Sentiment and Complaint Analysis of Airline Tweets via Natural Language Processing

Aidan Shin

Troy High School, Fullerton, CA

Abstract

Airline customers often share their experiences on social media, from glowing reviews to complaints about delays, cancellations, or lost luggage. In this study, we analyze a collection of airline-related tweets labeled by sentiment and, for negative tweets, by the specific reason for dissatisfaction. We explore the data with word clouds, train a DistilBERT model to classify sentiment and negative-reason categories, and use a multinomial logit model to highlight words most associated with positive or negative feedback.

Keywords: airline customer feedback, Twitter, sentiment analysis, natural language processing, distilBERT, word cloud, multinomial logit model

1 Introduction

Air travel has become an essential component of global mobility, connecting people and markets across the world. At the same time, the airline industry faces intense scrutiny from consumers, who increasingly share their travel experiences on social media platforms such as Twitter. These posts offer an unfiltered glimpse into customer perceptions, ranging from praise for exceptional service to frustration over delays, cancellations, and lost baggage. For airlines, understanding the content and tone of such feedback is vital for improving service quality, brand reputation, and customer loyalty. For researchers, the availability of large-scale labeled text data creates an opportunity to explore and advance methods in natural language processing (NLP).

This paper presents a multi-faceted analysis of airline-related tweets with sentiment and reason annotations. The dataset categorizes tweets into three sentiment classes—negative, neutral, and positive—with negative tweets further classified by the underlying cause of dissatisfaction, such as flight delays, cancellations, customer service issues, or baggage problems. Such structured labeling allows us not only to study general sentiment but also to uncover the specific pain points that drive customer dissatisfaction.

Our analysis proceeds in two stages. First, we employ DistilBERT, a lightweight transformer-based model, to perform sentiment classification and to predict the reason categories for negative tweets. This allows us to evaluate the performance of state-of-the-art deep learning models on sentiment and fine-grained classification tasks. Second, we apply a multinomial logit model to identify the words most strongly associated with positive or negative sentiment, providing interpretable statistical evidence that complements the deep learning results.

2 Literature Review

In recent years, the analysis of airline customer sentiments through social media platforms, particularly Twitter, has gained significant attention due to its potential for real-time feedback and service improvement. Kumar and Zymbler (2019) conducted one of the earlier studies in this domain, where they trained machine learning models on airline-related tweets to predict sentiment. Their approach demonstrated that automated analysis could provide a reliable gauge of customer satisfaction, emphasizing the utility of machine learning in capturing nuanced opinions expressed in natural language [1].

Following this, Wu and Gao (2022) extended the work by analyzing tweets to detect irregularities in passenger sentiment, highlighting the predictive potential of sentiment analysis for operational and customer service insights. Their work underscored the importance of temporal patterns in sentiment trends, showing that spikes in negative sentiment could correlate with service disruptions or delays [2].

Tusar and Islam (2021) contributed a comparative study of various classification algorithms on airline tweets, using natural language processing techniques to evaluate model performance. Their findings indicated that algorithm selection plays a critical role in ac-

curately classifying sentiments, suggesting that hybrid approaches might be necessary for enhanced performance [3].

In a more recent study, Shitole and Vaidya (2023) compared multiple machine learning algorithms for tweet classification, concluding that Support Vector Machines (SVM) achieved the highest accuracy among the methods tested. This highlights SVM's robustness in dealing with high-dimensional text data typical in social media platforms [4].

Hasan and Fattah (2024) introduced an innovative approach by incorporating tweet metadata alongside textual content. By leveraging features such as posting time, user influence, and engagement metrics, their model improved predictive performance, demonstrating that non-textual data can provide valuable context for sentiment analysis [5].

Rusta et al. (2019) explored the effect of different feature extraction methods, including term frequency (TF), term frequency-inverse document frequency (TF-IDF), and word embeddings (word2vec). Their comparative analysis suggested that while traditional TF-IDF features perform well, word2vec embeddings can capture semantic relationships between words, potentially enhancing sentiment classification accuracy [6].

