Sentiment Analysis and Topic Modelling in

US Airlines Tweets

**Abstract**

Customer reviews on social media platforms, such as Twitter, are typically unstructured. To effectively analyze this feedback at scale, sentiment analysis, a component of (NLP), is employed. This study employed a hybrid model by integrating TF-IDF and LDA for calculating text similarity. This approach achieved commendable results, with a coherence score of 35% and a silhouette score of 45%, effectively capturing both positive and negative sentiments in comparison to BERTopic. Additionally, a powerful BERT pre-train embedding technique was employed for predicting sentiment classification in the Appendix, Figure 23-24.

1.Introduction

This study analyzes positive and negative sentiments in tweets using advanced NLP techniques to help the airline industry anticipate customer return behaviors. This analysis supports the development of tailored value propositions that address consumer preferences and concerns, ultimately enhancing service quality.

The comprehensive analysis involves several preprocessing steps, followed by the application of (TF-IDF) and BERTopic for feature exploration. The TF-IDF scores are normalized using the 'L2' norm and subsequently utilized in (LDA) for topic modelling. Integrating TF-IDF with LDA provides a more sophisticated approach to topic modelling. This hybrid methodology considers not only the semantic relationships that may exist within the text but also the varying influence of individual words on the overall semantic representation of the documents, as reflected by their distinct weights[1].

By prioritizing terms that significantly contribute to semantic meaning, the model can generate insights that are closely aligned with user sentiments and preferences. Additionally, the results include a comparison between the hybrid models and advanced BERTopic, providing a thorough evaluation of their effectiveness in capturing customer sentiments.

2. Data Collection & Pre-processing

|  |
| --- |
| **About the Dataset**  The "Twitter US Airline Sentiment" dataset comprises 14,640 tweets pertaining to six major U.S. airlines: American, United, US Airways, Southwest, Delta, and Virgin America. This dataset exhibits a highly imbalanced distribution of sentiments, with 62.7% of the tweets classified as negative, 21.2% as neutral, and 16.1% as positive[2].  After removing duplicate entries, the dataset is reduced to a total of 14,604 tweets. For the purposes of this analysis, the focus will be exclusively on positive and negative sentiments. Consequently, positive sentiments will be encoded as 1, while negative sentiments will be encoded as 0. |

*Table 1: Dataset Information*

Prior to tokenization, several essential text preprocessing steps are undertaken, as illustrated in Figure 8. These procedures collectively serve to reduce noise, ensuring that the text is presented in a consistent format, emphasizing core content, and facilitating more accurate tokenization.

After tokenization, *lemmatization* is applied as a key component of the preprocessing pipeline. These steps reduce words to their root forms and eliminate suffixes, thereby enhancing semantic clarity and providing more precise insights into the text data.

Subsequently, TF-IDF vectorization is employed to identify significant terms within individual documents while considering their frequency across the entire corpus. The TF-IDF scores are normalized using the 'L2' norm. To evaluate the performance of the TF-IDF features, K-fold cross-validation is utilized in conjunction with logistic regression, resulting in F1 scores of 0.93 for negative sentiments and 0.65 for positive sentiments.

A screenshot of a computer program

Description automatically generated

*Figure 8: Text Pre-Processing*

2.1 TF-IDF Word Cloud

### 2.1.1 TF-IDF – Positive Tweet

**American** often receives positive feedback with terms like "service," "thank," and "good," indicating strong customer satisfaction. However, when service issues arise, customers tend to compare American to JetBlue.

**A close up of words

Description automatically generated**

*Figure 9: TF-IDF – Positive Tweet*

### 2.1.2 TF-IDF – Negative Tweet

A close-up of words

Description automatically generated**Customer service** emerged as the most highlighted aspect in customer feedback, particularly in relation to **flight delays, cancellations, and language barriers**. These operational challenges and communication difficulties can significantly impact their overall travel experience.

*Figure 10: TF-IDF – Negative Tweet*

## 2.2 BERTopic

In this section, we will explore how BERTopic leverages transformer models to capture semantic relationships, thereby distinguishing itself from traditional TF-IDF approaches.

