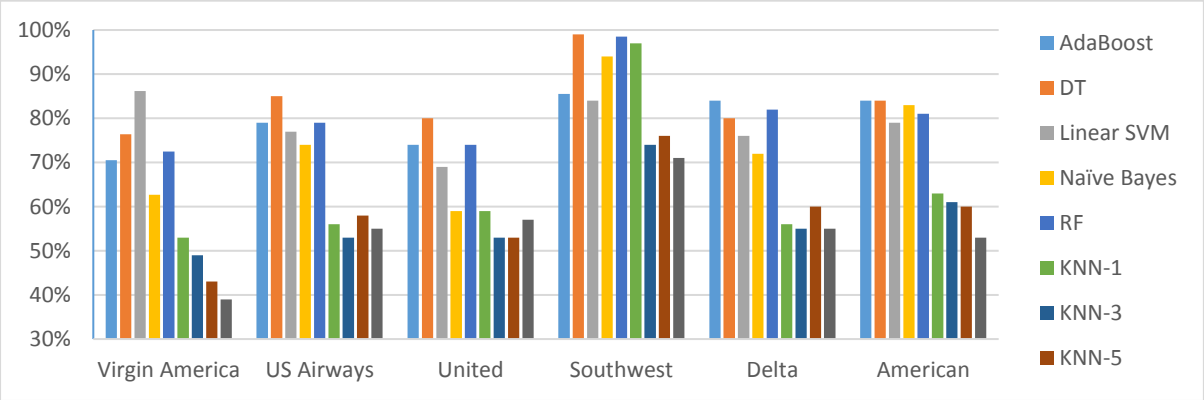

Fig. 1: Accuracy results on different datasets before level-1 feature selection using various classifiers

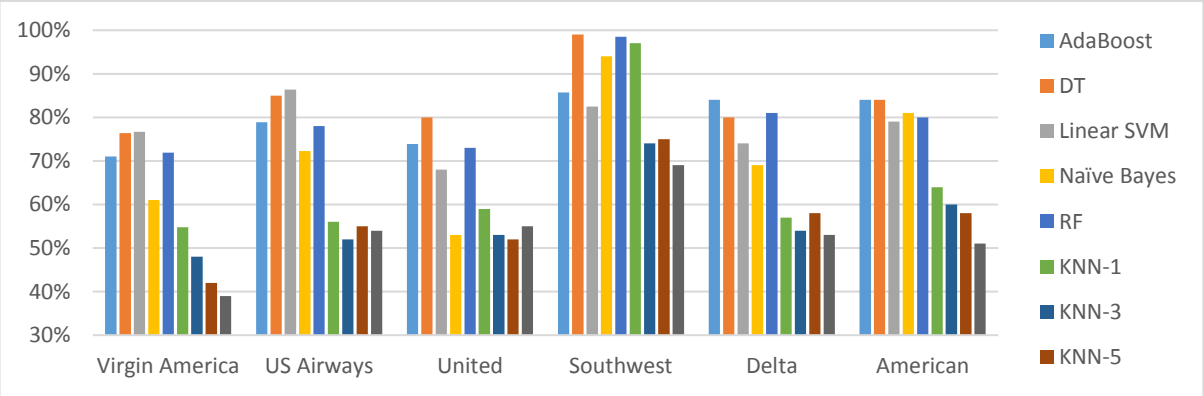

Fig. 2: F1 Measure results on different datasets before level-1 feature selection using various classifiers

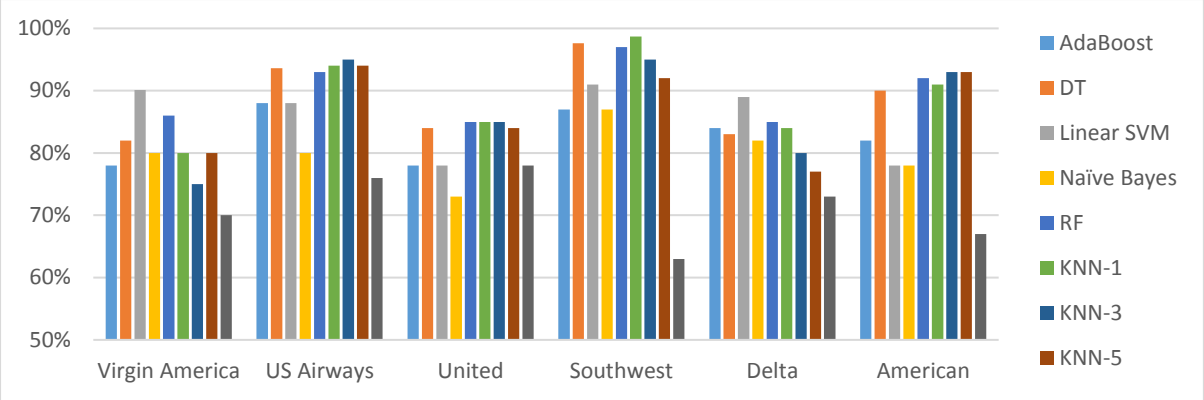

Fig. 3: Accuracy results on different datasets after level-1 feature selection using various classifiers