Chaudhary and Lohiya (2024) provided a comprehensive review of methodologies and challenges in applying machine learning for airline tweet sentiment analysis. They summarized the key motivations, ranging from customer satisfaction monitoring to operational decision support, and highlighted best practices in model selection, feature engineering, and evaluation metrics [7].

Finally, Panda and Sheetlani (2019) performed a literature review across multiple studies, concluding that no single machine learning method consistently outperforms others across datasets. Their work underscores the need for context-specific model tuning and the exploration of ensemble methods to achieve reliable performance [8].

Overall, the literature illustrates a progressive refinement in techniques, from basic sentiment classification to multi-faceted approaches integrating metadata and advanced feature representations. Despite the advancements, challenges remain, particularly in handling sarcasm, multilingual tweets, and domain-specific jargon, which continue to motivate

research in this field.

3 Data Description

The airline tweets dataset was obtained from Kaggle ("Twitter US Airline Sentiment" dataset), using this link: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment. Originally collected in February 2015 through a crowdsourcing platform, it contains 14,640 tweets about several U.S. airlines. Each tweet is labeled with its sentiment (negative, neutral, or positive), and for negative tweets, an additional column specifies the reason for the negativity, categorized into one of nine possible reasons. Below are sample tweets illustrating each sentiment category, as well as each reason associated with negative tweets. All tweets are presented with their original grammar and orthography preserved; however, any reference to specific airlines has been redacted (replaced by a sequence of `x` characters).

Examples of Positive Tweets

`@xxxxx yes, nearly every time I fly xxx this ‘‘ear worm’’ won’t go away :)`

`@xxxxx This is such a great deal! Already thinking about my 2nd trip to @Australia & I haven’t even gone on my 1st trip yet! ;p`

`@xxxxx Your crew on 4028 tonight was outstanding. God bless them and the medically trained passengers on board.`

Examples of Neutral Tweets

`@xxxxx will you be making BOS to LAS nonstop permanently anytime soon?`

`@xxxxx Just submitted my response on the link you sent.`

`@xxxxx Is there any way to get entry to the Las Vegas event to see @Imaginedragons perform? #DestinationDragons`

Examples of Negative Tweets

@xxxxx Dear xxx, I fly you a lot. Most of the time, amazing. Today: unacceptable.I am sitting here in Midway delayed 20 min....

@xxxxx Been standing at the gate for 45 min trying to go standby bc I will miss my connection. No help! Do NOT fly xxxx!

@xxxxx I have been checking consistently, and called multiple times. As a loyal xxx customer, I'm disappointed.

Examples of Negative Tweets for Each Reason

Bad Flight: @xxxxx it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse.

Late Flight: @xxxxx See? We were told repeatedly that the pilot was Late Flight and kept getting Late Flightr. After we boarded, there was a defibrillator issue.

Customer Service Issue: @xxxxx been on hold over an hr to rebook a Cancelled Flighted flight. Do you have anyone working???

Flight Booking Problems: @xxxxx why won't the site let me book tickets for nov for jfk to kin?

Lost Luggage: @xxxxx Your Baggage system has hung up on me twice because you have too many callers. I NEED TO FIND MY HUSBAND'S (@SweetingR) BAGS.

Flight Attendant Complaints: "@xxxxx @cnnbrk she tried they are not doing anything said they would talk to stewardess about serving drunks drinks how does that help

Long Lines: @xxxxx One hour to check in is 45 minutes too long. MIA FAIL.

Cancelled Flight: @xxxxx I tried that. You offered to charge me an additional $1k for a new ticket or be stranded until Thurs. 1st time, last time.

Damaged Luggage: "@xxxxx Had to spend 5 hours worrying that items in carryon would be broken/stolen since I couldn't carry them on plane or lock bag.

Below, we present a word cloud of the 100 most frequently used words in these tweets (see Figure 1). Apart from the company names themselves, words such as "flight, get, customer, cancelled, now, can, service, thanks, help" appear most frequently. The association of these words with positive or negative sentiments will be examined later in this paper.