### 2.2.1 BERTopic: Positive Tweet

Figures 11 depict positive words such as "**love","thank,"help and"great**" across topics 0 to 3, showing that their intertopic distances are relatively close, as illustrated in Figure 14. Additionally, Figure 13 indicates that **customer service and American Airlines** exhibit a higher similarity score of 0.76, suggesting a strong connection in the context of customer feedback. This highlights the positive associations customers have with service experiences related to American Airlines.

**A group of colorful bars

Description automatically generated with medium confidence**

A chart with colorful lines

Description automatically generatedFigure 11:*BERTopic positive Bar Chart with unique Eight Topic*

*Figure 12 :**BERTopic positive Hierarchical clustering*

Figure 13 :*BERTopic positive Similarity Matrix*

A screenshot of a computer

Description automatically generated

*Figure 14* ***:*** *BERTopic positive Intertopic Distance Map*

**A screenshot of a map

Description automatically generated**

### 2.2.2 BERTopic-Negative Tweet

In contrast to positive sentiments, negative sentiments such as "**cancel," "delay,"** and **"carrier issues**" are depicted in Figure 15. Additionally, inflight issues, including **Wi-Fi and delay**, show a similarity score above 0.50 in Figures 17. The intertopic distances among Topics 0 to 4 indicate a connection to customer service, with **Southwest Airlines** exhibiting the highest distance of 1211, particularly related to flight cancellations, as shown in Figure 18.

A graph of a bar graph

Description automatically generated with medium confidence

*Figure 15:**BERTopic Negative Bar Chart with unique Eight Topic*

**A screenshot of a computer screen

Description automatically generated**

*Figure 16* ***:*** *BERTopic Negative Hierarchical clustering*

*Figure 17 :**BERTopic Negative Similarity Matrix*

**A screenshot of a computer screen

Description automatically generated**

*Figure 18* ***:*** *BERTopic Negative Intertopic Distance Map*

**A screen shot of a map

Description automatically generated**

# 3. Topic Modelling Building

a. Latent Dirichlet Allocation (LDA)

LDA is a powerful tool for identifying topics within textual data. One of its primary strengths lies in the flexibility to tune hyperparameters, particularly the topic density parameter, beta. This parameter governs the distribution of words across topics, enabling researchers to adjust the granularity of topic extraction. By carefully fine-tuning beta, users can enhance topic coherence and improve the model's capacity to capture meaningful themes within the corpus[3, 4].

*Table 2: LDA model Key Parameters & Libraries*

|  |  |
| --- | --- |
| **Implementation Process** | The implementation process began with the application of the TF-IDF vectorizer, configured to extract both unigrams and bigrams (ngram\_range=(1, 2)). A maximum of 100 features was specified, while a minimum document frequency of 5 was employed to filter out infrequent terms. Subsequently, the TF-IDF matrix was normalized using L2 normalization to ensure that the magnitude of the vectors did not unduly influence the distance metrics utilized in the (LDA) model.  To enhance the robustness of the model, hyperparameter tuning was conducted through a randomized search combined with cross-validation. This systematic approach facilitated the identification of optimal parameters for the LDA model. Ultimately, the model was fitted to the normalized TF-IDF matrix, and the best parameters were determined based on the results of the randomized search. |
| **Key Parameters** | * **n\_components (5,10,15)** Specifies the number of topics to be extracted. The choice of this parameter is critical, as too few topics may oversimplify the dataset, while too many may introduce noise. * **doc\_topic\_prior (α = 0.1,0.5.1.0):** This parameter represents the document-topic density. It controls how concentrated the topics are within documents. A smaller value leads to more sparse topic distributions, meaning each document is likely to have a few dominant topics, while larger values allow for a more uniform distribution across topics. * **topic\_word\_prior (β = 0.01,0.1,0.5):** Influences the distribution of words across topics. A lower value indicates that topics will have fewer words associated with them, potentially leading to more focused topics. * **learning\_decay:** Controls how quickly the learning rate decreases during training. A lower value means faster convergence but may lead to overfitting. * **learning\_offset** :Provides a baseline for the learning rate. Higher values can help stabilize learning in the early stages. * **max\_iter**: Determines how long the algorithm will run to converge. More iterations allow for better fitting but increase computation time. |
| **Libraries** | * **Pandas:** Used for data manipulation and cleaning, facilitating the organization and handling of the dataset. * **Scikit-learn :** This library was instrumental in implementing the TF-IDF vectorizer, LDA model, and hyperparameter tuning through RandomizedSearchCV. * **Numpy :** A core library for numerical operations, helping with the handling of arrays and mathematical computations. |

