IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

Sentiment Analysis on Airline Service Reviews using Data Mining based Classification Techniques

Rajat Yadu[1] and Ragini Shukla[2]
Research Scholar, CSIT, Dr. C. V. Raman University, Bilaspur, India[1]
Professor, CSIT, Dr. C. V. Raman University, Bilaspur, India[2]

Abstract: *Sentimental analysis is the field where online reviews, opinions, and sentiments from users are available and provide considerable amounts of information about the services, facilities, and status of the service provider in the market. In addition to examining the classification accuracy of standard data mining methods, this research evaluates the sentiments expressed about six social media microblog traveller networking site datasets relevant to Indian airlines. Using standard data mining classifiers, Bayes Net and SVM performed with high accuracy rates. In this paper main analysis of the classification performance of passenger sentiments for six airline services has been performed. However, we found that the Bayes Net performed best accuracy rate using WEKA tool but in case of using Rapid Miner tool SVM has produced the highest accuracy rate for our research among the other common standard classifiers. On the basis of thoroughly favourable service reviews from passengers, Go Air is consistently the airline that is most highly recommended.*

Keywords: Sentiment Analysis, Standard Classifier, Social Media Micro Blog Datasets, Classification, SMO

I. INTRODUCTION

Today's social media encompasses a variety of online venues, and one of the greatest sources for gathering data for air travel is the microblog traveller website. In our study, we used four distinct social media microblog travel websites to gather customer service ratings of six airlines. The analysis of these customer reviews will aid businesses or other passengers in improving the calibre or services of airlines. Regarding the aviation sector Sentimental analysis is the newest method for determining how accurately customers' service ratings are supplied; it is a quicker and more affordable technique to ascertain how customers feel about the services offered and how microblogging has affected the aviation industry. Sentiment analysis is a discipline devoted to the extraction of irrational feelings and emotions from text. In this study, we use sentimental analysis to a collection of airline service review data. This work improves the passenger services provided by airlines. We have worked with sentence-level service reviews and defined the polarity label as positive or negative based on a few key terms like good, terrible, best, worst, comfortable, poor, and punctual.

Sentiment analysis commonly referred to as opinion mining, uses technologies for natural language processing to analyse the attitudes, emotions, and opinions hidden behind the words. Although the neutral class is usually disregarded in statistical categorization methods. Sentiment analysis became more popular as social media platforms like blogs and social networks grew. Online opinion has evolved into a kind of virtual currency for firms wanting to advertise their products, find new possibilities, and manage their reputations as a result of the explosion of reviews, ratings, recommendations, and other forms of online expression.

Utilization of online social networks is widespread today. Predictive analytics will benefit from mining the text found in online social networks. A difficult research topic is predicting information from unstructured data seen in social networks. Sentiment analysis, often known as text analysis, is the process of extracting, detecting, or otherwise characterising the sentiment content of the text unit using statistics and machine learning techniques. Weka tool will be used to illustrate sentiment analysis utilising Decision trees and Support vector machines, which are machine learning techniques. (V, 2014). Reviews are cleaned and pre-processed before being sent to WEKA for additional processing in a small number of literatures. Research is done on various sized feature sets for different standard machine learning with reviews represented as a feature set (Rani et al., 2021).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

Using data mining techniques in the WEKA tool and Rapid Miner, which does the actual categorization of feelings in the form of classifier accuracy rate and offers the capability to propose an airline to passengers, we have explored the sentimental analysis of airline review datasets in this paper. Computational linguistics and natural language processing techniques are employed in text sentiment analysis to classify the polarity of text at the text and sentence levels. There are normally two basic approaches for text sentiment analysis, and we have used the machine learning approach, which entails creating feature vectors as word count.

II. MATERIAL AND METHODS

The dataset of social media microblog airline service reviews for six different airlines is discussed in this section, along with the pre-processing of the data using the StringToWordVector feature extraction techniques to convert the string into vectors and the TF-IDF technique to remove delimiters notations in Weka and in case of Rapid Miner Word2Vec with N gram technique has been used as feature extractor and made vector as word count for the occurrence of same reviews. Following this procedure, we discuss about feature selection strategies, where we frequently take into account the three primary features—sentiments, text, and score—when using classification algorithms. The suggested method employs data mining techniques to analyse sentiment using 6 distinct airline service reviews. In total, we have collected 42000 reviews; the numbers of reviews are as follows: 20000, 1304, 5261, 3550, 1665, and 10952. The sentiment classification is carried out using two data mining tools.

