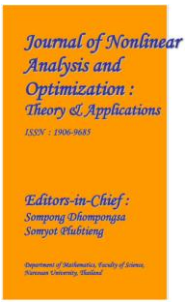
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.15 : 2024
ISSN : **1906-9685**



Paper ID: ICRTEM24_130 **ICRTEM-2024 Conference Paper**

OPINION MINING AND EMOTION CLASSIFICATION ON AIRLINE REVIEW

[#1]N.SAHITHI, ***UG Student***
[#2]N.AKSHAYA, ***UG Student***
[#3]P.APOORVA , ***UG Student***
[#4]DR. E. GURUMOORTHI, ***Associate Professor***
Department of Information Technology,
CMR COLLEGE OF ENGINEERING AND TECHNOLOGY, HYDERABAD.

ABSTRACT: Opinion Mining, also known as Sentiment Analysis, is the process of identifying, extracting, and categorizing opinions about various topics. This paper describes a complete technique to sentiment categorization and opinion mining in airline reviews. The method we use combines four machine learning algorithms: Support Vector Machine (SVM), XGBoost (XGB), Naive Bayes (NB), and AdaBoost. The goal is to correctly characterize the attitudes communicated in tweets about airlines, providing important data for airlines seeking to understand customer ideas and feedback. The approach begins with acquiring data from Kaggle, namely the Twitter US Airline Sentiment dataset. After preparing the data to increase its quality, we build and train the sentiment analysis model. Finally, we create a front-end interface that people can easily explore and interact with. Our technology improves the field by precisely categorizing tweets based on their mood, allowing airlines to make more educated judgments. Bar charts are a form of data visualization tool that helps us see how sentiment is distributed among multiple airlines. The primary purpose of sentiment classification is to thoroughly study internet documents such as blogs, comments, reviews, and news stories before classifying them as positive, negative, or neutral terms.

Keywords: Airline reviews, sentiment analysis, machine learning, Classification technique.

1. INTRODUCTION

Sentiment analysis has the ability to aid consumers in the airline choosing process by examining feedback from online platforms such as review sites and social media. This technology has been successfully deployed in a variety of industries, including automotive, education, and entertainment. This article investigates the use of Twitter data to automatically analyze the sentiment of customer evaluations for US airlines. Twitter serves as a repository of user-generated content, providing fast insights and permitting quick responses to events.

Twitter's unique limit on the amount of tweets allows for concise expression and speedy information distribution. The evaluation of tweets has the potential to improve information retrieval and allow for more informed decision-making. As a subset of data mining, text mining excels in extracting insightful

information from unstructured textual data. The use of sentiment analysis on Twitter data is a new advance in the field of text extraction. Scholars appreciate both the particular challenges and the multiplicity of potential implementations inherent in this methodology. Sentiment analysis allows airline industry stakeholders to gain a deeper understanding of customer preferences and perceptions, which can then be used to guide strategic decision-making and, eventually, boost overall customer happiness.

2. EXISTINGSYSTEM

In their study, Oza et al. [2] introduced a novel item-based recommender system that utilized machine learning techniques in addition to Google's BERT text classification method to train the model for sentiment analysis of airline reviews. This reduces the amount of data needed to train the system for specific tasks. By employing the graph and word cloud techniques, we were able to ascertain the rationale for the hostile or negative remarks. The y-axis remains constant, whereas the x-axis shows the frequency of the terms.

EkkaPujoAriesanto et al. [3] proposed a machine learning (ML) approach for opinion mining and sentiment analysis. The paper explains how online evaluations can be used to measure and predict consumer sentiment. The study's qualitative content was classified into core and supplementary services using aspect-level sentiment analysis. Similar neural networks (NN), SVMI, and ML models were used. It was discovered that using word embedding for feature extraction improved the performance of a support vector machine (SVM) model trained and evaluated on a dataset of airline tweets.

Ankita Rane and Dr. Anand Kumar introduced classification approaches for sentimental analysis and opinion mining [4]. Examples include Knn classification, decision trees, random forests, and support vector machines. The dataset used in this study consists of a variety of tweets collected from the regular Kaggle dataset. This page describes a variety of classifiers that use different classification techniques. While these techniques are frequently used for text classification, they can also be used for sentiment analysis of tweets.

