Sentiment Classification and Topic Modeling for Airline Review Tweets

Meichan Huang, Ryan Tervo
School of Information Studies
Syracuse University
Syracuse, NY, 13210
mhuang01@syr.edu; rtervo@syr.edu

Abstract

This project analyzes a corpus of 14,640 tweets to gauge public sentiment toward six major airlines in the U.S. Through the use of sentiment classification and Latent Dirichlet Allocation (LDA) topic modeling, the project aims to develop an algorithm that can accurately classify sentiment in tweets automatically. Additionally, the project investigates the topic patterns of positive and negative tweets to gain insight into customers' perceptions of airlines. By leveraging machine learning techniques, this study seeks to provide a deeper understanding of the factors that drive sentiment in the airline industry to gain insights into customer preferences and opinions.

1. Introduction

This study seeks to help airlines identify customer frustrations in the travel experience. Two separate, but related, modeling techniques will be used. First, sentiment analysis will be used to identify tweets with negative sentiment.

The second is topic modeling. Topic modeling will be used to identify specific areas that customers are unhappy with. If successful, this will automate an otherwise labor-intensive function and enable airlines to continuously monitor public sentiment towards their travel experience.

A plethora of literature has been published on the sentiment classification system of Twitter data for US airline services analysis using different algorithms and feature engineering (e.g. Rane & Kumar, 2018; Wan & Gao, 2015), and fairly recently, researchers started to tap into the possibility of using topic modeling for online review for airlines (Korfiatis, et al., 2019; Kwon et al., 2021). However, very little research has discussed the potential benefits of combining sentiment analysis and topic modeling in the airline service reviews. Therefore, in this project, we proposed the following three research objectives: (1) What is the best algorithm for predicting sentiment of tweets in the airline industry? (2) What frequent topics are associated with positive sentiment vs. negative sentiment? (3) Do six airline companies receive similar or different topics in travelers' negative comments?

2. Method

2.1 Data Set

The dataset used in this project was obtained from Kaggle.com and comprises 14,640 tweets scraped in February 2015. It focuses on the sentiment of travelers toward six major airlines during that time (Figure Eight, 2015). The data is in csv. format, with 15 columns, containing information regarding the tweets (ID, time, user, re_tweets), the tweet text with all the handles and hashtags, and URL links. The data has also been pre-labeled with sentiments (positive, negative, and neutral) by human annotators, who also categorized negative comments based on reasons such as "late flight" or "rude service".

It is worth noting that the dataset has an unbalanced distribution of data points across the different sentiment classes (Figure. 1). Specifically, there are 9,178 negative comments, 3,099 neutral comments, and 2,363 positive comments.

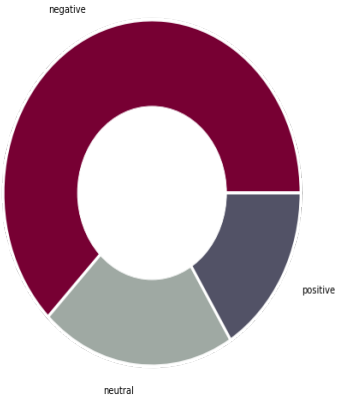

Figure 1. *Number of sentiments in each class*

Moreover, the number of comments received by each airline varies (Figure. 2), with United Airlines receiving over 3500 reviews and Virgin America receiving approx. 500 reviews.

