ISSN 1870-4069

US Airlines Twitter Opinion Analysis: Classifying Positive or Negative Comments

Raul Jimenez-Cruz, Galo Ruiz-Soto,
Miguel Gonzalez-Mendoza

Tecnologico de Monterrey,
School of Engineering and Sciences,
Mexico

{r.jimenez.c, a01799399, mgonza}@tec.mx

Abstract. This study investigates the classification of sentiment in tweets related to airlines, aiming to determine whether opinions are positive or negative. The dataset includes features such as the airline mentioned, sentiment ranking, geolocation, and the sentiment label. Six classifiers were evaluated for their effectiveness in sentiment classification. The preprocessing phase involved lemmatization, the removal of stopwords to clean the text data, and the generation of bigrams to mitigate the sparsity of the sparse matrix. Given the dataset's imbalance with an Imbalance Ratio (IR) of 1.68, the balanced accuracy metric was employed to ensure a fair assessment of classifier performance. The classifiers' outputs were geographically mapped to provide a visual representation of sentiment distribution, facilitating a more tangible analysis of the results. Among the classifiers, Logistic Regression achieved the highest accuracy (0.7281), while Multinomial Naive Bayes obtained the best balanced accuracy (0.7920). This study demonstrates the importance of robust preprocessing and the selection of appropriate evaluation metrics in handling imbalanced datasets, contributing valuable insights into the performance of different classifiers in sentiment analysis tasks within the domain of natural language processing and machine learning.

Keywords: Sentiment analysis, airline, classification, algorithm, prediction, machine learning, text analysis.

1 Introduction

In February 2015, travelers provided opinions about their experiences using six different US airlines: American, United, Southwest, Delta, Virgin America, and US Airways. These opinions were classified as positive, negative, or neutral. Airlines recognize that detecting and understanding customer emotions is crucial for providing a superior customer experience. Satisfied customers are more likely to repurchase tickets from the same airline, as emotions can significantly influence future buying decisions. Identifying emotions promptly can help airlines adjust their services to improve customer moods, thereby reducing churn rates. Anticipating user emotions is key for securing upskilling opportunities and retaining customers.

In this study, neutral sentiments were merged with positive sentiments to address the challenge of class imbalance inherent in the dataset. This approach simplifies the classification task, making it more manageable and allowing for more robust model performance, particularly when using balanced accuracy as a metric. The conversion to a two-class problem aligns with the research objective of assessing overall customer satisfaction, where neutral opinions are often considered closer to positive than negative in this context. This paper aims to predict the sentiments of US airline users by analyzing their emotions expressed on Twitter.

The study will involve a comprehensive analysis of the data to identify any issues that could impact the results. Various classification models will be employed, including K Nearest Neighbors (KNN) with 1 and 3 neighbors, Multinomial Naïve Bayes, Random Forest, Gradient Boosting, and Logistic Regression, to predict the sentiments. The performance of these models will be assessed using accuracy metrics and confusion matrices. In particular, the balanced accuracy metric will be used to address potential biases given the imbalanced nature of the dataset. As in [10] have argued, accuracy should not be taken as an absolute measure. Thus, the study will explore alternative performance measures to ensure robust model evaluation.

2 Methodology

2.1 Dataset

The dataset is publicly available on Kaggle [8]. The dataset comprises 15 columns: 'tweet_id', 'airline_sentiment', 'negative_reason', 'airline', 'airline_sentiment_ confidence', 'retweet_count', 'text', 'negativereason_confidence', 'airline_ sentiment_gold', 'name', 'negativereason_gold', 'tweet_coord', 'tweet_location', 'user_timezone', and 'tweet_created'.

However, most of these columns are irrelevant for the current task. For instance, 'tweet_id', 'name', 'tweet_coord', and 'tweet_created' do not contribute to the sentiment classification task. The dataset contains opinions from Twitter users about US airlines, categorized into three classes: positive, negative, and neutral, as indicated in the 'airline_sentiment' column, which is the target variable for prediction. Due to the irrelevance of most columns to the problem, only the 'airline_sentiment' and 'text' columns were initially retained.

However, the neutral class was merged with the positive class to address class imbalance, reducing bias observed in initial experiments where the neutral class skewed towards positive. This adjustment resulted in a class imbalance ratio of 1.68. In addition to the 'text' column, the 'airline_sentiment_confidence' and airline columns were included as features.