# 4. Unsupervised Topic Modelling Evaluation

LDA models undergo Randomized 3-fold Cross-Validation (CV), and the best parameters were fitted to the predicted labels for comparison against the true labels. The evaluation metrics of these models are summarized in Table 3.

|  |  |  |
| --- | --- | --- |
| **LDA Models** | **Negative LDA** | **Positive LDA** |
| Best Parameter by 3 folds Randomised CV | * doc\_topic\_prior=0.5 * learning\_decay=0.5 * learning\_offset=20.0 * max\_iter = 0.00 * n\_components=5 * random\_state=4 * topic\_word\_prior=0.01 | * doc\_topic\_prior=0.1 * learning\_decay=0.5 * learning\_offset=30 * max\_iter=20 * n\_components=15 * random\_state=4 * topic\_word\_prior=0.5 |
| **Performance Metrics**   * NMI * ARI * Homogeneity * Completeness * V-measure * Silhouette Score | NMI = 0.30  ARI = 0.00  Homogeneity = 0.18  Completeness = 1.00  V-Measure = 0.30  Silhouette Score = 0.37  **Evaluation:**  **NMI** measures the amount of information obtained about one clustering from the other. A value of 0.30 suggests a moderate level of agreement between the predicted clusters and the true labels.  **ARI** assesses the similarity between two data clustering’s, adjusting for chance. A score of 0.00 indicates that the clustering results are no better than random assignment, meaning there is essentially no agreement between the predicted clusters and the true labels.  **Homogeneity** measures whether each cluster contains only members of a single class. A score of 0.18 indicates low homogeneity, meaning that many clusters contain mixed classes.  **Completeness** measures whether all members of a given class are assigned to the same cluster. A score of 1.00 indicates perfect completeness, meaning that all instances of the true labels are contained within the identified clusters. However, this high score could be misleading without high homogeneity, as it suggests that while the clusters contain all members of a class, they also contain members of other classes.  **V-Measure** is the harmonic mean of homogeneity and completeness. A score of 0.30 indicates that while there is some completeness in the clustering, the low homogeneity suggests that the clusters are not meaningful or distinct.  **Silhouette** core of 0.37 suggests that the clusters have some degree of separation, but the score is relatively low, indicating that many data points may be poorly assigned or close to cluster boundaries.  The negative LDA results indicate that while there is some completeness in the clustering (all class members are included), the clusters themselves are poorly defined, containing a mix of different sentiment classes. The low NMI and ARI suggest that the clustering model is not effectively capturing the underlying structure of the data. | NMI = 0.52  ARI = 0.001  Homogeneity = 0.350  Completeness = 1.00  V-Measure = 0.52  Silhouette Score = 0.50  **Evaluation:**  **NMI** score of 0.52 suggests a moderate level of separating between predicted clustering and actual labels, indicating that there is some correspondence, but also significant room for improvement.  **ARI** score of 0.001 is low, suggesting minimal similarity between the clusters formed by the model and the ground truth. It indicates that the model may not be effectively capturing the underlying topic structure.  A **homogeneity** score of 0.35  indicates that the clusters contain a mix of different classes, suggesting that the model is not effectively grouping similar data points together.  Even though it has low homogeneity score, it still has **completeness** score of 1.0 in captures same true labels.  **V-Measure** score of 0.52 suggests that while the model captures the same labels together, it does so with substantial impurity in the clusters.  **Silhouette** score of 0.50 indicates that there is a reasonable degree of separation between the clusters.  The positive LDA model results present a mixed picture of the clustering quality. While it achieves perfect completeness and a moderate NMI and V-Measure, the low ARI and homogeneity scores indicate that the clusters may not be effectively capturing distinct topics. The silhouette score suggests that the model's clusters have some separation but are not well-defined. |