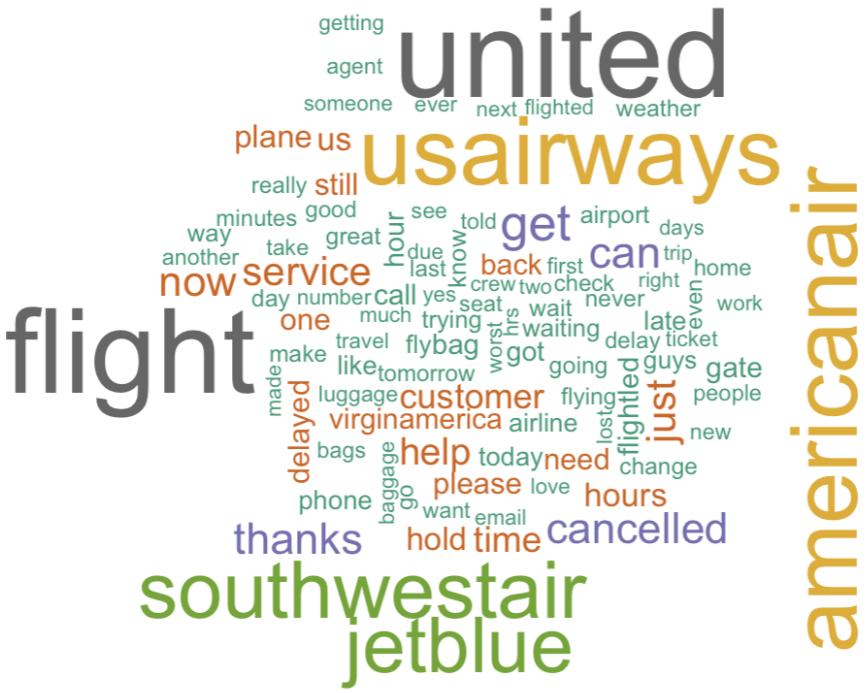

Figure 1: Word cloud for the 100 most frequently used words in the tweets.

4 Bidirectional Encoder Representations from Transformers (BERT) Model

4.1 Model Definition

The Bidirectional Encoder Representations from Transformers (BERT) model is a language model that learns rich contextual representations of text by examining how each word relates to every other word in a sentence, capturing meaning from both directions at once. At its core, a language model is a statistical tool that assigns probabilities to sequences of words, with the goal of capturing meaning and structure in text. BERT works by first

converting each word in a sentence into a vector representation, often called an *embedding*. These embeddings capture semantic information, such as words with similar meaning being represented by nearby vectors in the embedding space. Since word order also matters in language, BERT adds *positional encodings* to the embeddings, which are patterns of numbers that provide the model with information about the position of each token in the sequence. The resulting input vectors are then processed through a series of layers that use a mechanism called *self-attention*. Self-attention allows the model to compare each word in the sequence to every other word and decide which ones are most relevant for interpreting its meaning.

Training BERT involves teaching it to predict missing words in a sentence, a task known as *masked language modeling*. In this task, certain words are hidden (or "masked"), and the model learns to guess them based on the surrounding context. Another important aspect of training is learning from probability distributions rather than only from hard labels. Instead of simply being told which word is correct, the model is also guided by probability patterns that describe how likely alternative words are in the same position. This richer signal helps the model to capture subtle patterns in language. Mathematically, this guiding process is often expressed using the Kullback–Leibler (KL) divergence, which measures the difference between two probability distributions. If p is the target probability distribution and q is the model's predicted distribution, the divergence is written as

$$\mathrm{KL}(p, q) \;=\; \sum_i p_i \, \log\!\left(\frac{p_i}{q_i}\right),$$

which encourages the model to produce probabilities close to the target. Through this combination of embeddings, positional information, self-attention, and training based on masked language modeling and distributional learning, BERT acquires the ability to create rich contextual representations of text that can be used for tasks such as classification, sentiment analysis, and question answering.