The 'airline' column underwent label encoding to convert textual data into numerical format. The dataset contains 14,640 messages: 9,178 negative, 3,099 neutral, and 2,363 positive before merging classes. After combining the neutral and positive classes, the distribution is more balanced but still poses an imbalance challenge. There are no missing values in the columns. The messages are distributed across six airlines: United (3,822 messages), US Airways (2,913 messages), American (2,759

Table 1. Messages divided by airline.

Airline	Negative	Neutral	Positive
American	1,960	463	336
Delta	955	723	544
Southwest	1,186	664	570
US Airways	2,263	381	269
United	2,633	697	492
Virgin America	181	171	152

messages), Southwest (2,420 messages), Delta (2,222 messages), and Virgin America (504 messages). A breakdown of negative messages reveals 2,910 concerning customer service issues, 1,665 about late flights, 847 regarding canceled flights, 724 about lost luggage, and 1,190 messages that are ambiguous. Table 1 provides a detailed breakdown of the dataset. The preparation of the data ensures that the features selected are pertinent for the classification task and addressing class imbalance through the combination of neutral and positive classes.

2.2 Preprocess

The first step in the methodology was to clean the text field of the dataset. This involved several preprocessing actions to ensure the text data was suitable for analysis. Initially, a dictionary of contractions was created and used to expand contractions into their full forms. Following this, the text field was cleaned by removing numbers and punctuation. Further preprocessing included lemmatization and the elimination of stopwords, with all text converted to lowercase to ensure uniformity. Lemmatization is the process of reducing words to their base or root form, which helps in normalizing the text and reducing the dimensionality of the data [9].

Stopwords are common words that are typically removed from text data because they do not contribute significant meaning and can lead to noise in the analysis [12]. Once the tweets were cleaned, a new DataFrame was created containing the ‘text’, ‘airline_sentiment_confidence’, and ‘airline’ columns. The ‘airline’ column was subjected to label encoding, converting the categorical text data into numerical labels. The ‘airline_sentiment’ column, which served as the target variable, was similarly label encoded. The next crucial step was vectorizing the ’text’ column.

Term Frequency-Inverse Document Frequency (TF-IDF) was employed for this purpose. TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents [11]. An important modification was made: bigrams were utilized to capture more context and reduce the feature space. A bigram is a sequence of two adjacent words in a text, which helps in understanding the context better than individual words [7]. Additionally, words appearing fewer than three times were excluded to further streamline the feature set. The resulting sparse matrix from the TF-IDF vectorization was then converted into a DataFrame.

Table 2. The top 10 most important features determined by random forest classifier.

Feature	Importance
airline sentiment confidence	0.551708
airline	0.152976
flight cancelled flightled	0.009168
cancelled flightled flight	0.007575
cancelled flight flight	0.004096
flight booking problem	0.003839
flight cancelled flighted	0.003822
reflight booking problem	0.003039
worst customer service	0.002816
great customer service	0.002627

This DataFrame, now containing the TF-IDF features, was merged with the 'airline_sentiment_confidence' and label-encoded 'airline' columns, consolidating all relevant features for the classification task. This preprocessing pipeline aimed to prepare the data for effective sentiment classification by enhancing the quality and relevance of the features. After preprocessing, the next step was to analyze the new distribution of classes and the data, ensuring the changes adequately addressed the class imbalance and prepared the dataset for model training and evaluation. By carefully cleaning the text data, encoding categorical variables, and using TF-IDF with bigrams, the preprocessing aimed to create a robust feature set for the subsequent classification models. These steps were essential to mitigate noise and imbalance, thus improving the accuracy and reliability of the sentiment predictions.

2.3 Analysis of Preprocessed Patterns

The first step in the analysis involved visualizing the distribution of the classes to observe the imbalance. The negative class contains 9,178 patterns, while the combined positive class has 5,462 patterns. Next, a Random Forest classifier was trained on the entire dataset to determine which features were most important. The top 10 most important features across the dataset are shown below: Table 2 above displays the features and their respective importance scores as determined by the Random Forest classifier.

The most important feature is 'airline_sentiment_confidence', followed by 'airline'. The remaining features, although having lower importance scores, indicate specific terms related to customer issues such as flight cancellations and booking problems. This analysis helps in understanding which features contribute most significantly to the sentiment classification task. The analysis of the most representative words for each class revealed distinct patterns.

