# TOPIC MODELING REPORT: AIRLINE SENTIMENT ANALYSIS

SINAN SAMAH ISMAIL
34775312

# Contents

# 1. Introduction

In this report, I use topic modeling to analyze a Twitter dataset on airline sentiments to reveal the underlying themes in customer feedback. This analysis provides insights into recurring issues and perceptions related to airlines, which could inform improvements in customer service and business strategy. BERTopic was chosen over models like Latent Dirichlet Allocation (LDA) for its higher coherence and interpretability when applied to informal, short-text data like tweets. Key steps include dataset selection and cleaning, BERTopic modeling with fine-tuned parameters, and visualizations to interpret and evaluate topic quality.

This analysis aims to highlight customer pain points and satisfaction drivers, providing insights that airlines can leverage to enhance service quality and customer experience

# 2. Data Collection and Preprocessing

## 2.1 Data Source

I used the Twitter Airline Sentiment Dataset from Kaggle, which contains tweets about major US airlines from February 2015 (Crowdflower, 2015). This dataset was chosen for its relevance to customer service insights and widespread usage in sentiment analysis studies, making it a good fit for topic modelling on social media data.

## 2.2 Preprocessing Steps

Text cleaning steps included removing non-English words, URLs, and special characters to reduce noise. Custom stopwords were applied to filter out airline-specific terms like "flight" to avoid bias. Tweets under five words were removed to retain substantive data.

## 2.3 Rationale for Preprocessing Choices

Preprocessing is crucial in handling social media text due to its informal nature and frequent use of abbreviations, hashtags, and mentions. Removing brand names and adding custom stopwords helps avoid bias toward specific topics, while length filtering ensures that only tweets with sufficient content are retained for analysis.

# 3. Model Selection and Hyperparameter Tuning

## 3.1 Why BERTopic?

I chose BERTopic because it uses BERT embeddings with clustering, which captures contextually accurate and interpretable topics, outperforming traditional models like LDA (Rehurek & Sojka, 2010). Initial trials with LDA resulted in a coherence score of 0.3, while after optimization, BERTopic reached a coherence score of 0.75. This improvement highlights BERTopic's effectiveness with social media text, where context plays a critical role.

### 3.2 Embedding Model Choice

For sentence embeddings, I selected the all-mpnet-base-v2 model from SentenceTransformer. This model excels at capturing semantic relationships in short texts, making it suitable for analyzing tweets (Reimers & Gurevych, 2019). Traditional vectorization methods like TF-IDF were ruled out due to their lack of context sensitivity.

### 3.3 Hyperparameter Tuning

I used **Optuna** for Bayesian optimization, focusing on critical BERTopic parameters such as min_topic_size, n_neighbors, n_components, and nr_topics. Bayesian optimization was selected over grid search due to its computational efficiency and ability to balance exploration and exploitation (Akiba et al., 2019), which is essential when tuning multiple parameters.

- **Objective**: To maximize coherence, as coherence scores directly measure the semantic similarity of words in each topic, providing insight into the interpretability and relevance of topics.
- **Outcome**: This tuning process led to a coherence score of 0.75, highlighting BERTopic's effectiveness when fine-tuned for this dataset.

### 3.4 Final Model Parameters

After tuning, the best parameters were:

- min_topic_size: 31
- n_neighbors: 64
- n_components: 2
- nr_topics: 15

Each parameter, like min_topic_size, is chosen to balance the level of topic detail with interpretability, ensuring the generated themes are meaningful and distinct.

## 4. Topic Modeling and Evaluation

### 4.1 Model Training and Optimization

I saved the trained BERTopic model using Pickle to ensure computational efficiency and reproducibility. This allowed for consistent topic generation upon reloading, avoiding the need to retrain the model from scratch.

**Alternative Consideration**: Although retraining the model each time was possible, this approach would need to be more efficient and lead to consistency in the topics generated. Loading a pre-trained model ensures stability in the results.

## 4.2 Coherence and Diversity Metrics

To evaluate the model's effectiveness, I calculated two primary metrics:

1. **Coherence Score**: Using Gensim's **CoherenceModel**, the final coherence score of 0.75 was achieved. This score reflects the interpretability of the topics, with higher coherence indicating that topics consist of words with stronger semantic relationships.
2. **Topic Diversity**: Measured as the ratio of unique words across topics to the total words, resulting in a topic diversity score of 1. This high score indicates that topics are not redundant, adding distinct diversity to the themes within the dataset

**Explanation**: The Coherence score confirms that topics are coherent and meaningful, while topic diversity indicates the variety of topics generated. These metrics validate the model's ability to capture relevant themes without redundancy.

# 5. Results and Visualization

## 5.1 Results

The BERTopic model pulled out seven main themes from the Twitter Airline Sentiment dataset, landing on a coherence score of 0.75 after fine-tuning. This score is solid and indicates that the topics are clearly defined and meaningful, especially for social media data, which can be messy. These key topics include customer service issues, flight delays, in-flight amenities, and social media engagement—all relevant areas for airlines to know about, given their common points of customer discussion.