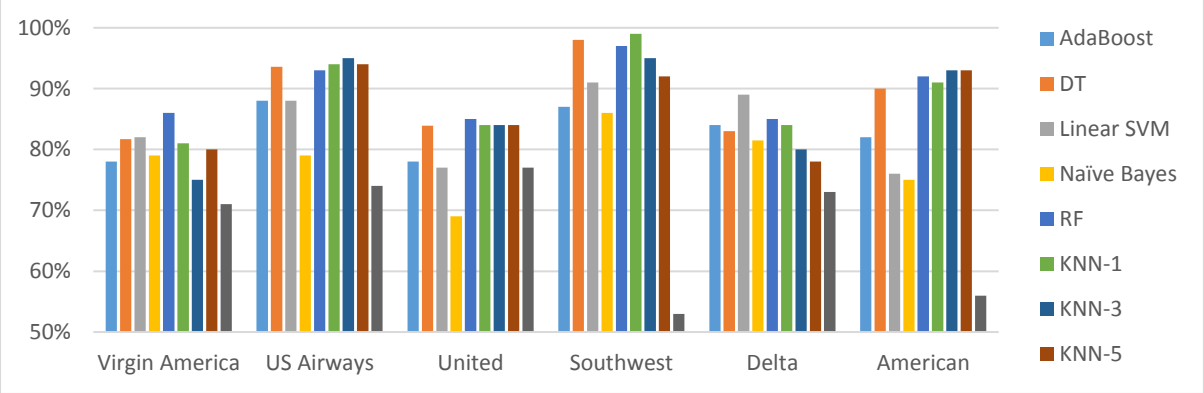

Fig. 4: F1 Measure results on different datasets after level-1 feature selection using various classifiers

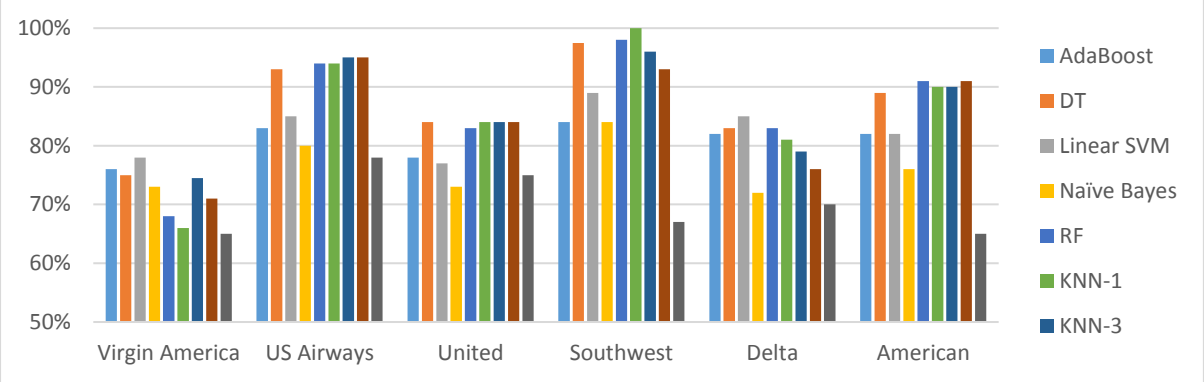

Fig. 5: Accuracy results on different datasets after level-2 feature selection using various classifiers

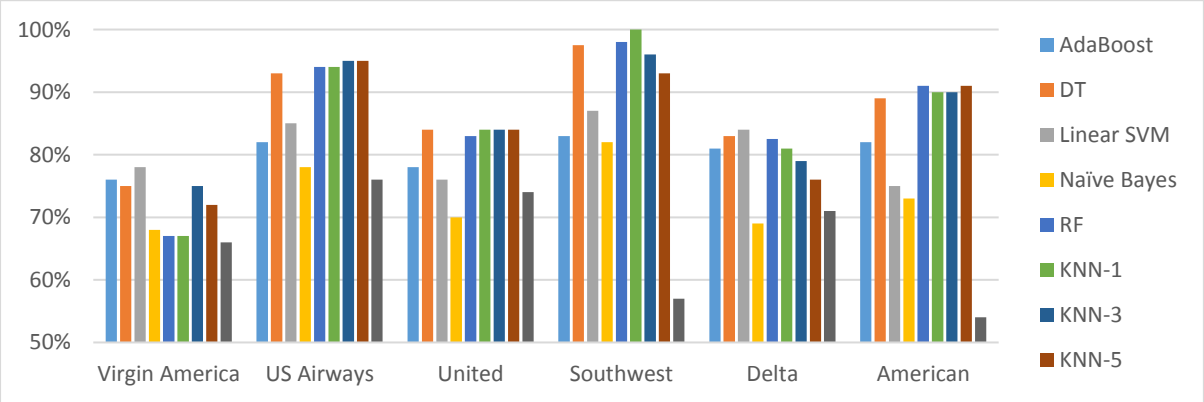

Fig. 6: F1 Measure results on different datasets after level-2 feature selection using various classifiers

