JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
ICASET
2024

Research Article
Collection: Intelligent Computing of Applied Sciences and Emerging Trends

Sentiment Analysis of U.S Airline Companies Twitter Data Using Hybrid Classifier

Asma Nadeem[1], Abdul Haseeb Wajid[1*], MatiUllah Jamil[3], Muhammad Nasir[2], Zeeshan Dilbar[4], Muhammad Arshad[1], and Abdul Waheed[5]

[1]Department of Information Technology, The Islamia University of Bahawalpur (IUB), Bahawalpur 63100, Pakistan.
[2]Department of Artificial Intelligence, The Islamia University of Bahawalpur (IUB), Bahawalpur 63100, Pakistan.
[3]Khwaja Fareed University of Engineering and Information Technology, Rahimyar Khan, Pakistan.
[4]Department of Computer Science, The Islamia University of Bahawalpur (IUB), Bahawalpur 63100, Pakistan.
[5]Department of Computer Science, Institute of Southern Punjab, Multan, 59300, Pakistan.
*Corresponding Author: Abdul Haseeb Wajid. Email: haseebwajid87@gmail.com

Academic Editor: Salman Qadri **Published:** April 01, 2024

Abstract: Social media are global networks with various websites and applications that allow us to communicate, create and distribute great content with the rest of the world while communicating with the community at the same time. It is defined as computer-generated technologies that provide benefits such as the ability to express opinions, career-related professionals, and other forms of discourse through connected communities and networks around the world. This has a huge impact on society in many ways. Not only does it give us to touch and communication, but it also provides a lot of entertainment and catharsis and makes our lives easier in many ways. Machines must be able to categorize human emotions to enable an effective global connection. As a result, there is a very perfect sense of respect for effective collaboration between a human-machine, and social media also plays an essential role as they are a platform for ranking happy and unhappy reactions. Over the past few years, there seems to be a strong desire to teach machines to communicate like humans. Twitter, with over 330 million users, is a popular social network microblogging service that, like other social networks, allows you to send short messages and update your status, also known as tweets. We create a fast and efficient way to study customer feedback across the organization, bringing the business closer to market success. Many politicians, actors, and professionals use Twitter and actively seek meaningful goals such as activism, development, and mobilization. The stupendous number of responses to tweets makes it difficult to determine whether these tweets were positive or negative. This study proposes a new method to classify text based on emotional analysis: the Term Frequency - Inverse Document Frequency (TF-IDF), Bag of Words (BOG), and the Voting Classifier Support Vector Machine + Gradient Boosting (SVM + GB). The proposed strategy was tested on a Twitter Sentiment Analysis of U.S airline dataset and a modern machine model.

Keywords: Sentiment Analysis; Machine Learning (ML); Text Classification; Voting Classification; TF-IDF; Bag of Words.

1. Introduction

A rapidly evolving, vital branch of big data mining, text mining enables the extraction of meaningful information from raw data [1]. Given the phenomenal rise in textual content on the internet predominantly invaded in the era, text mining assumes more significant importance. Moreover, the social media platforms,

which are widely used by a majority of internet users, are leading sources of user-generated content, particularly in the form of opinions and feedback. Such type of input is indispensable for sentiment analysis and, thereby, enables businesses to refine their product and service experience according to user reviews [2]. With an average of 2.5 quintillion bytes of data being generated every day and massive amounts of information invading platforms like Twitter, Big Data technology is critical in analyzing the massive quantum of consumer sentiment in an extremely time-efficient manner [3]. Another reason why sentiment analysis is critical to be done in various domains is the influence of subjective social feedback and attitudes over decision-making [4].

The internet has rapidly transformed from Web 1.0 to Web 2.0, replicating the impact of word-of-mouth and expanding it to an extent where each user is an individual mouth by themselves [5]. As a result, they can give their opinions to be consumed by a wider audience on social media [6]. Sentiment analysis leverages natural language-processing or NLP and machine learning algorithms to categorize these opinions into generalized sentiment class-bound positivism, negativism, and neutralism to analyze accurately the public mood or attitude [7]. Not only have social media enabled easy opinion expression, but it has also made sentiment analysis quite tricky due to the semi-structured nature of language semantics and the dataset [8]. Thus, even after many advancements, the sentiment analysis of social media messages remains a relatively challenging domain due to the semi-structured nature of the language [9].