For the positive class, terms such as ”fleet fleek,” ”customer service,” and ”thank much” were predominant, highlighting a focus on positive customer experiences and gratitude. In contrast, the negative class was dominated by terms related to service issues and disruptions, such as ”customer service,” ”cancelled flightled,” and ”late flight.” These findings suggest that positive tweets often emphasize appreciation and specific positive interactions, while negative tweets predominantly address complaints about service failures and delays. This lexical analysis provides valuable insights into the differing nature of customer feedback based on sentiment, which can be crucial for tailoring responses and improving service quality.

2.4 Classification

In this section, six different classifiers were implemented and evaluated: KNN 1, KNN 3, Multinomial Naive Bayes, Random Forest, Gradient Boosting, and Logistic Regression. Aggarwal and Zhai [1] provide a comprehensive survey of text classification algorithms, highlighting the effectiveness of machine learning techniques in text mining. Each classifier was subjected to k-fold cross-validation with k=10 to ensure robust performance evaluation. Below is a brief description of each classifier along with relevant references.

1. **K-Nearest Neighbors (KNN):** KNN 1 and KNN 3: KNN is a non-parametric method used for classification and regression. In KNN, the input consists of the k closest training examples in the feature space. KNN 1 uses the closest neighbor, while KNN 3 uses the three closest neighbors [3].
2. **Multinomial Naive Bayes:** This classifier is based on Bayes' theorem and is particularly suited for classification with discrete features. The multinomial variant is specifically useful for text classification where word frequencies are used as features [6].
3. **Random Forest:** This ensemble learning method constructs multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees. It is known for its robustness and ability to handle overfitting [2].
4. **Gradient Boosting:** Gradient Boosting builds models sequentially, with each new model attempting to correct the errors of the previous models. It combines the predictions of multiple base estimators to improve robustness [5].
5. **Logistic Regression:** This linear model estimates the probability that an instance belongs to a particular class. It is particularly useful for binary classification problems but can be extended to multiclass problems [4].

The classifiers were configured with specific hyperparameters to optimize their performance. For Multinomial Naive Bayes, alpha was set to 0.001 with fit_prior was True. Random Forest used criterion was gini, number of estimators were 100. Logistic Regression was configured with max_iter in 300, solver was liblinear and class_weight was balanced. Given the dataset's class imbalance, where neutral sentiments significantly outnumber negative ones, we opted to merge neutral with positive sentiments.

Table 3. Performance measures.

Model	Accuracy	Sensitivity (Recall)	Specificity	Balanced Accuracy
KNN 1	0.5210	0.5210	0.3706	0.4458
KNN 3	0.4596	0.4596	0.2294	0.3445
Multinomial Naive Bayes	0.7001	0.7001	0.8839	0.7920
Random Forest	0.7152	0.7152	0.7712	0.6966
Gradient Boosting	0.7050	0.7050	0.8590	0.6526
Logistic Regression	0.7281	0.7281	0.7636	0.7158

This decision was guided by the objective of simplifying the classification task and improving model performance. Balanced accuracy was chosen as the primary metric to ensure that the classifier's performance was fairly evaluated across both classes, particularly in an imbalanced setting.

3 Results

Finally, the performance metrics, including accuracy, balanced accuracy, sensitivity (recall), and specificity, were recorded for each classifier. The results are summarized below: The table 3 above shows the average performance metrics obtained through 10-fold cross-validation for each classifier. Logistic Regression achieved the highest average accuracy (0.7281) but Multinomial Naive Bayes achieved the best balanced accuracy (0.7920), indicating its robustness in handling the class imbalance. Random Forest and Gradient Boosting also performed well, with Gradient Boosting achieving a balanced accuracy of 0.8590. These results provide insights into the effectiveness of different classifiers in predicting user sentiments based on Twitter data.

4 Analysis and Discussion

The results obtained from the classification models provide insightful observations about the performance and applicability of various machine learning algorithms in sentiment analysis of airline tweets. The models implemented were KNN ($k = 1$ and $k = 3$), Multinomial Naive Bayes, Random Forest, Gradient Boosting, and Logistic Regression.

Feature Importance: The feature importance analysis using Random Forest revealed that airline_sentiment_confidence was the most critical feature, followed by airline. Bigram features such as 'flight cancelled flightled' and 'customer service' also ranked highly, indicating their relevance in sentiment classification.