The main challenge in sentimental analysis is data collecting, which is challenging because of privacy issues such the worry about disclosing personal information. Real-time passenger reviews from airline service providers were gathered via social media. The machine learning team uses this dataset to carry out empirical research on machine learning techniques. We gathered reviews from many air travel websites, including Tripadvisor, Skytrax, mouthshut, Makemytrip, and trustpiolet, using the Web Harvey tool data extraction tool. By identifying single abbreviations and lemmatization, as well as correcting and halting word removal, these data sets illustrated the value of pre-processing data. With the exception of blank spaces, the text has been tokenized by breaking it into a series of letters or words. Stop word elimination Delete any common words that convey a non-discriminatory sentiment. Steaming using a term just as grammatical tool Special characters, such as punctuation marks, is removed that are not necessary.

Data preparation allows us to turn basic information into a form that is useful. A variety of content that is unequally dispersed among its users is included in the final text-based information made available by websites. We remove many stop words from our dataset that appear in two different classifications as part of the pre-processing of our data. This method begins with pre-processing steps used to clean the reviews and balance the review data. It then employs a few feature extraction techniques, such as StringToWordVector and Word2Vec with N gram technique, which converts string attributes into a set of numeric attributes representing word occurrence information from the text contained in the strings. Skytrax is primarily concerned with airport rankings and the presentation of World Airports Awards, as opposed to TripAdvisor and Google Maps, which concentrate on customer information sharing. Like ASQ awards, Skytrax awards are commonly used by winning airports as marketing tools. Skytrax started compiling online opinions in July 2002, but the rate of increase in airport opinions has been sluggish (Lee & Yu, 2018). This dataset has undergone pre-processing, a data mining method (Kaur & Malik, 2021). We used a variety of feature selection techniques after pre-processing. The selection of features is one of the most crucial responsibilities for text classification problems (FS). In our research the airline dataset has contained few features for analysis. Choosing the best attributes to represent reviews is the aim of the text feature selection process. Through the use of this technology, data mining algorithms' performance can be improved and dataset sizes can be reduced. In order to improve run time and accuracy, we employ feature selection algorithms to remove unsuitable features from statistical reviews. When using our technique, keywords and words that feature in reviews with more insightful votes are given more weight (Joyce & Deng, 2019). In our study, we used six attributes as the initial set of data, including Name, Text, Sentiment, Description, Date, and Score, but after applying a variety of feature selection techniques, such as Latent Semantic Analysis, CfsSubsetEval, ClassifierAttributeEval, CorrelationAttributeEval, GainRatioAttributeEval, and InfoGainAttributeEval with different search algorithms, such as greedy search , CIsearch Algorithm , Hill climbing with search methods as Best Search , Greedy Search and Ranker in Weka and IG , GR have used in Rapid Miner then we have get two better features to get better accuracy for classification such as sentiment and score.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

We describe the standard classification techniques utilized in this paper towards the end of this section. In the experimental process we have used the common standard classifiers with its performance accuracy rate between both the tools. We go over each technique and describe how it was used to our experiment. We used the following data mining classification algorithm approaches for six social media microblog airline service reviews datasets under the various categories:

2.1 Bayes Classifier

We introduce the naive Bayes algorithm and use it for text categorization, the process of giving an entire text or document a label or text categorization category.

- BayesNet - A graphical representation of the joint probability between groups of random variables serves as a Bayesian network. Bayesian classifiers are those used in statistics. They have the ability to predict probabilities of class membership, such as the likelihood that a given tuple belongs to a particular class. The foundation of Bayesian categorization is the Bayes theorem. The Naive Bayesian classifier, a straightforward Bayesian classifier, has been found by classification algorithms to perform comparably to decision trees and certain neural network classifiers.
- Naive Bayes- Popular text classification algorithms include Naive Bayes. Sentiment classification is the first step in the Naive Bayes classifier learning process. This approach of text categorization uses generative models, which are well-established. In this project, we use the TF-IDF technique to extract the features from textual data before applying the Naive Bayes classifier. The NB classifiers' odd inability to operate on unbalanced classes is one of their distinctive limitations. Because they assume conditional independence among language features, the primary theoretical drawback of NB approaches is this (El-Rashidy et al., 2014).