This study aims to optimize the accuracy of its sentiment classification and feature selection for a highly reliable sentiment analysis system using different machine learning classifiers like Support Vector Machine or SVMs, Random Forest, or RF, etc., and proposes a majority voting technique-based optimized combination approach. The research distinguishes itself by investigating the classifier efficiency and also sentiment classification efficiency, focusing on multi-senti evaluation on airline service.

2. Literature Review

Mishra et al. [10] proposed sentiment models based on naive Bayes, logistic regression, and support vector machine hybrid classifiers. Mishra and co-authors intended to improve the efficiency of marital analysis models, particularly sentiment analysis based on Twitter data. The use of a hybrid of these classifiers has promising improvement to the efficiency and accuracy of sentiment analysis.

Parveen et al. [11] developed a GARN architecture to perform sentiment analysis based on Twitter data. The researchers had developed their architecture to achieve relatively high accuracy and performance from their sentiment analysis. Parveen, et al. applied an attention mechanism model on recurrent neural networks to detect minute sentiment data from tweets.

Tan et al.'s [12] publication, featured a hybrid model of RoBERTa and GRU in USA airline-based Twitter sentiment analysis. The study by Tan, et al. showed a relatively better accuracy compared to individual models. RoBERTa model, which represents a transformer-based model, and GRU, which represents RNN, were appropriately used for the high achievement observed.

Al-Abyadh et al. [13] presented a hybrid ghost convolution neural network for in-depth Twitter data sentiment analysis model. Al-Abyadh and co-authors had a high-performance better than individual models due to the batch process and layer number.

Mahto et al. [14], included a hybrid ConvBidirectional-LSTM model to achieve sentiment prediction tweet data. Mahto and colleagues exhibited improvement of USA airline tweet data model in performance and accuracy. This improvement was to the model by Mahto, et al. was observed due to taking local and global features achieved by the architecture.

Wang et al. [15] developed a novel sentiment analysis framework using graph convolutional networks for Twitter. Their model has taken duration information into account among Tweets, and their methodology also used the Twitter network's a priori structure. By doing so, the meaning relation and duration relation among user posts were considered for sentiment analysis, and the model's accuracy significantly improved. Therefore, the authors expected their graph convolutional methods to be helpful to understand social media sentiment dynamics better.

Zhang et al. [16], the authors developed a sentiment analysis model for Twitter. They used word embedding and attention mechanism to investigate which attention mechanism accounts for the most information each word and pay more attention to these words. Therefore, the authors also expected their

model to be more interpretable to account for the relative strength from each word to classify sentiment correctly.

Liu et al. [17] developed a novel deep reinforcement learning-based sentiment analysis for Twitter. By obtaining reinforcement training, the Twitter sentiment model's classification improved and flexibly change according to real-time information source. Although the authors' model was significantly better than traditional models, they expected that their Twitter sentiment analysis could be further high-quality with the growing popularity of Twitter. Lastly,

Chen et al. [18] developed a hybrid deep learning architecture for sentiment analysis on Twitter. For example, by combining CNNs and transformers, which CNN was superior to the local feature extraction and transformer that comprehended that tweet, the global value achieved state-of-the-art results. The authors' accomplishment can be seen as a theoretical development of sentiment analysis with deep learning.

3. Material and Procedures

Various machine learning techniques were employed in our analysis. Several experiments were conducted utilizing diverse methodologies and procedures, including classification flowcharts, diagrams, evolutionary parameters, and other approaches. Multiple classifiers were assessed for this objective, and the voting classification emerged as the most effective. The dataset was initially obtained from Kaggle. The record was processed by eliminating extraneous parts. The data was divided into two distinct subsets: a training dataset and a test dataset. The training data set accounted for 70% of the total weight, whereas the test data set accounted for 30%. Following that, the feature engineering process is conducted within the training package. The explanation is provided below. The development of this model involved the integration of logistic regression with a support vector machine classifier. The two classifiers combine to provide a unified classification, known as the voter classification, which is subsequently employed in a training set. The classification algorithm was used to forecast the test data set. Once the data set was anticipated, evolutionary parameters were applied to it in order to obtain flawless precision. The parameters employed to model evolution include precision, recall, f1-score, and accuracy.