Abdelrahman [5] et al. proposed a large number of classifiers for sentiment analysis in their data set. K-nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Naive Bayes (NB) are some of them. After the preprocessing phase, which removes noise from tweets by removing stopwords and punctuation, a corpus of cleansed tweets is created. The BoW model is then built to extract features from tweets, with the data divided into 70% training and 30% testing after the phase.

Rana Alqahtani et al. [6] proposed two widely used machine learning classifiers intended exclusively for sentiment analysis and opinion mining. Naïve Bayes and Logistic Regression are the two methods. An explanation of deep learning is presented, with specifics on the CNN (Convolution Neural Network), BERT, XLNET, and ALBERT models used by USairlinetweets. During the validation phase, the K-Fold Cross-Validation approach is used. K-nearest neighbors (KNN) and Naive Bayes (NB) are two machine learning techniques used to collect tweets as a dataset from the Twitter API.

Furqan Rustam et al. [7] offered a range of feature extraction and classification approaches. The chosen dataset was used to evaluate the performance of CC, LR, ETC, SVC, DTC, RF, ADB, GNB, SGDC, and GBM using term frequency (TF), term frequency-inverse document frequency (TF-IDF), and word features. For the purposes of this study, dataset pre-processing was performed. Various libraries and tools, including a natural language toolkit, were used in this stage. This study investigated two pre-processing strategies: partial pre-processing and complete pre-processing.

Heba Allah et al. [8] proposed using traditional approaches combined with two trending techniques (CNN/LSTM) to achieve an 82% accuracy rate. So (SACF) data analysis relies on the use of often embedding models for feature extraction. It uses the deep learning algorithms RNN, LSTM, GRU, CONVD, and BER. The application of transfer learning to big data sets. The tweet dataset has an approximate quantity of 65,944. The suggested study used a large dataset gathered from an airline application, as well as deep

learning models, to investigate user opinions.

3. PROPOSED SYSTEM

The proposed approach for sentimental categorization and opinion mining on airline reviews uses four machine learning algorithms: AdaBoost, Support Vector Machine (SVM), XGBoost (XGB), and Naive Bayes (NB). By properly categorizing attitudes indicated in airline reviews, this technology enables airlines to get significant insights into customer comments and opinions. Our methodology consisted of a methodical and cohesive sequence of events: first, we preprocessed the dataset to improve its quality; then, we built and trained the sentiment analysis model using machine learning algorithms; and finally, we implemented a user interface on the front end. We were able to properly categorize tweets based on sentiment and provide significant insights to users attempting to evaluate and interpret the sentiments communicated in textual material using this all-encompassing methodology. Figure 1 depicts the proposed system's block diagram.

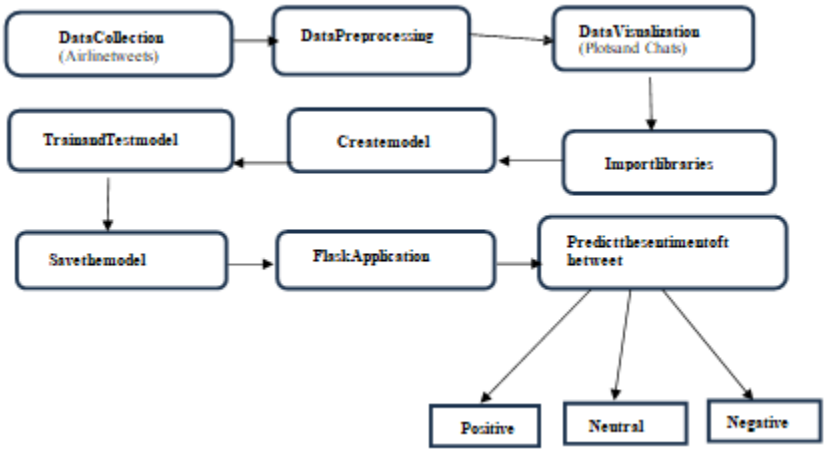


Figure 1 shows a block diagram of the proposed system.

Data Collection: The initial stage of the suggested approach involves gathering data from the web portal Kaggle.com. The dataset used is Twitter US Airline Sentiment [1]. It has 14640 rows and includes messages from six different airlines: American, United, Delta, United States Airways, and Virgin America.

Data Preprocessing: Following data capture, data preparation begins. The Twitter US Airlines Sentiment CSV file is loaded into a Jupyter notebook. NumPy and Pandas are used to help with data cleansing tasks such as stopword removal, notice cleaning, and unnecessary content removal, among others.