2.2 Lazy Classifier

Machine learning techniques known as lazy learning postpone generalisation of the training data until the system is queried. Instance-based Learning is another name for this kind of learning. When working with enormous datasets that only have a few attributes, lazy classifiers are extremely helpful.

- IBK- Machine learning techniques known as lazy learning postpone generalisation of the training data until the system is queried. Instance-based Learning is another name for this kind of learning. When working with enormous datasets that only have a few attributes, lazy classifiers are extremely helpful.
- Kstar - K-Star, also referred to as the K * algorithm, is a K Nearest Neighbour (KNN)-based instance-based classifier. Its objective is to form k clusters out of n data points. The entropic distance metric used by K* is based on the probability that one occurrence will lead to another. Information theory helps determine the distance between instances, and the application of entropy is essential for a distant instance. Entropic distance is used to locate the data set's most comparable examples as a result.
- LWL- The leaves of a logistic model tree are where logistic regression functions are employed. This method can deal with binary and multi-class variables, missing values, as well as numerical and nominal features. It produces tiny, accurate trees. It uses the CART pruning technique. There are no tuning parameters required.

2.3 Functions Classifiers

- SVM/SMO (Sequential Minimal Optimization): Support vector machines (SVMs) are supervised learning methods that generate input-output mapping functions from a set of labelled training data. The mapping function can be either a classification function (used to categorize the input data) or a regression function (used to estimation of the desired output). For classification, nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. Then, the maximum-margin hyper planes are constructed to optimally separate the classes in the training data. SVM is the best algorithm for binary classification (Kaur & Malik, 2021).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

2.4 Tree Classifiers

- Decision Stump- A single level decision tree is used. There is only one internal node, and it is linked directly to the terminating nodes. A single root node decides how to classify inputs based on a single feature. A potential feature value is shown in each leaf, along with the class label that should be assigned to inputs with that value. Choose the feature and build the tree to implement this technique.
- LMT- The leaves of a logistic model tree are where logistic regression functions are employed. This method can deal with binary and multi-class variables, missing values, as well as numerical and nominal features. It produces tiny, accurate trees. It uses the CART pruning technique.
- Random Forest- As an ensemble learning technique for classification, regression, and other tasks, random forests, often referred to as random decision forests, create a huge number of decision trees during training and then forecast the class that is the mean prediction of all the individual trees (Panda, 2019).
- Random Tree - Leo Breiman and Adele Cutler designed random forests, which may be accessed at http://www.stat.berkeley.edu/users/breiman/RandomForests/. The approach can deal with both classification and regression issues.

The software tools we have used in this research are WEKA and Rapid Miner where the first tool is WEKA (Waikato Environment for Knowledge Analysis) which is a well-known machine learning software suite also used for sentimental analysis. The WEKA workbench includes a set of visualization tools and algorithms for data analysis and predictive modelling, as well as graphical user interfaces for quick access to these features (Karrar & Mutasim, 2016) and the second tool is Rapid Miner which is similarly implemented in Java, provides powerful analytics through template-based frameworks that require users to write very little code. Rapid Miner is a data mining and machine learning environment.

III. RESULT AND DISCUSSION

In the dataset, 66% of the sentiments which have been analysed used for training and the rest for testing in Weka tool and In case of Rapid Miner the dataset 80% of the sentiments which have been analysed used for training and the rest of for testing. The summary of results and classification accuracy obtained using standard classifiers are tabulated in Table I implemented in Weka tool, where Bayes Net has produced the highest accuracy rate for maximum all airline datasets, but Go Air airline has achieved the maximum accuracy rate, allowing us to recommend Go Air as the best airline in terms of services and sentiment classification.