Word Representation Analysis: The representative words for each class highlighted distinct patterns. For the positive class, phrases like "great flight" and "thank much" were prominent, reflecting satisfaction and gratitude. In contrast, the negative class was dominated by terms like "customer service", "cancelled flight", and "late flight", pointing to common issues faced by passengers. To further enrich our analysis, we leveraged the 'tweet_coord' column to map the sentiment geographically.

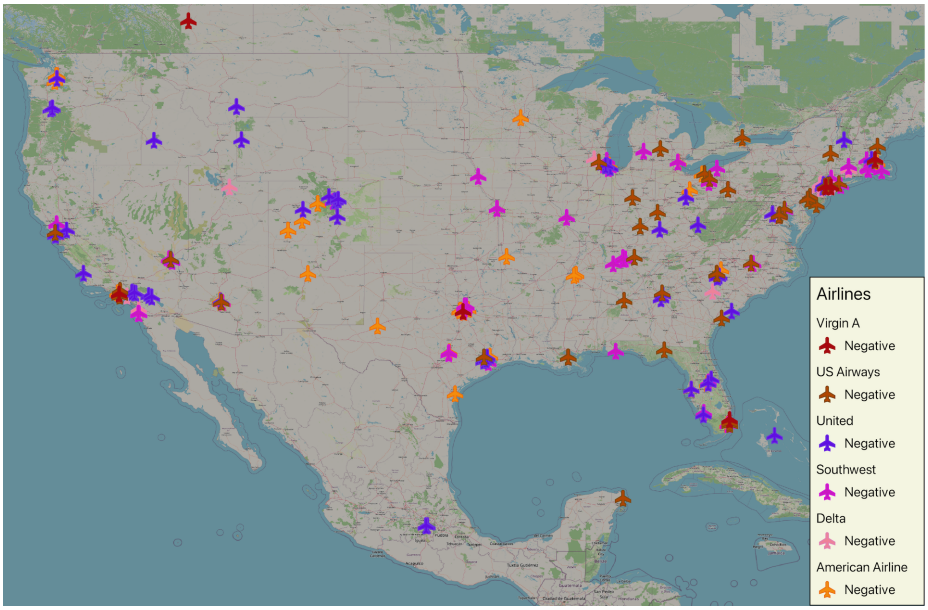

Fig. 1. Positive distribution by airline.

This visualization provides deeper insights into regional sentiment trends and highlights areas with higher concentrations of negative or positive sentiments.

4.1 Negative Sentiments Fig. 1

1. Concentrations of negative sentiments are evident in large metropolitan areas, particularly on the East and West Coasts.
2. Virgin America shows a significant number of negative tweets in Los Angeles and New York, while US Airways has more negative sentiments spread across the Midwest and Northeast.
3. There seems to be a stronger presence of negative sentiments in regions with heavy air traffic.

4.2 Positive Sentiments Fig. 2

1. Positive sentiments are more geographically dispersed, with a noticeable presence in the Midwest and along the East Coast.
2. Virgin America and Southwest Airlines receive positive sentiments in key cities like San Francisco and Dallas.
3. The spatial distribution of positive tweets suggests that positive experiences are more spread out, possibly reflecting regional differences in service quality or customer expectations.