Data Visualization: The suggested system visualizes data using libraries such as Wordcloud, PIL, and Matplotlib. In this stage, the algorithm calculates estimates of negative, positive, and neutral feedback for each airline, as well as the total number of sentiment-based confidence feedback per airline. Following that, a DataFrame is created to enable visualization, displaying the tweet count as a bar chart with a dashed line (see Fig. 3).

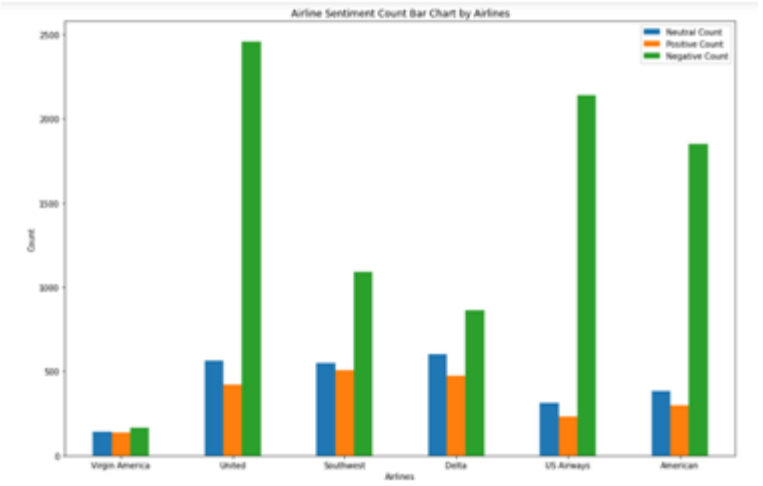


Fig3: Tweet count in a chair line

Import libraries: Python, pandas, matplotlib, Jupyter Notebook, Seaborn, Scikit-learn, Textblob, imutils, xgboost version 1.1.1, Django, and Flask are among the libraries that have been imported.

Create model: Using word frequencies, our system creates a classifier that predicts airline mood in future tweets. This eliminates the need for manual categorization, allowing airlines and interested parties to easily assess the sentiment of specific tweets. A classifier of this type allows for efficient analysis of future comments, aiding in attempts to improve quality.

Train and Test model: Our approach divides the dataset, which has 14,640 rows, into two different sets: training and testing. In general, the training set is divided between the first 11,712 rows and the remaining 2,928 rows in an 8:2 ratio. To enable effective model training, Twitter content is converted to vectors using the TfidfVectorizer.

Save the model: According to the machine learning map page, the system analyzes four different models: linear SVC (Support Vector Classification), Naive Bayes, XGBoost classifier, and AdaBoost classifier. Following that, the model with the highest level of precision is picked and kept for implementation.

Flaskapplication: To increase user interaction with the sentiment analysis model, the proposed system included a front-end interface that was integrated to the model using Flask, a lightweight Python web framework. Users can use this front-end interface to input tweets or text excerpts and receive sentiment analysis results from the model in real time. By integrating Flask with the sentiment analysis model, the system created an accessible and user-friendly platform for interacting with classification results.

Predict the sentiment of the tweet: When the model is completed with the airline's name and a tweet as inputs, the proposed system will produce the mood of the tweet as output.

The Support Vector Machine (SVM) distinguishes between linear and nonlinear data. SVM is a binary classifier that differentiates objects based on hyperplanes. Hyperplanes serve as a border to differentiate data elements that will be allotted to distinct classes.[9]

Equation (1) represents the formula for the linear SVM classifier used in instrumental classification.

Show a binary classification problem with two classes, +1 and -1, involved. A training dataset is available, which includes the input feature vectors X and the matching class labels Y.[10]

Consider the linear SVM classifier:

$$\hat{y} = \begin{cases} 1 & : \ w^T x + b \geq 0 \\ 0 & : \ w^T x + b < 0 \end{cases} \quad (1)[10]$$

Multinomial Naive Bayes is a probabilistic classifier that calculates the probability distribution of textual data. As a result, it is extremely compatible with natural language processing (NLP) jobs involving data including discrete frequencies or counts of events.[11]