*Table 3: Positive & Negative Best Parameters with Performance Metrics*

# 5. Topic Modelling Results and Interpretation

In this section, the results of (LDA) topic modelling for both positive and negative sentiments are visualized through pyLDAvis and T-SNE plots. Additionally, a comparison between BERTopic embeddings and traditional LDA is conducted. While LDA focuses on identifying topics based on word co-occurrences within the corpus, BERTopic leverages embeddings to capture semantic similarities, resulting in potentially richer and more nuanced topic representations. This comparison underscores the strengths and weaknesses of both approaches in effectively analyzing sentiment within the dataset.

## 5.1 Positive LDA Visualizations

Topic 1 to 6 in figure 20 had closely distance suggesting themes of appreciation and compliments related to flight agents resolving customer service issues.

A diagram of a number of colored dots

Description automatically generated with medium confidenceFigure 19 also presents the T-SNE plot, which demonstrates some overlap between Topics 11,13 and 14. This overlap indicates that while there are distinct themes, certain customer experiences may share common sentiments.

*Figure 19: Positive LDA T-Sne plot*

A screenshot of a computer

Description automatically generated

*Figure 20 : Intertopic Distance positive LDA*

## 5.2 Negative LDA Visualizations

The analysis of negative topics identified that the top four themes predominantly revolve around customer service and carrier-related issues. Topics 3 and 4 focus on offline concerns, such as email responses, while Topics 1 and 5 address in-flight customer service issues. Collectively, these topics highlight significant challenges faced by customers in their interactions with service providers.

There is a noticeable overlap among Topics 0,3 and 4 suggesting that some negative sentiments may be interconnected. These findings underscore critical areas where improvements are needed, particularly in customer service and operational efficiency, to mitigate negative customer experiences.

A diagram of a graph

Description automatically generated with medium confidence

*Figure 21: Negative LDA T-Sne plot*

Figure xx: Top four Negative LDA Topic

## A screenshot of a computer Description automatically generated5.3 Comparison of BERTopic and LDA

*Figure 22: Negative LDA word Intertopic Distance*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **BERTopic** | | **LDA** | |
| Sentiments | **Positive** | **Negative** | **Positive** | **Negative** |
| Coherence score | 0.34 | 0.38 | 0.36 | 0.34 |
| Av. Coherence | 0.36 | | 0.35 | |
| Silhouette | 0.11 | 0.05 | 0.50 | 0.39 |
| Av. silhouette | 0.08 | | 0.45 | |

*Table 4a: Comparison between BERTopic & LDA Metrics*

|  |  |
| --- | --- |
| **BERTopic** | **LDA** |
| The average coherence score of 0.36 suggests that BERTopic embedding ability to generate well defined topic in negative sentiment (0.38).  While coherence is strong, if the silhouette scores are low, it may indicate that the topics are not well-separated. This could lead to overlap in themes that are difficult to interpret. | An average coherence score of 0.35 shows that the topics generated are interpretable.  The average silhouette score of 0.45 indicates that the clusters formed by LDA are well-defined especially for positive sentiment (0.50). However, LDA may struggle with capturing the subtlety and richness of negative sentiments, as reflected in its coherence scores for negative topics. |

*Table 4b: Evaluation between BERTopic & LDA*

In summary, BERTopic’s embedding approach in negative sentiment demonstrates its robustness in capturing complex themes, achieving a coherence score of 38%.

In contrast, LDA excels in positive sentiment, demonstrating its ability to delineate well-defined topics with a high silhouette score of 50%. While LDA benefits from hyperparameter tuning to optimize its performance, it shows stronger topic separation and distinctness, as indicated by its higher silhouette scores, despite having slightly lower coherence compared to BERTopic.

Conversely, although BERTopic exhibits better coherence, its significantly lower silhouette scores suggest that the topics may not be well-separated, potentially complicating interpretation. Overall, LDA emerges as the better choice once hyperparameter tuning is applied.