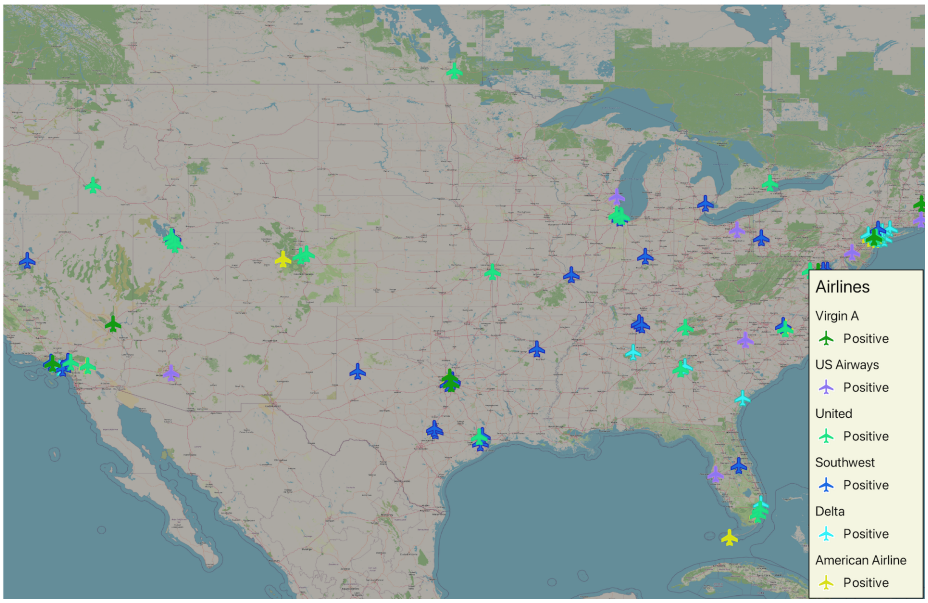

Fig. 2. Positive distribution by airline.

This analysis indicates that airlines could focus on specific regions to improve customer satisfaction, particularly where negative sentiments are concentrated. The geographical patterns could help target marketing efforts or service improvements in those areas.

5 Conclusions

This study demonstrates the effectiveness of various machine learning models in classifying sentiments expressed in airline tweets. Multinomial Naive Bayes emerged as the best-performing model, closely followed by Gradient Boosting. The incorporation of feature engineering techniques, such as bigrams and TF-IDF vectorization, significantly contributed to the models' performance. Addressing the class imbalance through techniques like balanced class weighting combining the neutral and positive class and using balanced accuracy as a performance metric provided a more nuanced understanding of the models' capabilities. Future research could explore several avenues to enhance the current work:

1. **Incorporating Deep Learning Models:** Utilizing advanced deep learning architectures like LSTM and BERT for sentiment analysis might improve accuracy and capture more complex patterns in the text data.
2. **Enhancing Geospatial Analysis:** Integrating detailed geospatial analysis by mapping tweets to specific locations can provide deeper insights into regional sentiment trends and potentially uncover regional-specific issues.

3. **Temporal Analysis:** Adding a temporal component to analyze how sentiments evolve over time could help identify patterns related to specific events or seasons.
4. **Handling Neutral Sentiments:** Developing more sophisticated techniques to handle and differentiate neutral sentiments could provide a clearer picture of customer feedback.
5. **Real-time Sentiment Analysis:** Implementing a real-time sentiment analysis system could help airlines respond more promptly to customer feedback and improve service quality dynamically.
6. **Updated geospatial data analysis:** Add a new dataset updated and analyze where the sentiments changed and why.

In summary, the study lays a solid foundation for sentiment analysis in the airline industry, demonstrating the potential of machine learning models in deriving actionable insights from social media data. Further advancements in this field can significantly enhance customer experience and operational efficiency for airlines.

References

1. Aggarwal, C. C., Zhai, C.: A survey of text classification algorithms. Mining Text Data, pp. 163–222 (2012) doi: 10.1007/978-1-4614-3223-4_6
2. Breiman, L.: Random forests. Machine Learning, vol. 45, no. 1, pp. 5–32 (2001) doi: 10.1023/a:1010933404324
3. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27 (1967) doi: 10.1109/tit.1967.1053964
4. Cox, D. R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 20, no. 2, pp. 215–232 (1958) doi: 10.1111/j.2517-6161.1958.tb00292.x
5. Friedman, J. H.: Greedy function approximation: a gradient boosting machine. The Annals of Statistics, vol. 29, no. 5 (2001) doi: 10.1214/aos/1013203451
6. Gauch, J. M., Gauch, S., Bouix, S., Zhu, X.: Real time video scene detection and classification. Information Processing and Management, vol. 35, no. 3, pp. 381–400 (1999) doi: 10.1016/s0306-4573(98)00067-3
7. Jurafsky, D., Martin, J. H.: Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models. Pearson (2024)
8. Kaggle: Twitter US Airline Sentiment (2019) www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
9. Manning, C. D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008) doi: 10.1017/cbo9780511809071
10. Provost, F. J., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 445–453 (1998)
11. Ramos, J.: Using TF-IDF to determine word relevance in document queries. In: Proceedings of the First Instructional Conference on Machine Learning, pp. 1–4 (2003)
12. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management, vol. 24, no. 5, pp. 513–523 (1988) doi: 10.1016/0306-4573(88)90021-0