4.2 Measures of Performance for Multinomial Classifiers

For multinomial classification problems, model performance is typically evaluated using both class-specific and aggregated metrics. Class-specific metrics—precision, recall, and F1-score—are defined in terms of true positives (TP, correctly predicted instances of a class), false positives (FP, instances incorrectly predicted as that class), and false negatives (FN, instances of that class incorrectly predicted as another). Because these metrics vary

across classes, they are often summarized using macro and weighted macro averages, while overall accuracy provides an aggregate measure of correct classification across all classes. For each class,

$$\text{Precision} = \frac{TP}{TP + FP},$$

that is, of all those predicted as that class, what proportion is predicted correctly?

$$\text{Recall} = \frac{TP}{TP + FN},$$

that is, of all true instances of that class, what proportion is predicted correctly?

$$F1\text{-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

that is, the F1-score is the harmonic mean of precision and recall. Macro metrics for precision, recall, and F1-score are defined as the arithmetic mean of the respective class-specific metrics over all classes. Weighted macro metrics are the class-specific metrics averaged with weights proportional to class support (the number of true instances in each class).

4.3 Application to Airline Tweets Data

We first split the data into training and testing sets, allocating 70% for training and 30% for testing, while stratifying by sentiment (negative, neutral, or positive). The BERT model is then trained on the training set (in practice, we use Python to fit its more efficient successor, DistilBERT), and the testing set is used to evaluate its sentiment predictions. The trained model can also be applied to custom sentences to observe how they are classified.
Next, we focus on tweets labeled as negative. We create a subset containing only these tweets, split it again into 70% training and 30% testing sets, and train the model to predict the reason for negative sentiment. As before, the fitted model can be tested in custom sentences to see which reason for the negative sentiment it predicts.

4.4 Results

Tables 1 and 2 below contain the confusion matrix and performance measures for the first trained BERT model that predicts the sentiment of the tweets as negative, neutral, or positive.

Table 1. The Confusion Matrix for the Sentiment Model.

		PREDICTED		
		Negative	Neutral	Positive
ACTUAL	Negative	2586	163	65
	Neutral	219	585	80
	Positive	57	79	558

Table 2. Performance Measures for the Sentiment Model.

	Precision	Recall	F1-score	Support
Negative	0.90	0.92	0.91	2814
Neutral	0.71	0.66	0.68	884
Positive	0.79	0.80	0.80	694
Macro Average	0.80	0.79	0.85	4392
Weighted Macro Average	0.85	0.85	0.85	4392
Overall Accuracy	0.85			4392

The model achieved an overall accuracy of 85%. Its best performance was on the negative class, with 90% precision and 92% recall, and its weakest performance was on the neutral class, with 71% precision and 66% recall. This disparity may stem from negative tweets containing more distinct sentiment-related keywords, whereas neutral tweets tend to lack clear indicators.

The trained model is then tested on custom sentences. Instead of producing a single label, it assigns a score to each category, and the category with the highest score is taken as the prediction. Below, we present results for two sentences in each sentiment category, one that is classified correctly and another that is misclassified.

The first sentence was correctly classified, but the second sentence was incorrectly classified as positive when it is negative. This is likely due to the presence of words the model associates with being positive, like "awesome" and "thanks", even though the overall sentiment is actually negative.

Results

Negative: "The customer service was horrendous and didn't listen to me, I demand my money back!": `[ 5.1386557, -2.089425, -2.7402763] -> negative`

Negative: "The flight was not awesome, thanks but no thanks." `[-1.1340543, -2.1395345, 3.2384398] -> positive`

Neutral: "When will my flight arrive?": `[-1.5347573, 3.5893867, -2.50458] -> neutral`

Neutral: "Is it true that the Frequent Flyers program will be discontinued soon?": `[ 1.8420892, 1.4021966, -3.6714497] -> negative`

Positive: "I love the service that I experienced!": `[-2.0549004, -1.9490422, 3.938163] -> positive`

Positive: "The workers barely put me on hold and weren't awful, would recommend.": `[ 4.2612185, -1.8605127, -2.227903 ] -> negative`

For the fourth sentence, the model again predicted a negative label even though the sentiment is neutral. The negative and neutral scores were very close, but the model ultimately chose negative, perhaps influenced by the word "discontinued," which carries a negative connotation. Note that even when the scores are close, the model is still counted as entirely incorrect because accuracy depends only on the final predicted label. The third and fifth sentences were correctly classified, but for the last sentence, it predicted a negative label even though the sentiment is positive, likely because the sentence contains several negative words despite its overall positive meaning.