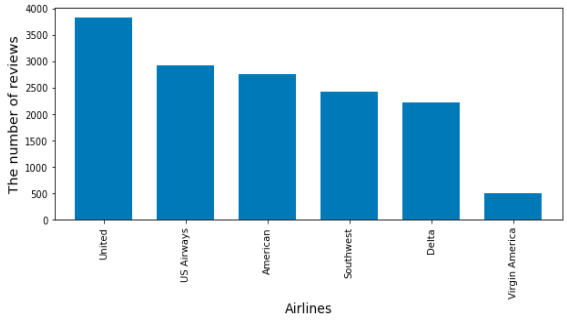

Figure 2. *The number of reviews by airlines*

2.2 Data Preprocessing

After downloading the CSV file, we loaded the dataset into a pandas data frame for preprocessing as our project is focused on sentiment classification and topic modeling, which requires the retention of the text and its respective sentiment labels. Since Twitter data is often inconsistent and has some features that could interfere with the machine's processing of the text, we applied several pre-processing techniques to clean the tweets, such as using Regex to remove potential noise such as URLs, punctuations, and special characters (emojis). Additionally, we applied lowercase conversion and WordNet stemming to remove capitalization and lemmatize the tokens. Stopwords were not removed in this process, considering that they are important in the formation of N-grams. Instead, we utilized the vectorizers' parameter tuning to remove default English stopwords and a list of customized stopwords, including the airline handles (i.e. @southwestair, @usairways).

After pre-processing, we further processed the data by vectorizing it, as both classification algorithms and LDA topic modeling require a document-word matrix as the primary input. We experimented with several vectorizers for the classification tasks, testing and tuning each to determine the best performance. The classifiers were trained using 80% of the data and tested using the remaining 20%. For the LDA modeling, CountVectorizer was used since LDA is based on term count and document count.

2.3 Classification Algorithms

We employed two distinct types of algorithms in order to determine the optimal classification method. Naive Bayes, which is a probabilistic algorithm commonly used for text classification, was one of the algorithms we used. Naive Bayes is known for its simplicity, speed, and ability to operate with relatively little training data. It is also well-suited for handling high-dimensional data and can handle missing values. Additionally, we employed Support Vector Machines (SVMs), which are effective for handling high-dimensional data, especially when the data is linearly separable. SVMs are capable of handling both categorical and continuous data and are ideal for small to medium-sized datasets.

Specifically, six model configurations were investigated to determine the best method to identify negative tweets. The Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) algorithms were used in combination with three vectorizers. To fit these two algorithms, the data were vectorized using two different Term-frequency Counts (CountVectorizer) settings and a term frequency-inverse document frequency (TF-IDF) Vectorizer with parameters fine-tuned to various degrees, i.e. min_df, max_df, stopwords removal, and so on.

2.3.1 Classification Vectorizer Configuration

The classification models attempt to correctly categorize tweets as either 'positive', 'negative', and 'neutral'. The word vectorizer's parameters were modified to maximize the model's performance. In order to evaluate the model's performance the model's accuracy, precision, recall, F-1 scores, and the top 10 positive and negative words were used. Using trial and error the following settings appeared to provide the best overall performance in terms of performance and top 10 negative and positive words.

Table 1. *Vectorizer 1 Settings*

Parameters	Setting
encoding	'latin-1'
binary	False
min_df	3
max_df	1500
ngram_range	(1, 5)

Table 2. *Vectorizer 2 Settings*

Parameters	Setting
encoding	'latin-1'
binary	False
min_df	2
max_df	1500
ngram_range	(1, 3)

Table 3 *Vectorizer 3 Settings*

Parameters	Setting
encoding	'latin-1'
binary	False
min_df	5
smooth_idf	True
sublinear_tf	True
ngram_range	(1, 4)
max_features	2000

2.3.2 Classification Model Interpretation

The sentiment classification algorithms are designed to predict the tweet's sentiment as either 'negative', 'positive', or 'neutral'. This is accomplished by first vectorizing the tweets. The model is then trained to associate those vectors (features) with their respective labels. The trained models are then evaluated using test data to determine how well they can predict the correct label.

In this use case both the overall model accuracy and specifically how well the model can predict 'negative' tweets are considered the most important. Due to the importance of being able to identify one particular label ('negative') the specific recall, precision, and F1 score will be used to evaluate the model. These scores allow us to compare multiple model's performance.