Table I: Performance of Standard classifiers in WEKA tool

Standard Classifiers in WEKA tool	Accuracy Rate					
	Air India	Air Asia	Go Air	Spice jet	Vistara	Indigo
BayeNet	99.89%	99.91%	**99.94%**	99.85%	99.85%	99.82%
NaiveBayes	97.76%	97.83%	97.86%	97.40%	98.73%	99.80%
SMO (SVM)	99.84	97.91%	99.48%	97.52%	97.35%	99.84%
(KNN) IBK	99.88%	99.81%	99.93%	99.78%	99.82%	99.80%
(KNN) Kstar	97.76%	97.51%	97.86%	99.84%	99.85%	99.81%
(KNN) LWL	99.88%	99.90%	99.93%	99.84%	99.82%	99.81%
Decision Stump	99.88%	99.90%	99.93%	99.83%	99.83%	99.81%
Random Forest	99.76%	99.76%	99.93%	99.80%	99.82%	95.66%
Random Tree	99.91%	99.90%	95.66%	99.83%	99.81%	99.81%

The Table II has shown the accuracy rate of same standard classifiers which have use in first data mining tool but the name of few classifiers have different such as SMO known as SVM in Rapid Miner tool. This table also described the Go Air airline has achieved the maximum accuracy rate by the SVM classifier but the overall accuracy rate has produced by the Bayes Net in Weka tool and this case also allowing us to recommend Go Air as the best airline in terms of services and sentiment classification.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

Table II: Performance of Standard classifiers in Rapid Miner tool

Standard Classifiers in Rapid Miner tool	Accuracy Rate					
	Air India	Air Asia	Go Air	Spice jet	Vistara	Indigo
BayeNet	76.59	71.88%	78.70%	62.48%	81.63%	77.55%
NaiveBayes	76.59	71.88%	76.50%	62.48%	81.63%	77.55%
SVM	92.24%	81.12%	**96.70%**	82.96%	81.28%	93.12%
KNN	87.92	91.67%	96.30%	72.90%	77.91%	92.67%
Decision Stump	78.61	78.61%	88.90%	72.79%	66.02%	78.61%
Random Forest	65.50	70.91%	78.04%	63.27%	64.35%	66.91%
Random Tree	87.94	91.99%	78.31%	83.72%	64.85%	87.95%
LR	93.32%	81.96%	96.27%	81.70%	80.08%	77.91%

The below figure I shown the graphical representation of accuracy rate for the all the six airlines sing Weka data mining tool. Out of all the classifiers the Bayes Net method gave the highest accuracy rata for Go Air airline among the six airlines.