Further, Tables 3 and 4 present the results of modeling the reasons for negative tweets. The labels used in this model are as follows: 1 = "Flight Attendant Complaints", 2 = "Bad Flight", 3 = "Customer Service Issue", 4 = "Lost Luggage", 5 = "Late Flight", 6 = "Damaged Luggage", 7 = "Cancelled Flight", 8 = "Long Lines", and 9 = "Flight Booking Problems". Negative tweets labeled as "Can't Tell" were removed from the dataset prior to training the model.

Table 3. The Confusion Matrix for Modeling Reasons Behind Negative Sentiments.

ACTUAL \ PREDICTION	1	2	3	4	5	6	7	8	9
1	57	13	44	6	18	0	2	0	1
2	7	79	20	8	30	0	4	0	8
3	21	17	714	24	58	0	22	0	40
4	7	2	30	162	17	0	5	0	0
5	10	29	53	7	401	0	6	0	3
6	2	3	0	12	3	0	0	0	0
7	4	6	26	4	17	0	186	0	12
8	5	4	9	5	23	0	0	0	2
9	0	5	68	0	11	0	8	0	57

Table 4. Performance Measures for Modeling Reasons Behind Negative Sentiments.

	Precision	Recall	F1-score	Support
1	0.50	0.40	0.45	141
2	0.50	0.51	0.50	156
3	0.74	0.80	0.77	896
4	0.71	0.73	0.72	223
5	0.69	0.79	0.74	509
6	0.00	0.00	0.00	20
7	0.80	0.73	0.76	255
8	0.00	0.00	0.00	48
9	0.46	0.38	0.42	149
Macro Average	0.49	0.48	0.48	2397
Weighted Macro Average	0.67	0.69	0.68	2397
Overall Accuracy	0.69			2397

This model achieved an overall accuracy of 69%. It performed particularly poorly on category 6 (Damaged Luggage) and category 8 (Long Lines), with 0% accuracy for both—likely due to the very small number of examples in these categories, which made the model unlikely to predict them. Overall, the lower accuracy suggests that identifying

specific reasons for negative tweets is more challenging than predicting general sentiment categories.

Next, we test our own sentences, and the results are shown below.

Results

Flight Attendant Complaints: "The flight attendants were very rude to me and refused to serve me snacks.": `[2.6968806, 0.01893388, 0.84494245, -0.8049284, -0.519119, -1.1598305, -0.90577173, -0.16752988, -1.1670196] -> Flight Attendant Complaints`

Bad Flight: "The flight was very low quality, with lots of turbulence and a messy cabin." `[-0.4667249, 2.6291175, -0.33898532, -1.7089324, 0.7651745, -1.6477805, 0.06079473, -0.47331858, -1.2240868] -> Bad Flight`

Customer Service Issue: "The customer service was horrendous and didn't listen to me, I demand my money back!": `[-0.31218076, -1.1314658, 4.722972, -0.698938, -0.57097685, -1.643983, -0.60682726, -0.701203, -0.11647189] -> Customer Service Issue`

Lost Luggage: "The airport lost my bags, would not recommend.": `[-0.39649984, -1.2392728, -0.41867024, 3.775128, -0.76235855, 0.15429991, -0.30775326, -0.35913384, -0.87179023] -> Lost Luggage`

Late Flight: "The flight randomly got delayed for 3 hours and I was unable to make it to my appointment." `[-1.099647, -0.98199224, -0.3567949, -0.73104644, 3.8831253, -2.0131075, 0.05645005, -0.07423633, -1.2502313] -> Late Flight`

Damaged Luggage: "I had instruments in my luggage, but they were all damaged when I got them back. I demand to be compensated for this!": `[-0.1449628, -0.59634924, 0.09826794, 2.5973651, -0.4798279, -0.07439104, -0.61499524, -0.52286494, -0.69575477] -> Lost Luggage`

Cancelled Flight: `[-1.3964986, -0.9451135, 0.20890956, -0.9620407, 0.7439341, -1.7599922, 3.8737276, -1.1164099, -0.3722449] -> Cancelled Flight`

Continues on the next page.