# 6. Conclusion / Limitations & Future Work

Attracting repeat customers is crucial for the long-term success of the airline industry. This study integrates a hybrid model that combines TF-IDF with LDA and compares it with BERTopic.

The comprehensive analysis results show that the hybrid model performs better than BERTopic, with optimized parameters for both positive and negative sentiments through hyperparameter tuning. For positive sentiment, LDA achieved 15 topics with parameters α=0.1 and β=0.5 resulting in a coherence score of 0.36 and a silhouette score of 0.50. In contrast, for negative sentiment, LDA identified 5 topics with α=0.5 and β=0.5 achieving a coherence score of 0.34 and a silhouette score of 0.39. This indicates LDA's superior ability to generate well-defined topics and effectively distinguish between different customer sentiments. Despite BERTopic's overall poorer performance, it still achieved a high coherence score of 0.38 for identifying negative sentiments.

Future work could explore models like RoBERTa to overcome the token limit constraints of BERT. Additionally, integrating other popular unsupervised embeddings, such as GloVe, with RNNs could leverage their unique strengths in text representation and sequence modeling.

In summary, this study contributes to enhancing customer satisfaction by highlighting the interrelationship between positive and negative sentiments. The airline service industry often faces negative tweets that emphasize underlying issues such as meal quality, Wi-Fi connectivity, and weather-related to flight delays. Effectively addressing these concerns through customer service management can transform negative experiences into positive sentiments, ultimately improving overall customer perception.

7. References

[1] J. Wang, W. Xu, W. Yan, and C. Li, "Text Similarity Calculation Method Based on Hybrid Model of LDA and TF-IDF," in *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, Normal, IL, USA, *March 4,* 2020: Association for Computing Machinery, pp. 1–8.

[2] Kaggle,*Twitter US Airline sentiment*, 2019. [Dataset]. Available: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment/data?select=database.sqlite.[Accessed: Oct 1, 2024]

[3] M. A. Vázquez, J. Pereira-Delgado, J. Cid-Sueiro, and J. Arenas-García, "Validation of scientific topic models using graph analysis and corpus metadata," *Scientometrics,* vol. 127, no. 9, pp. 5441-5458, September 2022, [Online]. Available: https://doi.org/10.1007/s11192-022-04318-5. [Accessed Oct.15, 2024]

[4] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology,* vol. 7, pp. 1-13, May 2022. [Online]. Available : https://doi: 10.3389/fsoc.2022.886498. [Accessed Oct. 15, 2024]

# 8. Appendix

## 8.1 Abbreviations

**TFIDF:** Term Frequency Inverse document frequency

**NLP:** Natural Language Processing

**NLTK:** Natural Language toolkit

**BERT:** Bidirectional Encoder Representations from Transformers

**ARI:** Adjusted rand Index

**NMI:** Normalized mutual information

**LDA:** Latent Dirichlet Allocation

**\*\* Figures order starts from Appendix**

## 8.2 List of Figures and Tables in the Report

|  |  |  |  |
| --- | --- | --- | --- |
| **Figure 1** | US Twitter Sentiment Analysis Experimental Design | **Figure 12** | BERTopic positive Hierarchical clustering |
| **Figure 2** | Distribution of Target variable: Airline Sentiment | **Figure 13** | BERTopic positive Similarity Matrix |
| **Figure 3** | Six Airline Sentiment Distribution | **Figure 14** | BERTopic positive Intertopic Distance Map |
| **Figure 4** | Negative Tweets Reason | **Figure 15** | BERTopic Negative Bar Chart with unique 8 Topic |
| **Figure 5** | Negative Tweets Reason across six airlines | **Figure 16** | BERTopic Negative Hierarchical clustering |
| **Figure 6** | Vader Sentiment scoring result | **Figure 17** | BERTopic Negative Similarity Matrix |
| **Figure 7** | Frequency of words | **Figure 18** | BERTopic Negative Intertopic Distance Map |
| **Figure 8** | Text Pre-processing | **Figure 19** | Positive LDA T-Sne plot |
| **Figure 9** | TF-IDF – Positive Tweet | **Figure 20** | Intertopic Distance positive LDA |
| **Figure 10** | TF-IDF – Negative Tweet | **Figure 21** | Negative LDA T-Sne plot |
| **Figure 11** | BERTopic positive Bar Chart with unique 8 Topic | **Figure 22** | Negative LDA word Intertopic Distance |
|  |  | **Figure**  **23a**  **23b**  **23c** | BERT Pre-trained Prediction Train-Test Accuracy  BERT Pre-trained Prediction Train-Validation Loss  BERT Pre-trained Prediction Classification Report |
|  |  | **Figure 24** | BERT Pre-trained Sentiment Prediction |