Equation (2) is the formula for the multinomial Naive Bayes classifier used in incremental classification to

determine probability.
The likelihood of detecting a specific set of word counts in a document is described using the probability mass function (PMF) of the multinomial distribution, as conferred by:

$$P(D|c) = \frac{T_c!}{\prod_{i=1}^{V}(x_i!)} \prod_{i=1}^{V} \left(\frac{\theta^{x_i}_{c,i}}{x_i!} \right) \longrightarrow \quad (2)[11]$$

here,

is the word count for document D. indicates the likelihood that the term i will appear in a document of class C. represents the total quantity of words in papers of class C.
Both XGBoost and Adaboost are important in the field of sentiment classification, capitalizing on their respective advantages. While XGBoost excels at capturing intricate text relationships, which improves accuracy, Adaboost's iterative training method improves performance by incorporating weighted predictions. Both algorithms handle huge datasets efficiently, which is critical for practical sentiment analysis jobs across different domains. Their combined strengths and wide range of options highlight their usefulness in giving accurate sentiment analysis results.
Software prerequisites include platforms such as Anaconda and Jupyter Notebook, Python as the programming language, and NLTK and pip asstools. Furthermore, Flask serves as the foundation for the development process.

ResultAnalysis

The proposed sentiment analysis and opinion mining approach for airline evaluations will improve performance. The xgboost and adaboost techniques are used for implementation. The model is used to predict if a tweet is positive, negative, or neutral. The front end of this solution is linked to the model using a flask. We can input our tweets by picking an airline from the dataset.
The figures that follow show the outputs of the proposed system. A neutral tweet is presented as such; a good tweet is labeled "We hope you will fly with us again"; and a negative tweet is labeled as such.

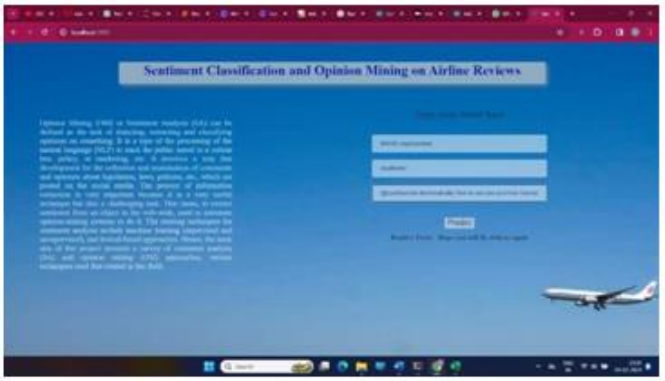


Figure 6: A page of results showing neutral tweets.

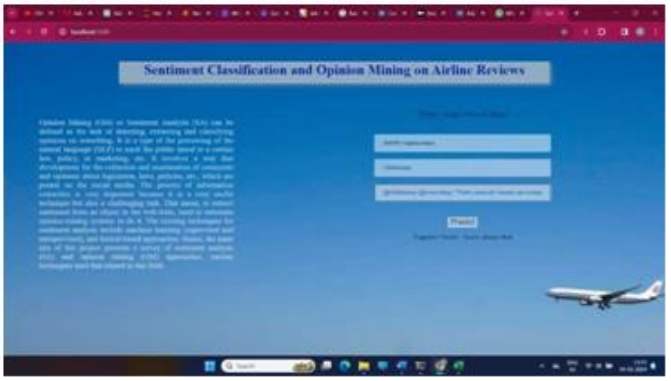


Figure 7: A page of results showing a nasty tweet.