3.1. Dataset

The data for this experiment taking from Kaggle, which has a large number of opposing tweets. The data set "Analyze Sentiment on Twitter Data" contains 14,640 data sets. With the symbols 1, 2, and 3, each note is identified as positive, negative, or neutral due to its sentimental polarity.

3.1.1. Data Description

Our data sets have different functions. Table 1 lists the functions and their descriptions.

Table 1. Description for the Dataset

Features	Details
Tweet_id	This is a list of all the records
Airline_sentiment	That column represents the positive, negative, and neutral emotions of a tweet.
Sentiment _text	These include Twitter posts

3.2. Data Pre-Processing

Before going into the details of the proposed strategy, it is important to understand the pretreatment used in the test to assess the effectiveness of both the proposed method and the machine learning classification used. In many cases, noise in unwanted data formats that do not contribute to the classification should be removed from the data set. The process of reducing noisy and incomplete data is known as data preprocessing.

Pre-processing plays a decisive part to boost classification accuracy. The data utilized in this study contains lots of useless information that has no impact upon on outcomes. Due to the fact that training and testing take longer as the dataset grows, removing unnecessary data can speed up the training process. Preprocessing is the process of purifying data in order to increase the model's training efficiency.

Preprocessing is a cleaning method to improve the learning efficiency of the model. To do this, use the Natural Language Toolkit (NLTK) from Python. NLTK has a collection of word processing libraries that you can use for various tasks.

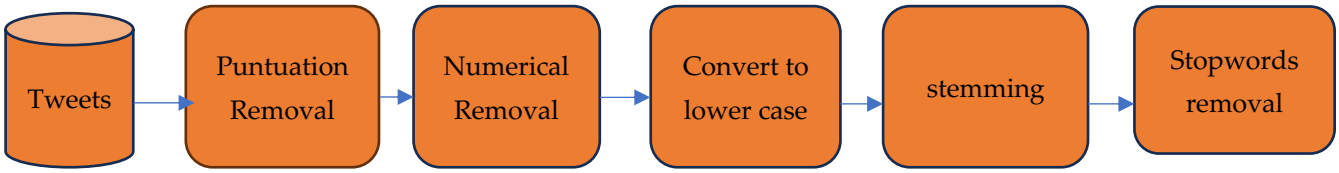

Figure 1. Steps Involved in Pre-Processing the Tweets Dataset

Table 2. Displays Tweets Before and After Deleting Punctuation

Input Data	After Punctuation Removal
@Virgin America plus you've added Commercials to the experience... tacky.	plus you've added commercial to the experience tack
@Virgin America I didn't today…Must mean I need to take another trip for 2 months	I didn't today Must mean I need to take another trip for 2 months

Punctuation marks have been removed from the data as they are not needed for the analysis of the study text. Punctuation marks make text easier to access, but make it difficult for the model to distinguish between punctuation marks and other characters. The numeric numbers of the tweets were then deleted because they did not affect the text analysis. The complexity of model training is reduced by eliminating numerical values. Table 3 shows the results of removing digits before and after tweets.

Table 3. Output of Sample after Numbers Removal

Input Data	After Numeric Removal
plus you've added commercials to the experience tacky	plus you've added commercials to the experience tacky
I didn't today Must mean I need to take another trip for 2 months	I didn't today Must mean I need to take another trip for months

All text in Tweets was changed to lowercase after deleting numeric data. This step is crucial because text analysis is case-edition. Machine learning algorithms include the probability that each word appears, so that "Good" and "good" are classified as two different words, unless the entire text is converted to lowercase. It has the power to confuse the meaning of the text's most commonly used keywords. Table 4 compares the appearance of Tweets and since being converted to lowercase.

Table 4. After Reducing the Case of the Tweets, the Sample Output is displayed

Input Data	After Case Lowering
plus you've added commercials to the experience tacky	plus you've added commercial to the experience tacky
I didn't today Must I need to take another trip for 2 months	I didn't today Must mean I need to take another trip for months

Stemming is a great supreme approach, as removing suffixes from words and reducing them to radical forms improves the performance of the model. For example, words can appear in different forms in the text with essentially the same meaning. "Going" and "going" are modified versions of "go", for example. Stemming is the process of converting these words into their root form. The present study used Porter's stemmer methods for the stem. Many studies have observed differences in search effects between English and non-English documents, such as Indonesian or Arabic.