2.4 Topic Modeling Toolkit

In order to conduct the evaluation of the topics in customers' opinions on airlines, this project used Latent Dirichlet Allocation (LDA) (Blei et al., 2003) for topic discovery. LDA is a statistical modeling technique that helps identify the hidden topics present in a given corpus or dataset. It posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. The project applied Scikit-learn's built-in LDA module (Pedregosa *et al.*, 2011) for text processing, document classification, clustering, and topic modeling.

2.4.1 Topic Modeling Interpretation

We conducted topic modeling using the LDA model using sklearn.de To determine the optimal number of topics that best encompassed the information in the corpus, we tested several different numbers of topics, including K20, K10, K5, K4, K3, and K2, each with a max_iter of 10. We computed the log-likelihood scores and perplexity scores for each model to evaluate its performance. Additionally, we conducted model comparisons using GridSearch to determine the best estimator for the number of components.

2.4.2 Human Judgment on Topic Coherence

The outputs of the LDA topic modeling were used to examine the differences in themes or topics. To assess topic cohesion and meaningfulness, we conducted a human judgment evaluation through the composition of the top 10 words in the topic clustering outputs, as well as the comparison of LDAVis outputs for each topic modeling, following the suggestion of Chang et al. (2013). Our evaluation served as a triangulation method for the statistical modeling outputs in sklearn, such as log-likelihood and perplexity scores. As Kapadia (2019) suggested, "optimizing for perplexity may not yield human interpretable topics", as is the case we found out later in the topic modeling process.

3 Results

In the following section, we discuss our final results with the classification and topic modeling. For more specific results, please see the attached html output of our codes.

3.1 Classification

The six model configurations were evaluated using a ten-fold cross validation. Given the focus on 'negative' tweets, we chose to report the 10-cv accuracy scores, recall, precision, and F1 scores that were based on the 'negative' category only. In Table 4 the models were evaluated using the data labeled as either 'negative', 'positive', and 'neutral'.

The six models were evaluated a second time in which the category labels were modified to be either 'negative' and 'notNegative'. The 'notNegative' category was the combination of 'positive' and 'neutral' tweets. Table 5 shows the results in which the precision, recall, and F1 scores that are based on the 'negative' category only.

Table 4 *Model Results using 3 Categories*

Model	Vectorizer	Accuracy	Precision	Recall	F1-Score
MNB	Count 1	75%	80%	88%	84%
MNB	Count 2	75%	79%	88%	84%
MNB	TF-IDF	74%	74%	96%	83%
SVM	Count 1	73%	82%	83%	83%
SVM	Count 2	74%	82%	84%	83%
SVM	TF-IDF	76%	81%	89%	85%

Table 5. *Model Results using 2 Categories*

Model	Vectorizer	Accuracy	Precision	Recall	F1-Score
MNB	Count 1	79%	81%	87%	84%
MNB	Count 2	80%	82%	87%	84%
MNB	TF-IDF	79%	78%	92%	85%
SVM	Count 1	78%	82%	82%	82%
SVM	Count 2	78%	83%	83%	83%
SVM	TF-IDF	80%	82%	86%	84%

All of the model configurations, using the data categorized using two or three labels, performed better than the trivial model. The best accuracy was achieved using the SVM model with the TF-IDF vectorizer which improved the accuracy of the trivial model by 13.1%. The trivial (no information) model would predict every tweet 'negative' and achieve an accuracy of only 62.9%.

In evaluating the SVM model using the TF-IDF vectorizer the F1 score was 85% (original categories) and 84% (two categories). The F1 score is of particular interest because it is the harmonic mean between the recall and precision scores.

3.2 Topic modeling

This report presents the results of topic modeling on positive and negative comments received by airline companies. Our findings show that positive tweets about airlines tend to be more general in nature, while negative comments can be classified into three major topics.

We also analyzed whether companies received negative feedback for similar or different reasons. Our results indicate that while the words associated with different topics varied across different airlines, there were some common trends in the negative aspects of airlines that are frequently reported.

For more details about the log likelihood scores and perplexity for all the modeling in this project, please refer to the attached html output of the coding and outputs.