|  |  |
| --- | --- |
| **Table 1** | Dataset Information |
| **Table 2** | LDA model Key Parameters & Libraries |
| **Table 3** | Positive & Negative Best Parameters with Performance Metrics |
| **Table 4a**  **Table 4b** | Comparison between BERTopic & LDA Metrics  Evaluation between BERTopic & LDA |

## 8.3 US Twitter Sentiment Analysis Experiment Design

A diagram of data processing

Description automatically generated

*Figure 1: US Twitter Sentiment Experimental Design*

## 8.4 Exploratory Data Analysis

Figures 2 and 3 illustrate the sentiment distribution within the dataset, revealing that negative sentiment comprises the largest portion at 62.7%, followed by neutral sentiment at 21.2% and positive sentiment at 16.1%. Within the negative sentiment category, the bar plot shows that United Airlines accounts for 18% of the negative tweets, US Airways represents 15.5%, and American Airlines makes up 13.4%.

A graph of negative and neutral expression

Description automatically generated

*Figure 2: Distribution of Target variable: Airline Sentiment*

A graph of different colored bars

Description automatically generated

*Figure 3: Six Airline Sentiment Distribution*

A graph of negative tweets

Description automatically generatedFigures 4 and 5 illustrate that American Airlines, US Airways, and Southwest Airlines received the highest number of negative tweets. Key indicators of customer dissatisfaction include issues related to customer service, late flights, cancelled flights, and lost luggage. These factors significantly contribute to the negative sentiments expressed by passengers, underscoring critical areas for improvement in airline service quality.

*Figure 4: Negative Tweets Reason*

A graph of different colored bars

Description automatically generated

*Figure 5: Negative Tweets Reason across six airlines*

The F1 score results from the VADER sentiment scoring indicate that positive sentiment achieved the highest score of 0.64, reflecting a strong ability to accurately identify positive sentiments in the dataset. In comparison, neutral sentiment received an F1 score of 0.40, while negative sentiment scored 0.49. Overall, the accuracy of the model is only 54%, suggesting that it may not be reliable for analysis. Consequently, VADER was not selected for this study, as its performance does not meet the necessary standards for effective sentiment analysis.

A screenshot of a test

Description automatically generated

*Figure 6: Vader Sentiment Scoring Classification report*

Figure 7 displays the frequency of words in the tweets, highlighting that terms predominantly associated with negative sentiment, such as "customer service," "delay," are significant contributors to the overall expressions of dissatisfaction among passengers.

A close up of words

Description automatically generated

*Figure 7: Frequency of words*

## 8.5 BERT Pre-trained Embedding Predictions

Figure 23c results of the BERT sequence classification predictions indicate a superior F1 performance for the positive class at 88%, compared to 83% for the negative class and 79% for the neutral class. At epoch 1, the training accuracy is 0.80, and the test accuracy is 0.78 in Figure 23a, suggesting that the model is generalizing well to unseen data. Additionally, in Figure 23b, the validation loss of 0.55 is higher than the training loss of 0.52, which is a common observation in early training stages.

A screenshot of a computer screen

Description automatically generatedA graph with blue and orange lines

Description automatically generatedA graph with blue and orange lines

Description automatically generated

*Figure 23c: BERT Pre-trained Prediction Classification Report*

*Figure 23b: BERT Pre-trained Prediction Train-Validation Loss*

*Figure 23a: BERT Pre-trained Prediction Train-Test Accuracy*

*Figure 24: BERT Pre-trained Sentiment Prediction*

A screenshot of a computer program

Description automatically generated