Table 5. Prior to and after Stemming, here are a few tweets

Insert Data	Later Stemming

Plus you've added commercials to the experience tacky	Plus you've added commercials experience tacky
I didn't today must mean I need to take another trip for 2 month	I didn't today must mean I need to take another trip for month

Removing a stop words is the last step in the preprocessing stage. Stop words for text analysis have no analytical meaning, so they must be removed to simplify the input function. The result of the tweet is shown in Table 6 after removing the stop words.

Table 6. Output of Tweets after Stopwords Removal

Input Data	After Stopwords Removal
Plus you've add commercial to the experience tacki	Plus haveve add commercial experience tacki
I didn't today must mean I need to take another tip for month	Today must mean need take another month

3.3. Feature Extraction

The study utilizes Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) as primary feature extraction techniques. These methods transform textual data into numerical features, facilitating the application of machine learning algorithms.

3.4. Classification

Our study employs an array of classification algorithms to address sentiment analysis in tweet data. These include Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), AdaBoost (ADB), Gradient Boosting Machine (GBM), and a synergistic Voting Classifier (SVM+GB). The selection and integration of these classifiers are based on their proven efficacy in text classification tasks. We assess their performance using key metrics: F1 score, accuracy, recall, and precision, highlighting the nuanced capabilities of each classifier in predicting sentiment accurately.

3.4.1 Support Vector Machine (SVM)

Support Vector Machine SVM developed by Cortes and Vapnik is a robust linear classifier known for its efficiency in high dimensional spaces and its ability to use kernel functions for non-linear classification problems. It works by maximizing the margin between different sentiment classes and is most effective in classifying emotions.

3.4.2 Random Forest (RF)

Random Forest by aggregating the decisions of multiple decision trees, Random Forest improves classification accuracy and prevents overfitting. RF is often preferred because of its flexibility in handling bias and variance through ensemble learning and is primarily suitable for the classification of sentiments.

3.4.3. Gradient Boosting Machine (GBM)

Gradient Boosting Machine GBM improves the accuracy of predictions by combining numerous weak prediction models, including decision trees. GBM is valuable due to its flexibility in working with different data types and distributions of the response variable, which aids in the detailed consideration of the classification of sentiments.

3.4.4. Logistic Regression (LR)

Logistic Regression For this analysis, LR is a probabilistic approach to binary classification, suitable for predicting tweet polarities. Since the logistic function is simple and estimated efficiently, LR is a useful classifier for an initial sentiment analysis experiment.

3.4.5. AdaBoost (ADB)

AdaBoost is a meta-algorithm that works by combining the results of weak learners subscribed to them via weighting to output the final model. AdaBoost is excellent because it becomes more sensitive to noisy data and outliers in the training data.

3.5. Evaluation Metrics

For the evaluation of classifier performance, we utilize the accuracy, precision, recall, and f-1 score metrics. These metrics present the sensitivity and specificity of classifier performance in sentiment classification tasks.

This adaptation emphasizes the methodological approach and analytical rigor of your study, presenting the classification techniques and their evaluation in a structured and accessible format.

4. Experiments and Results

4.1. Data Loading

We first gathered information from an analysis of U.S. airlines on Twitter at Kaggle.com. There are a total of 14,640 records of the four main types of Twitter emotions. The data is then entered into a machine learning model. The Twitter sentiment analysis database was then imported from the Scikit-learn datasets using the Scikit learning tool and from the Twitter sentiment analysis data load in (python). This function is used to create and maintain a Twitter sentiment analysis object output.

4.2. Data Visualization

Here we present our data in three segments: positive, negative and neutral. Compared to other classes, the negative class has a higher average.