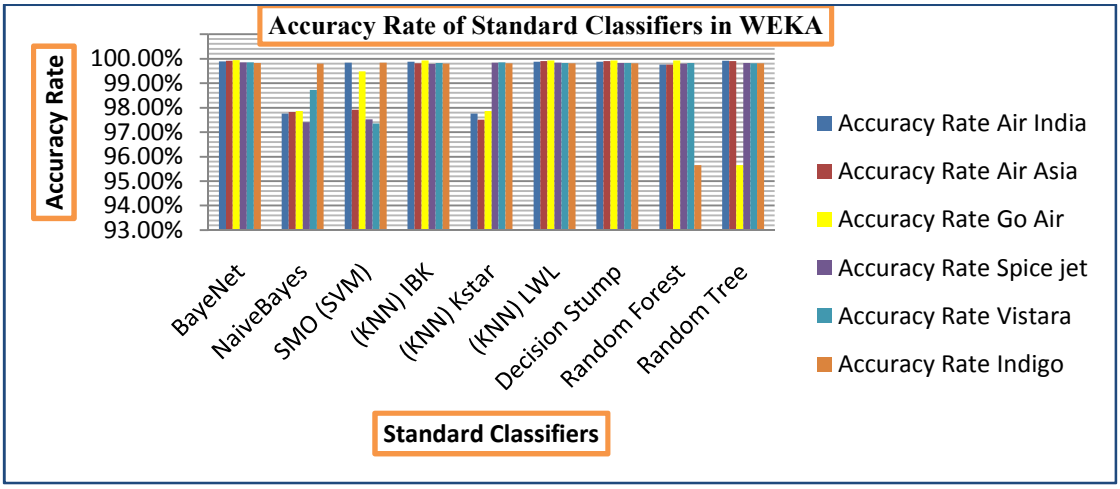

Figure 1: Classification accuracy of various classification techniques in WEKA

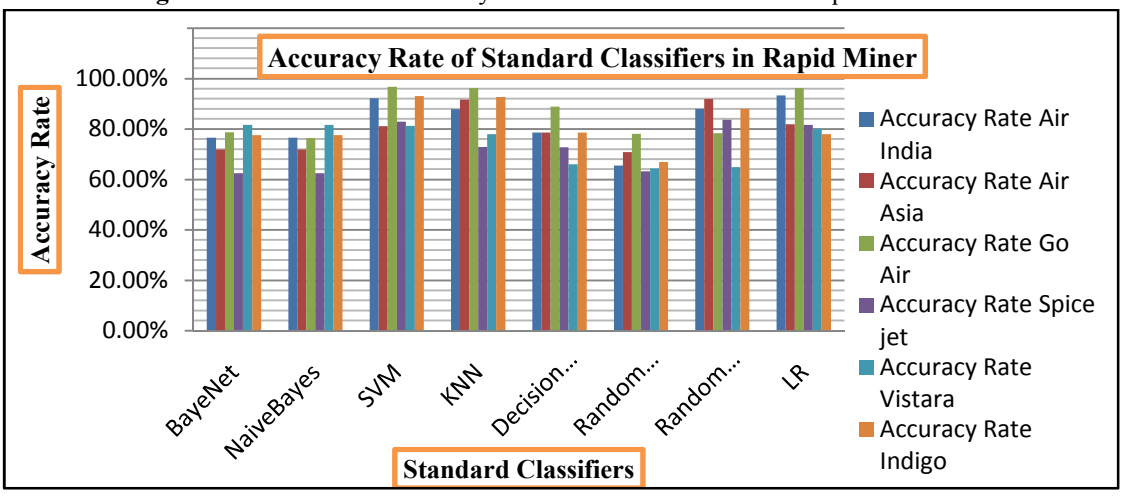

Figure 2: Classification accuracy of various classification techniques in Rapid Miner

In the above figure II the graphical representation of performance of standard classifiers of Rapid Miner tool has shown where all the classifiers performed better but overall the SVM classifier performed the highest accuracy for GoAir airline so we can recommend this airline for passengers with respect of our research. It has been noted that there is little variation in the accuracy rate for both positive and negative sentiment classes. However, the outcomes can be utilized to group a lot of data for future investigations.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

IV. CONCLUSION

The base of our research is mainly to analyse the sentiments of social media microblog airline review datasets of six airlines using data mining based standard classification techniques that how much the reviews have accurately positive or negative. For this purpose we have used two different data mining tools WEKA and Rapid Miner with most common classifiers in which KNN have the three forms as IBK , Kstar , LWL in case of WEKA and SMO (sequencing Minimal Optimizing) classifier known as SVM in Rapid Miner. In order to comprehend how airline services deliver services to clients and how we can anticipate the airline for the best travelling medium, we looked at a variety of social media transportation platforms in this study. We made an effort to analyse this goal using online resources and user reviews. In our study, we worked with three different attributes features (Text, Sentiment, and Score) that have used for training, and we used the StringtoWordVector and Word2Vec feature extraction method with tuned parameters and Word Tokenizer tokenization methods to generate filter vectors as numeric word occurrence count and various FS techniques specially Latent Semantic Analysis for feature selection in WEKA and IG in Rapid Miner. In this research we have found Baye Net classifier using WEKA and SVM classifier in Rapid Miner produced as the best accuracy rate with 99.94% , 96.70% for Go Air airline and we can have recommend this airline as best in services from our research experiments. In case of performance data mining tools WEKA has produced the slightly better as compare to Rapid Miner.

REFERENCES

[1]. Ariyawansa, C. M., & Aponso, A. C. (2016). Review on state of art data mining and machine learning techniques for intelligent Airport systems. Proceedings of 2016 International Conference on Information Management, ICIM 2016, 134–138. https://doi.org/10.1109/INFOMAN.2016.7477547.

[2]. Atubonengi, R., Anireh, V. I. ., & Matthias, D. (2021). Airline Reservation Using Sentiment Analysis with Naïve Bayes Classifier. Ijarcce, 10(8), 14–21. https://doi.org/10.17148/ijarcce.2021.10802.

[3]. Amazal, H., & Kissi, M. (2021). A New Big Data Feature Selection Approach for Text Classification. Scientific Programming, 2021. https://doi.org/10.1155/2021/6645345.