Results

Continued from the previous page.

Long Lines: "Why is there still only one terminal for this section? I have to wait forever to get to my flight.": `[-0.519021, -0.6651811, 0.60306764, -0.6721138, 1.7523743, -1.4529951, 0.8087218, -0.03857224, -0.86727756] -> Late Flight`

Flight Booking Problems: "I wanted to book some seats all in a row for my family, but the website wouldn't let me even though they all showed as available.": `[-0.48680884, 0.46946236, 1.3060435, -0.8938089, -0.77265763, -1.490014, -0.18414132, -0.9677125, 1.9347365] -> Flight Booking Problems`

We can see that the model performed fairly well, misclassifying only the sixth and eighth ones: it predicted Lost Luggage instead of Damaged Luggage for the sixth one, and Late Flight instead of Long Lines for the eighth one. This is likely because category 6 (Damaged Luggage) and category 8 (Long Lines) were scarcely represented in the training data, making it difficult for the model to learn and recognize it. The lower overall accuracy of this second model, compared with the first, may stem from the larger number of categories and their highly imbalanced representation.

5 Cumulative Logit Model

5.1 Model Definition

Suppose the dataset consists of predictor variables $x_1, \ldots, x_k$ and an ordinal response variable y taking values in c ordered categories. The *cumulative logit model* is defined as [9]:

$$\frac{P(y \leq j)}{P(y > j)} = \exp(\alpha_j + \beta_1 x_1 + \cdots + \beta_k x_k), \quad j = 1, \ldots, c-1,$$

where $\dfrac{P(y \leq j)}{P(y > j)}$ represents the odds of y falling in category j or below. Here, $\alpha_1, \ldots, \alpha_{c-1}$ are the intercepts (which vary across categories), while $\beta_1, \ldots, \beta_k$ are the common slope

coefficients. The fitted models then take the form:

$$\frac{\hat{P}(y \leq 1)}{\hat{P}(y > 1)} = \exp\Big(\hat{\alpha}_1 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\Big),$$

$$\frac{\hat{P}(y \leq 2)}{\hat{P}(y > 2)} = \exp\Big(\hat{\alpha}_2 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\Big),$$

$$\cdots$$

$$\frac{\hat{P}(y \leq c-1)}{\hat{P}(y > c-1)} = \exp\Big(\hat{\alpha}_{c-1} + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\Big).$$

5.2 Application to Airline Tweets Data

In this application, we treat sentiment as an ordered variable with three levels: negative, neutral, and positive. As predictors, we use the 20 most frequent words of each sentiment. When removing duplicates, this leads to a total of 36 words that are fitted. This setup leads to two fitted expressions (this is the model R fits):

$$\frac{\widehat{P}(\text{negative})}{\widehat{P}(\text{neutral or positive})} = \exp\{\widehat{\alpha}_1 - \widehat{\beta}_1\,\text{word}_1 - \cdots - \widehat{\beta}_{36}\,\text{word}_{36}\},$$

and

$$\frac{\widehat{P}(\text{negative or neutral})}{\widehat{P}(\text{positive})} = \exp\{\widehat{\alpha}_2 - \widehat{\beta}_1\,\text{word}_1 - \cdots - \widehat{\beta}_{36}\,\text{word}_{36}\}.$$

After fitting the cumulative logit model, we examine which slope estimates are significant at the 5% level (i.e., with p-values below 0.05). A positive coefficient indicates that the corresponding word is more likely to appear in tweets with negative sentiment, while a negative coefficient suggests that the word is more likely to be associated with positive sentiment. However, in R, the coefficients have the opposite sign, meaning that positive coefficients are associated with positive sentiment, and negative coefficients are associated with negative sentiment.