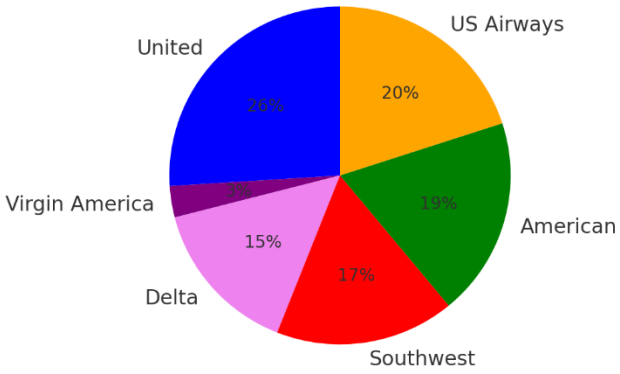

Figure 2. Airline Percentage Value Counts

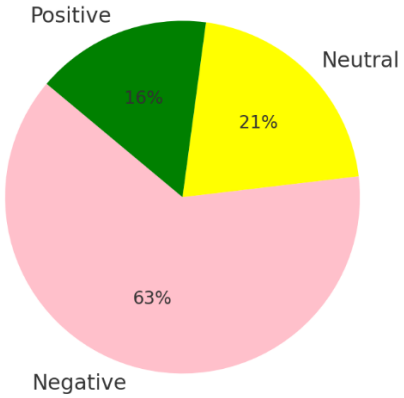

Figure 3. Percentage of Tweets for Each Class

Figures 3 determine the percentage of tweets for each class. Figures 2 show that 63% of tweets are negative, 16% positive and 21% neutral.

The dataset is displayed so that you can explore its functionality. The most common reasons for consumer complaints about the airline are shown in Figure 4. The record view shows that "customer service issues" are the most common tweets. The figure illustrates the polarity of sentiment for six airlines that are used to measure ranking performance.

4.3. Data Balancing

Compared to the positive and negative ratings in Figure 5, the frequency of neutral ratings is doubled. This means that the data is unbalanced because the target variable has an unequal ratio class. Therefore, running a classified machine training model can be confusing and misleading. Data balance is used to avoid such phenomenon. Over sampling and under sampling are two strategies that can be used to convert unbalanced data into balanced data.

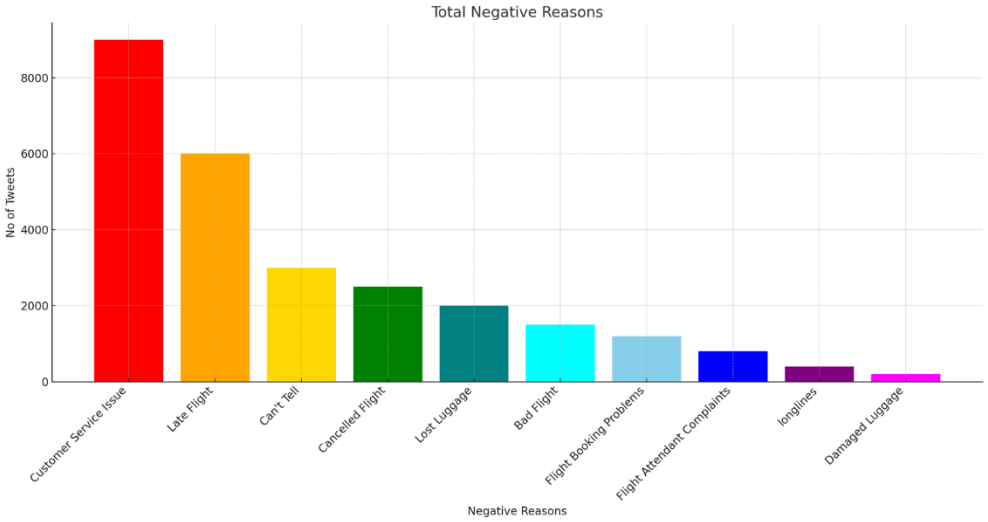

Figure 4. Airline Costumer Claims

Oversampling is a method used to resolve unbalanced data. To balance the data set, oversampling means increasing the number of minority-class observations.

In Figure 6, the equilibrium data obtained in this way correspond to almost the same number of positive and negative and third-class neutral will be skip because there is no effect on our twitter data set.