[4]. Bae, W., & Chi, J. (2022). Content Analysis of Passengers' Perceptions of Airport Service Quality: The Case of Honolulu International Airport. Journal of Risk and Financial Management, 15(1). https://doi.org/10.3390/jrfm15010005.

[5]. Baharum, Z. (2020). An Integrated Model on Airport Terminal Level of Satisfaction for Service Quality Evaluation: A Proposal. International Journal of Advanced Trends in Computer Science and Engineering, 9(1.3), 247–250. https://doi.org/10.30534/ijatcse/2020/3791.32020.

[6]. Chaudhari, B., & Parikh, M. (2012). A Comparative Study of Clustering Algorithms using Weka Tools. International Journal of Application or Innovation in Engineering and Management (IJAIEM), 1(2), 154–158.

[7]. El-Rashidy, N., Arafat, H., Elawady, R. M., Barakat, S., & Elrashidy, N. M. (2014). Different Feature Selection for Sentiment Classification Different Feature Selection for Sentiment Classification Different Feature Selection for Sentiment Classification. International Journal of Information Science and Intelligent System, 3(1), 137–150. https://www.researchgate.net/publication/328581496.

[8]. Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1). https://doi.org/10.1186/s40537-015-0015-2.

[9]. Ibarguren, I., Pérez, J. M., Muguerza, J., Gurrutxaga, I., & Arbelaitz, O. (2018). UnPART: PART without the 'partial' condition of it. Information Sciences, 465, 505–522. https://doi.org/10.1016/j.ins.2018.07.022.

[10]. Joyce, B., & Deng, J. (2019). Sentiment analysis using naive bayes approach with weighted reviews - A case study. 2019 IEEE Global Communications Conference, GLOBECOM 2019 - Proceedings. https://doi.org/10.1109/GLOBECOM38437.2019.9013588.

[11]. Karrar, A. E., & Mutasim, M. (2016). Comparing EM Clustering Algorithm with Density Based Clustering Algorithm Using WEKA Tool. International Journal of Science and Research (IJSR), 5(7), 1199–1201. https://doi.org/10.21275/v5i7.art2016420.

[12]. Kaur, G., & Malik, K. (2021). A Sentiment Analysis of Airline System using Machine Learning Algorithms. International Journal of Advanced Research in Engineering, 12(1), 731–742.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, January 2023
Impact Factor: 6.252

[13]. Lee, K., & Yu, C. (2018). Assessment of airport service quality: A complementary approach to measure perceived service quality based on Google reviews. Journal of Air Transport Management, 71(May), 28–44. https://doi.org/10.1016/j.jairtraman.2018.05.004.

[14]. Martínez, D. M., Ebenhack, B. W., & Wagner, T. P. (2019). Transportation sector energy efficiency. Energy Efficiency, 197–226. https://doi.org/10.1016/b978-0-12-812111-5.00007-x.

[15]. Panda, P. K. (2019). A LITERATURE REVIEW : CUSTOMER SATISFACTION ON AIRLINE TWEETS. 1, 896–904.

[16]. R, A. X. A., Mohan, V., & Venu, S. H. (2016). Sentiment Analysis Applied to Airline Feedback to Boost Customers' Endearment. International Journal of Applied and Physical Sciences, 2(2), 219–232. https://doi.org/10.20469/ijaps.2.50004-2.

[17]. Rani, S., Singh Gill, N., & Gulia, P. (2021). Analyzing impact of number of features on efficiency of hybrid model of lexicon and stack based ensemble classifier for twitter sentiment analysis using WEKA tool. Indonesian Journal of Electrical Engineering and Computer Science, 22(2), 1041. https://doi.org/10.11591/ijeecs.v22.i2.pp1041-1051.

[18]. Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00380-z.

[19]. Yoo, G., & Nam, J. (2018). A hybrid approach to sentiment analysis enhanced by sentiment lexicons and polarity shifting devices. The 13th Workshop on Asian Language Resources, 21–28. https://hal.archives-ouvertes.fr/hal-01795217.

[20]. Zayet, T. M. A., Ismail, M. A., Varathan, K. D., Noor, R. M. D., Chua, H. N., Lee, A., Low, Y. C., & Singh, S. K. J. (2021). Investigating transportation research based on social media analysis: a systematic mapping review. In Scientometrics (Vol. 126, Issue 8). Springer International Publishing. https://doi.org/10.1007/s11192-021-04046-2.