5.3 Results

The words that have the most negative slopes (while still being significant) are "hold", "hour", "delayed", "call", and "cancelled". The words that have the most positive slopes are "awesome", "great", and "thank". Below, we present sample tweets that utilize these words. Note that the full results of the model are in the Github link at the bottom of the paper, along with all the code used.

Examples of Negative Tweets With Key Words

@xxxxx Calls to 800 resulted in 2hrs of hold time amp; 2day wait to check suspect code share fare. Nothing investigated-my time wasted (2/2)

@xxxxx you Cancelled Flighted our flights for no reason amp; now we have been on the phone for AN HOUR on our vacation. Why?

@xxxxx Delayed due to lack of crew and now delayed again because there's a long line for deicing... Still need to improve service xxxxx

@xxxxx my last call your customer service agent called me back... As he was stuck on hold to air Canada.. Just hope he's booked me a flight

@xxxxx the amount of money I spent on hotels for a WEEK bc of flight Cancelled Flightlation, another flight doesn't make up for the money lost.

Examples of Positive Tweets With Key Words

@xxxxx awesome thank you very much for the help

@xxxxx Great landing in Denver, next Rapid City. Snow starting to fall...Florida Everglades is a faint sunburn away...here comes the cold!

@xxxxx. You guys made my day. Treated me well. Thank you!!!

6 Future Research Directions

The increasing use of social media presents a growing opportunity to gauge customer satisfaction in real time. Building on this study, future research could explore several directions. One avenue is the development and evaluation of alternative feature extraction methods to improve sentiment and category prediction. Another is the use of larger, more diverse, and balanced datasets to address issues related to underrepresented categories and enhance overall model accuracy. Additionally, combining NLP-based approaches with other sources of customer feedback could further improve prediction performance and provide more actionable insights for businesses.

Supplemental Materials

The dataset, the R code used to construct the word cloud, and the Python scripts for both BERT models are available at the following GitHub repository: https://github.com/aidan-shin/airline-sentiment

Acknowledgments

The author would like to express sincere gratitude to Dr. Olga Korosteleva, Professor of Statistics at CSULB, for her guidance and support throughout this project. Appreciation is also extended to the teachers at Troy, including Mr. Rodriguez and Mr. Hwang, whose instruction laid the foundation for this work. Thanks are also due to Dr. Oleg Gleizer and the Math Circle staff at UCLA for their encouragement. Finally, heartfelt gratitude is expressed to the author's family for their constant support and encouragement.

References

[1] Kumar, S. and M. Zymbler. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. *Journal of Big Data*, 6(1), 62.

[2] Wu, S. and Y Gao. (2022). Happy or grumpy? A machine learning approach to analyze the sentiment of airline passengers' tweets. arXiv:2209.14363v1

[3] Tusar, M. T. H. K. and M. T. Islam. (2021). A comparative study of sentiment analysis using NLP and different machine learning techniques on US airline Twitter data. arXiv:2110.00859v1

[4] Shitole, A. S. and A. S. Vaidya. (2023). Machine learning based airlines tweets sentiment classification. *International Journal of Computer Applications*, 185(20).

[5] Hasan, M. and S. A. Fattah. (2024). A machine learning approach to detect customer satisfaction from multiple tweet parameters. arXiv:2402.15992v1

[6] Rustam, F. Ashraf, I., Mehmood, A., Ullah, S., and G. S. Choi (2019). Tweets classification on the base of sentiments for US airline companies. *Entropy*, 21.

[7] Chaudhary, R. and H. Lohiya. (2024). Comprehensive review on machine learning based customer satisfaction from airline tweets. *International Journal of Engineering Applied Science and Management*, 5(9).

[8] Panda, P. K. and J. Sheetlani. (2019). A literature review: Customer satisfaction on airline tweets using machine learning, *International Journal of Advanced Research and Innovative Ideas in Education*, 5(1).