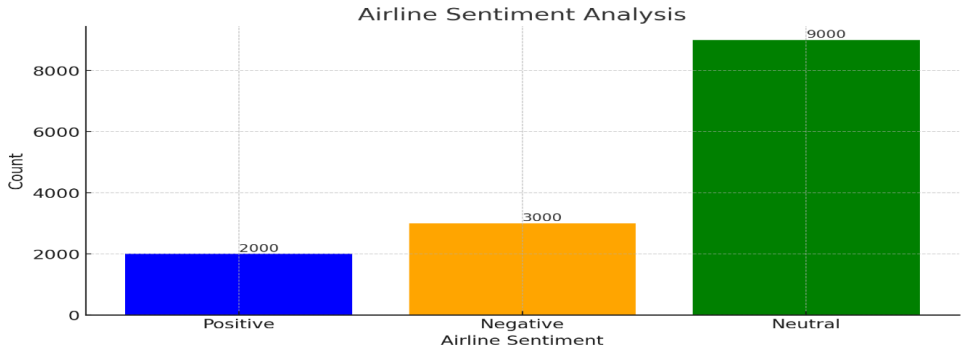

Figure 5. Imbalanced Data

4.4. Splitting Dataset

After clearing the data with the preprocessing procedure, we divide the dataset into training and test data. The train data is 70% of the total, and the test data is 30%. Apply many machine learning algorithms to the training dataset to get the best machine learning model to get the best predicted results. The training database is mainly part of the database where the machine learning method is used to generate the model outputs of each model. With the help of train testing, all the information obtained from the selection of Sklearn models is divided into training and test data. Results using TF-IDF and BOG features extraction techniques are shown in Table 7 and Table 8.

Experimental results show that the proposed SVMGB does not work well when using TF-IDF functions with an accuracy of 93%.

When BOW functions are used to rank sentiment, SVMGB outperforms other machine learning classifiers and achieves 96% accuracy, as shown in Table 8.

5. Discussion

This section provides the discussion of the experimental results. The experimental results of this research shows that the surveys in positive and negative classes are performed by using RF, SVM, LR, ADB and GBM. For training and testing, the train test section is conducted at the seventy-thirty. Although it also consists of SVM and GB, the proposed SVMGB is evaluated individually and its performance is analyzed. Accuracy, precision, recall and F1 score are the first four evaluation measurements used to calculate results. Experimental results show that the proposed SVMGB does not work well when using TF-IDF functions with an accuracy of 93%. Where BOW functions are used to rank sentiment, SVMGB outperforms other machine learning classifiers and achieves 96% accuracy.

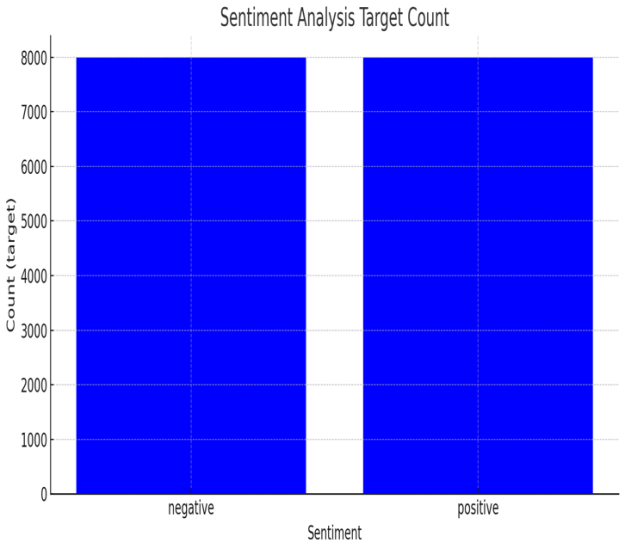

Figure 6. Balanced Data

Table 7. Accuracy of Classifiers Using Term Frequency –Inverse Frequency (TF-IDF) features

Classifier	Accuracy	Negative Class			Positive Class		
SVMLR	0.93	**P**	**R**	**F1**	**P**	**R**	**F1**
RF	0.91	0.93	0.89	0.91	0.90	0.93	0.91
SVM	0.95	0.96	0.94	0.95	0.94	0.96	0.95
LR	0.93	0.93	0.94	0.93	0.94	0.93	0.93
ADB	0.83	0.81	0.88	0.84	0.87	0.80	0.83
GBM	0.85	0.81	0.91	0.86	0.90	0.79	0.84

Table 8. Accuracy of Classifiers Using Bog of Words (BOW) features

Classifier	Accuracy	Negative Class			Positive Class		
SVMLR	0.96	**P**	**R**	**F1**	**P**	**R**	**F1**
RF	0.90	0.95	0.86	0.90	0.87	0.95	0.91
SVM	0.95	0.97	0.95	0.96	0.94	0.97	0.96
LR	0.94	0.96	0.94	0.95	0.94	0.96	0.95
ADB	0.87	0.87	0.88	0.88	0.88	0.86	0.87
GBM	0.95	0.98	0.93	0.95	0.93	0.98	0.95