5. CONCLUSIONS

In this paper we have studied the sentiment analysis based on the feedbacks of travelers regarding airline companies. Our proposed approach showed that both feature selection and over-sampling techniques are equally important as regards to boosting our results. The use of feature selection techniques has returned the best subset of features and reduced the computations needed to train our classifiers. Whereas, SMOTE has reduced the skewed distribution of the classes found in most of our smaller datasets without causing overfitting. Our results are a compelling evidence that the proposed model has high classification accuracy in predicting instances form the three classes (Positive, Negative, and Neutral).

As can be seen, some of the applied classifiers have outperformed the others. For example, Random Forest and Decision Tree have shown a high prediction level, and stability when applied on all datasets. While K-NN and Linear SVM have shown an acceptable level of performance regarding all the evaluation metrics. On the other hand, Kernel SVM has shown poor results in comparison with other classifiers

6. REFERENCES

[1] Esi Adeborna, Keng Siau, "An Approach To Sentiment Analysis-The Case of Airline Quality Rating," PACIS 2014 Proceedings. Paper 363.

[2] Hung T.Vo, Hai C.Lam, Duc Dung Nguyen, Nguyen Huynh Tuong,"Topic Classification and Sentiment Analysis For Vietnamese Education Survyey System",Asian Journal of Computer Science and Information Technology 6:3, May (2016).

[3] Soumi Sarkar, Taniya Seal, "Sentiment Analysis- An Objective View", Jornal of Research, Volume 02, Issue 02, April 2016.

[4] Ann Devitt, Khurshid Ahmad, "Sentiment Analyis And The Use of Extrinsic Datasets in Evaluation," International Conference on Language Resources and Evaluation,2008.

[5] Show-Jane abd Yue-Shi Lee, "Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset," Springer- Verlag Berlin Heidelberg ,pp.731-740 ,2006.

[6] Emanuel Lacic, Dominik Kowald, and Elisabth Lex, "High Enough? Explaining And Prediciting Traveler Satisfaction Using Airline Reviews", ACM-2016.

[7] Jayashri Khairnar, Mayura Kinikar, "Machine Learning Algorithms for Opinion Mining and Sentiment Classification", "International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013.

[8] Joel Gnanprakash, Prasad Kulkarni, " Social CRM in the Airline Industry: A Case Study of Indian Airline Companies", "Journal of Marketing Management and Cosnumer Behavior, Vol.1, Issue 1 (2016) 76-87".

[9] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing – Volume 10, pages 79-86".

[10] David Mc. A Baker, " Service Quality and Customer Satisifaction in the Airline Industry: A Comparison between Legacy Airlines and Low-Cost Airlines", "American Journal of Tourism Reseearch Vo1 2. No. 1,2013, 67-77".

[11] Ilias Vlachos, Zhibin Lin, "Drivers of airline loyalty :Evidence from the business travelers in Chine", ELSEVIER, Transportation Research Part E: Logistics and Transportation Review , Volume 71, November 2014, Pages 1–17".

[12] Bowen, B.D. and Headley, D.E., (2013, April).Airline QualityRating 2012.(22nd ed.)[Online].Avaiable: http://www.airlinequalityrating.com/reports/2012aqr.pdf

[13] Natesh Chawla, Kevin Bowyer, Lawrence Hall, W. Phillip Kegelmeyer, SMOTE: Synthetic Minority Over-sampling, Technique, Journal of Artificial Intelligence Research, vol. 16, pp.321-357, 2002

[14] Shukri, Sarah E., Rawan I. Yaghi, Ibrahim Aljarah, and Hamad Alsawalqah. "Twitter sentiment analysis: A case study in the automotive industry." In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on, pp. 1-5. IEEE, 2015.

[15] Qaisi, Laila M., and Ibrahim Aljarah. "A twitter sentiment analysis for cloud providers: a case study of Azure vs. AWS." In Computer Science and Information Technology (CSIT), 2016 7th International Conference on, pp. 1-6. IEEE, 2016.

[16] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18.

[17] Hamilton, New Zealand. "Correlation-based feature subset selection for machine learning." New Zealand (1998