3.2.1 Topics Distribution for Positive Tweets

The analysis indicates that travelers tend to provide a general overview of their positive experiences. Figure x illustrates that keywords associated with positive experiences include smooth flights and positive feedback regarding crew/attendant services.

Figure 3. *WordCloud on all positive comments*

Although the model with n_components = 2, when compared to other numbers of components, has the highest log likelihood (-93538.20085373441) and the lowest perplexity score (723.0571290935178), and indicated the better model fit for the data theoretically (citation needed), 3 topic clusters actually worked the best with topic coherence and topic interpretability, when human judgments were factored into for model evaluation. Based on the output of top 10 keywords in topic modeling, when the corpus was reduced to 3 topics, it encompasses a distinct set of topics without obvious overlapping.

Table 6. *Top 10 keywords with 3 topics in positive comments*

Topic	Top 10 words
1	flight, great, thanks, crew, time, gate, response, plane, home, good
2	service, thanks, customer, customer service, help, good, flight, today, thx, amp
3	thank, thanks, just, love, guy, airline, best, fly, yes, flying

Upon reviewing the LDAvis results, it becomes apparent that analyzing the top 30 words associated with each topic provides greater insight into the underlying themes. In general, the feedback from customers is positive regarding their experiences with airline customer service and flight experiences. Within the flight experience category, the quality of attendant/crew service and a seamless travel experience stand out as key themes within the dataset. It appears that positive feedback from travelers is primarily focused on the services provided by airlines and a smooth overall experience, while negative feedback is less centered on these aspects.

3.2.2 Topic Distribution for Negative Tweets

When examining the topics associated with negative tweets, three major themes emerge: (1) punctuality issues such as cancelled or delayed flights; (2) poor customer service characterized by rude agents and slow response times; and (3) problematic baggage handling leading to lost or damaged luggage. Even if the top 10 words for each topic varied slightly in different companies, the general themes seemed to align with the general negative comments received across the airline companies, as shown in Table 7:

Table 7. *Top 10 keywords with 3 topics in negative comments*

Topic	Top 10 words
1	time, hold, hour, agent, airline, aa, minute, worst, phone, wait
2	hour, plane, delayed, bag, late, delay, waiting, gate, hr, time
3	cancelled, customer, service, flightled, help, cancelled flightled, customer service, day, need, amp

When deciding which was the best estimator, n_component = 3, again, although this model did not have the highest log likelihood and lowest perplexity score compared to other models, there was no overlapping amongst topics, based on the intertopic distance map using LDAvis (see Figure. 4). All topics are semantic meaningful.

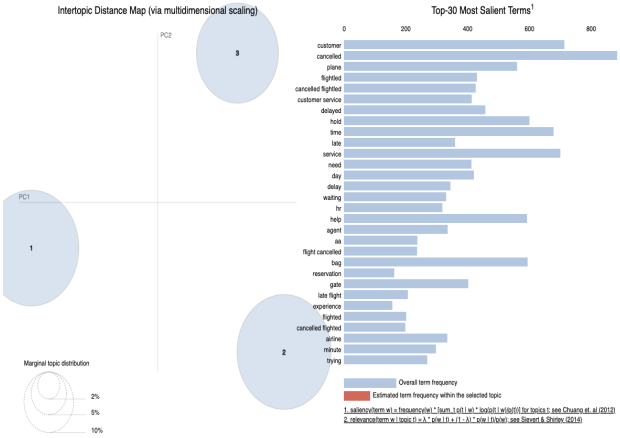

Figure 4. *LADvis output for negative comments*

In addition, the topic modeling effectively aligns with the overall themes identified in the human annotation of the original dataset's negative reviews, apart from the tweets that even human annotaters cannot tell, see Figure 5 below:

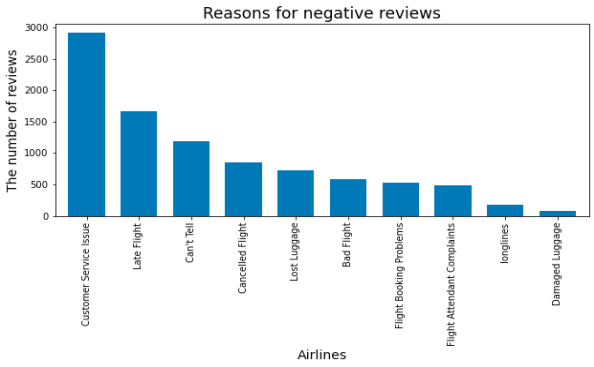

Figure 5. *Reasons for negative reviews assigned by human annotators in the original dataset*

3.2.3 Most Frequent Topics for Negative Tweets by Airline

Despite the variations in comments and word choices received by each airline, it appears that these differences do not significantly affect the overall topic composition of negative feedback expressed by travelers. In general, the identified topics appear to center on the same three

problematic areas of airline services: (1) customer service; (2) flight punctuality issues such as cancellations or delays; and (3) luggage handling. However, some minor variations were also noted, including issues with check-in services and email communication with customers, as shown in Table 8.

Table 8. *Comparison of Topic Keywords across different airlines (n_component = 3)*

Topics	United	US Airway	AA	Southwest	Delta	VA
1	service, customer, customer service, thanks, like, just, help, know, need, did	flight, hold, hour, time, service, customer, help, delayed, need, minute	flight, cancelled, did, ca, seat, help, just, flighted, cancelled flighted, hold	flight, cancelled, cancelled flightled, flightled, hold, flight cancelled, hr, flight cancelled flightled, time, cancelled flightled flight	just, need, help, right, fleek, fleet, fleet fleek, ca, blue, work	flight, cancelled, check, customer, jfk, virgin, today, airline, vx, ca
2	flight, bag, plane, cancelled, hour, gate, time, amp, seat, late	day, bag, customer, service, luggage, phone, people, doe, say, customer service	service, hour, customer, agent, customer service, gate, time, bag, phone, hold	hour, hold, flight, phone, help, ca, online, trying, hold hour, just	flight, delayed, time, service, customer, amp, cancelled, customer service, hour, min	email, website, site, doe, day, checkin, contact, iphone, problem, applied
3	flight, delay, hour, airline, delayed, ua, time, luggage, crew, worst	flight, cancelled, plane, flightled, cancelled flightled, delay, mile, sitting, flight cancelled, hour	flight, cancelled, flightled, cancelled flightled, hour, aa, late, dfw, flight cancelled, delayed	customer, service, customer service, bag, hour, flight, plane, gate, late, seat	flight, delay, jfk, plane, hour, bag, just, gate, hr, time	eat, help, flight, trying, time, luggage, ticket, just, hour, book

The following Figures 6. - 11. visually display the word clouds generated from negative tweets received by each airline. These graphics provide compelling evidence that certain topic themes are commonly shared across all airlines when it comes to the negative experiences reported by passengers.

Figure 6. *WordCloud on United's negative comments*

Figure 7. *WordCloud on US Airway's negative comments*

Figure 8. *WordCloud on American Airlines' negative comments*

Figure 9. *WordCloud on Southwest' negative comments*

Figure 10. *WordCloud on Delta's negative comments*

Figure 11. *WordCloud on Virgin America's negative comments*

5. Discussion and Conclusions

The objective of this project was to assess the public sentiment towards six major airlines in the U.S. using a corpus of 14,640 tweets and to understand the topics that generate these sentiments. To achieve this goal, sentiment

classification and Latent Dirichlet Allocation (LDA) topic modeling were employed to develop an algorithm capable of accurately classifying sentiment in tweets, and exploring topics the topic patterns of both positive and negative tweets, providing valuable insights into customers' perceptions of airlines.