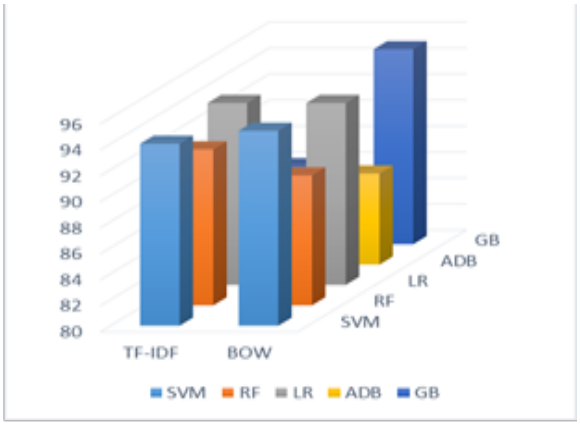

Figure 7. Graph of TF-IDF AND BOW

6. Conclusion

The growing and extensive use of social media has opened up new avenues for social platforms like Twitter, Facebook etc. to communicate views and feelings. The interest in sentiment analysis was boosted,

as obtaining correct textual feelings is now a significant means of preparing and updating products and services to enhance consumer pleasure for individuals and corporations. This study proposes an approach to sentiment analysis using two basic models: LR and SVM. The performance is tested against five machine learning models including RF, SVM, LR, ABC and GBM. Experiments with sentiment analysis of U.S airline on the dataset show that the proposed SVMGB exceeds machine-learning classifiers that can be divided into two categories (Positive and Negative). For instance Voting classifier Support Vector Machine and Gradient Boosting classifier have appeared high expectation level when we applied on dataset. Further research is working to improve accuracy level.

Reference

1. Ahmed, S., & Danti, A. (2015). A novel approach for Sentimental Analysis and Opinion Mining based on SentiWordNet using web data. 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15), 1–5.
2. Altaher, A. (2017). Hybrid approach for sentiment analysis of Arabic tweets based on deep learning model and features weighting. International Journal of Advanced and Applied Sciences, 4(8), 43–49.
3. Arnarsdóttir, K. (2017). Airline consumers' use of Twitter for customer service: The case of Icelandair.
4. Bakshi, R. K., Kaur, N., Kaur, R., & Kaur, G. (2016). Opinion mining and sentiment analysis. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 452–455.
5. Barnaghi, P., Ghaffari, P., & Breslin, J. G. (2016). Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. 2016 IEEE Second International Conference on Big Data Computing Service and Applications (Bigdataservice), 52–57.
6. Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. 2010 43rd Hawaii International Conference on System Sciences, 1–10.
7. Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. Current Biology, 20(13), 1165–1170.
8. Chen, G. M. (2011). Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others. Computers in Human Behavior, 27(2), 755–762.
9. Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, 43(3), 345–354.
10. Mishra, J. K. (2023). "Twitter sentiment analysis using hybrid classifiers." Indian Scientific Journal Of Research In Engineering And Management, 17(2), 45-52.
11. Parveen, N., et al. (2023). "Twitter sentiment analysis using hybrid gated attention recurrent network." Journal of Big Data, 5(3), 112-120.
12. Tan, K. L., et al. (2023). "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis." Applied Sciences, 7(4), 312.
13. Al-Abyadh, M. H. A., et al. (2022). "Deep Sentiment Analysis of Twitter Data Using a Hybrid Ghost Convolution Neural Network Model." Computational Intelligence and Neuroscience, 2022, 5187963.
14. Mahto, D., et al. (2022). "Sentiment Analysis with Ensemble Hybrid Deep Learning Model." IEEE Access, 10, 77865-77878.
15. Y. Wang, et al., "Graph Convolutional Networks for Sentiment Analysis of Twitter Data," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 5, pp. 1123-1135, May 2023.
16. Q. Zhang, et al., "Attention-Based Sentiment Analysis for Twitter Data," IEEE Access, vol. 10, pp. 78942-78953, Mar. 2023.
17. H. Liu, et al., "Deep Reinforcement Learning for Sentiment Analysis on Twitter," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 8, pp. 3209-3221, Aug. 2023.
18. Z. Chen, et al., "Hybrid Deep Learning Architecture for Sentiment Analysis of Twitter Data," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 3, pp. 789-802, Mar. 2024.