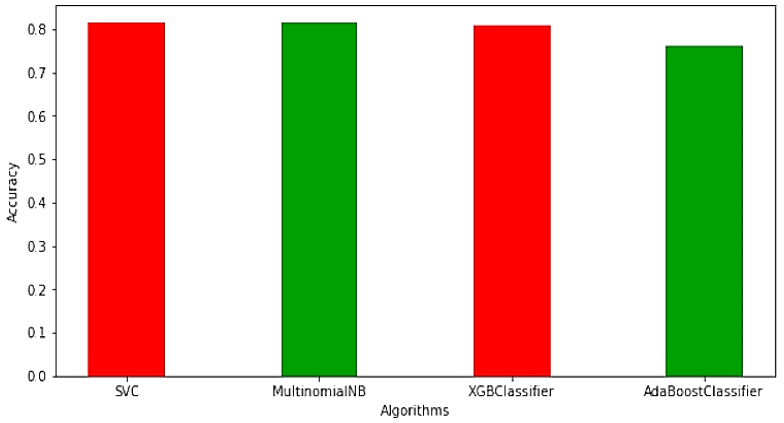
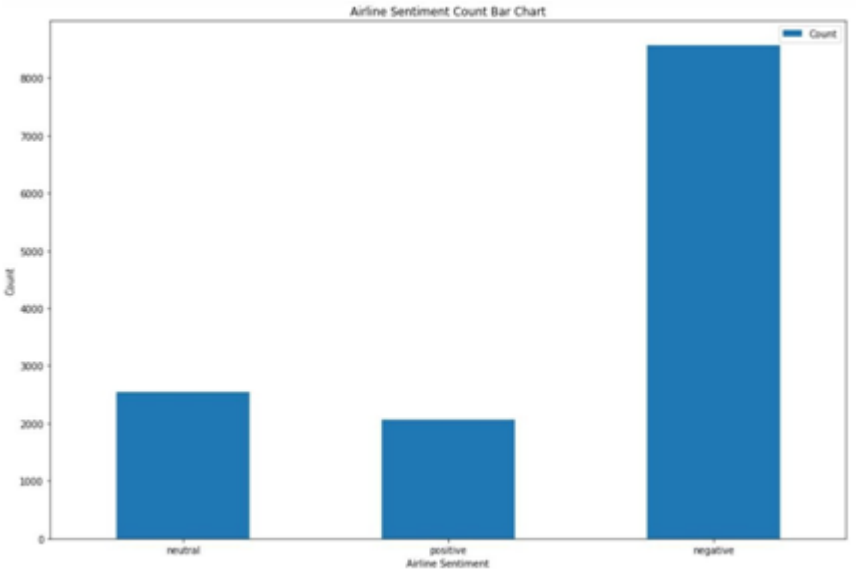


Figure 8: Precision.

Figure 8 depicts the relationship between numerous methods and accuracy. The graph allows us to observe the accuracy of each technique.



(Figure 9).Airline sentiment count bar chart

Figure 9 depicts a variety of counts for emotive analysis of the airline count bar. The values 2500, 2000, and 8500 represent neutral, positive, and negative sentiments, respectively.

4. CONCLUSION

Sentiment analysis is a new approach for understanding the wishes and needs of the general public. It is a more convenient and cost-effective way to understand public sentiment on a certain topic and evaluate the brand influence of microblogging. In this scenario, we addressed recent airline industry challenges and how the public perceived them in order to measure public mood. The sentiment classification and opinion mining project has greatly aided our understanding of passengers' sentiments and opinions. It serves as a valuable tool for improving airline services and ensuring that customers' expectations are met and exceeded. For this project, we used four classifiers and determined the accuracy of each. We found 81.45% for xgboost, which is the highest among all algorithms. We conclude that xgboost is the most precise algorithm.

REFERENCES

[1] Dataset : https://www.kaggle.com/crowdflower/twitter-airline-sentiment

[2] Patel, A., Oza, P., & Agrawal, S. (2023). Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model. Procedia Computer Science, 218, 2459-2467.

[3] Akhmad, E. P. A., Adi, K., & Widodo, A. P. (2023). Machine learning approach to customer sentiment analysis in twitter airline reviews. In E3S Web of Conferences (Vol. 448,
p. 02044). EDP Sciences.

[4] Rane, A., & Kumar, A. (2018, July). Sentiment classification system of twitter data for US airline service analysis. In 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 769-773). IEEE.

[5] Saad, A. I. (2020, December). Opinion mining on US Airline Twitter data using machine learning techniques. In 2020 16th international computer engineering conference (ICENCO) (pp. 59-63). IEEE.

[6] Al-Qahtani, R., & bint Abdulraman, P. N. (2021). Predict sentiment of airline tweet using ML models. (No. 5228). EasyChair

[7] 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), 1078.

[8] Samir, H. A., Abd-Elmegid, L., & Marie, M. (2023). Sentiment analysis model for Airlinecustomers' feedback using deep learning techniques. International Journal of Engineering Business Management, 15, 18479790231206019.

[9] Saad, Y. M. K. Omar, and F. A. Maghraby, "Predicting Drug Interaction With Adenosine Receptors Using Machine Learning and SMOTE Techniques," IEEE Access,vol. 7, pp. 146953–146963, 2019.

[10] https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[11] https://www.geeksforgeeks.org/multinomial-naive-bayes/