The sentiment analysis showed that SVM using the TF-IDF vectorizer performed best. The model achieved an accuracy of 76% using the three label categories. Evaluating the data using only two labels, 'negative' and 'notNegative', the accuracy improved to 80%. In addition, with the emphasis on 'negative' tweets the F1 score was used to determine how well the model is identifying 'negative' tweets. Achieving an F1 score of 84% shows that the model is performing very well in identifying 'negative' tweets.

The results of the topic modeling suggest that when discussing their positive experiences with airlines, travelers tend to provide less specific information. This result is not unexpected since people tend to express their compliments less frequently and may not always provide specific reasons for positive experiences unless they are particularly outstanding, such as exceptional service or an extremely friendly flight attendant.

However, negative comments align well with common real-life issues we experience with airlines, including flight cancellations or delays, dissatisfaction with customer service during the booking process, poor service provided by flight attendants, and lost or damaged luggage.The fact that all airlines received similar complaints in these areas suggests that airlines can use sentiment analysis and topic modeling of online reviews as a tool to identify common customer preferences and concerns. By scraping all online reviews, airlines can gain valuable insights into what their customers like and dislike about the services they provide, helping them to improve the customer experience and stay competitive in the industry.

The results also has meaningful applications in the airline industry. The results should be considered as a successful attempt in trying to automate the classification process of public sentiment and gain insights into these sentiments. That is, first, classifying large volumes of online reviews can be a time-consuming task. Applying classification algorithms can automate the process, saving time and increasing efficiency. Applying classification algorithms can ensure a consistent approach to analyzing online reviews, reducing the risk of bias and errors that may arise from manual analysis.

In addition, measuring the quality of services is a challenging task because different customers may have varying preferences. However, by extracting negative comments from airline passengers' online reviews and performing topic modeling, companies can gain an understanding in areas with shared negative customer experience, where they need to improve their services and enhance the customer experience, leading to more satisfied customers and increased revenue. The findings suggest that using quality features extracted from tweets can help predict variations in passenger preferences and competition, indicating their potential usefulness in evaluating airline service quality.

However current project also has its limitations, which will constrain the generation of the above results. First, the performance of the models might be far from perfect. For classification models, there are indeed better models suggested in previous literature, e.g. neural network models (RNN) and AdaBoost (Rane & Kumar, 2018), a technique in which multiple classifiers were combined; however, due to our limitation, we were not able to explore the possibility these models can provide.

Second, regarding the topic modeling, even if the current project has used both statistic outputs and human judgment for topic modeling, the performance of the topic modeling still needs further examination, since the determination of the number of topics can be arbitrary and depends on the researchers' ability to interpret the resulting topic solution.

Lastly, sklearn libraries provided different LDA modeling tools, with gensim being one of them. We tried the Gensim module and modeled the positive and negative comments with n_components = (5, 4, 3) respectively and the results slightly different but eventually, since we are more familiar with sklearn, we opted to use sklearn for our analysis. One advantage of using gensim is that Gensim's LDA has more built in functionality and applications for the LDA model such as Topic Coherence Pipeline or Dynamic Topic Modeling, in addition, the built-in parameter of coherence scores would help adding

more statistical measures that human judgment or log likelihood scores.

References

[1] Blei, D. et al., (2003) Latent dirichlet allocation. *Journal of machine learning research*,(4-5):993-1022.

[2] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems,* 22.

[3] Wan, Y., & Gao, Q. (2015, November). An ensemble sentiment classification system of twitter data for airline services analysis. In *2015 IEEE international conference on data mining workshop (ICDMW)* (pp. 1318-1325). IEEE.

[4] Kapadia, S. (2019). Evaluate Topic Models: Latent Dirichlet Allocation (LDA): A step-by-step guide to building interpretable topic models. Retrieved from: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[5] Korfiatis, N., Stamolampros, P., Kourouthanassis, P., & Sagiadinos, V. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications, 116,* 472-486.

[6] Kwon, H. J., Ban, H. J., Jun, J. K., & Kim, H. S. (2021). Topic modeling and sentiment analysis of online review for airlines. *Information,* 12(2), 78.

[7] 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.

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research,* 12, 2825-2830.