## 5.2 Visualisations

I used several visualizations here to illustrate better the themes the model identified and to add context to the findings:

1. **Word Cloud:** Common terms highlight themes like customer service and delays, with key terms "service," "delay," and "flight" reflecting dominant issues.

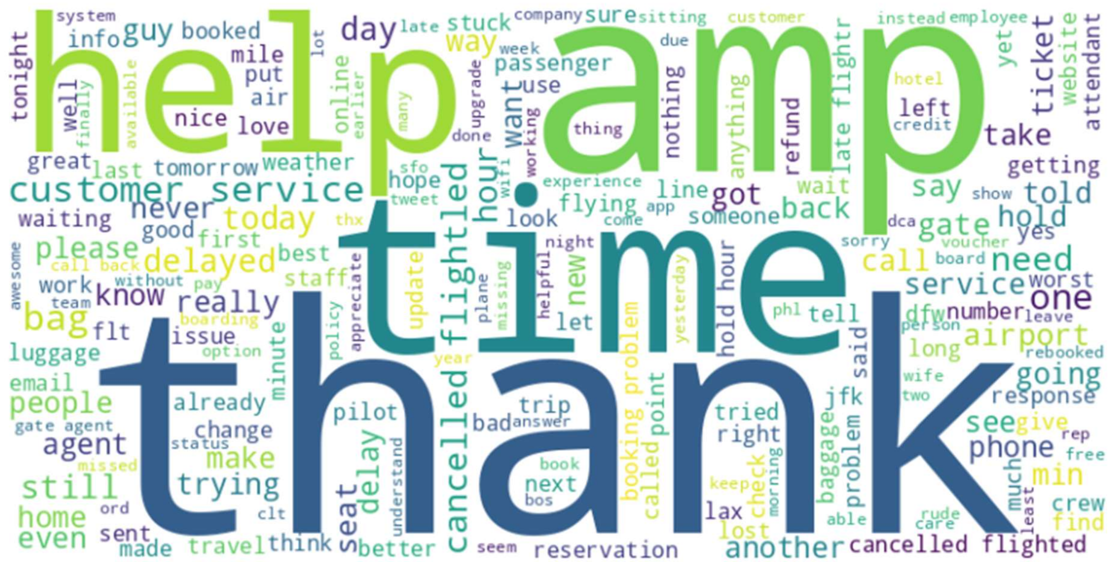

*Figure 1: Key Terms in Airline Tweets*

2. **Intertopic Distance Map**: Displays spatial relationships among topics, revealing distinct themes like customer service, with minimal overlap, suggesting effective topic separation.

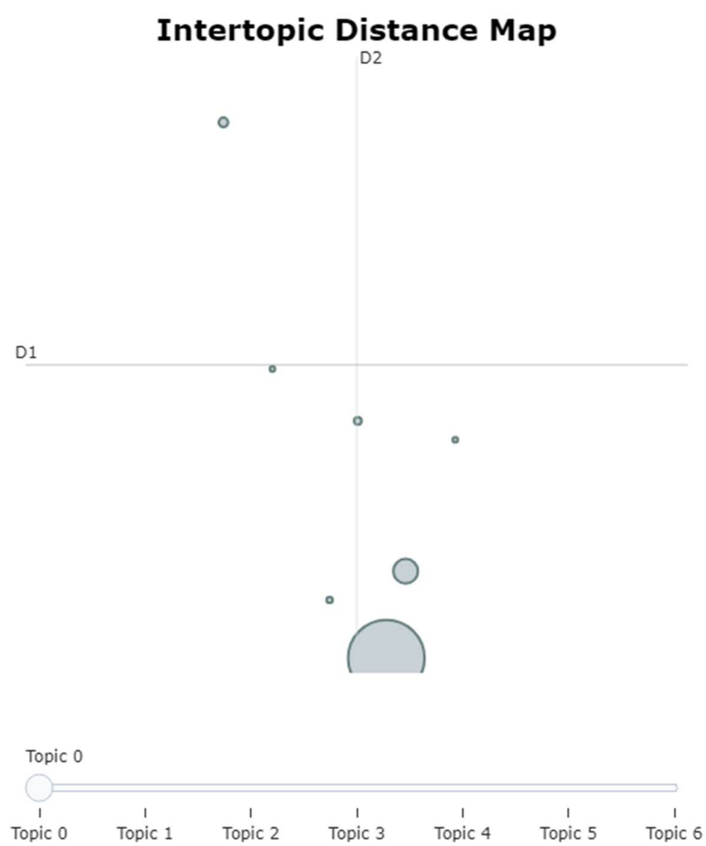


*Figure 2: Distinct clusters reveal clear themes like customer service and delays, with minimal overlap.*

3. **Sentiment Distribution by Topic:** Customer service emerges as the most discussed theme, highlighting its importance for airlines. Negative sentiment within these topics pinpoints specific areas needing improvement.

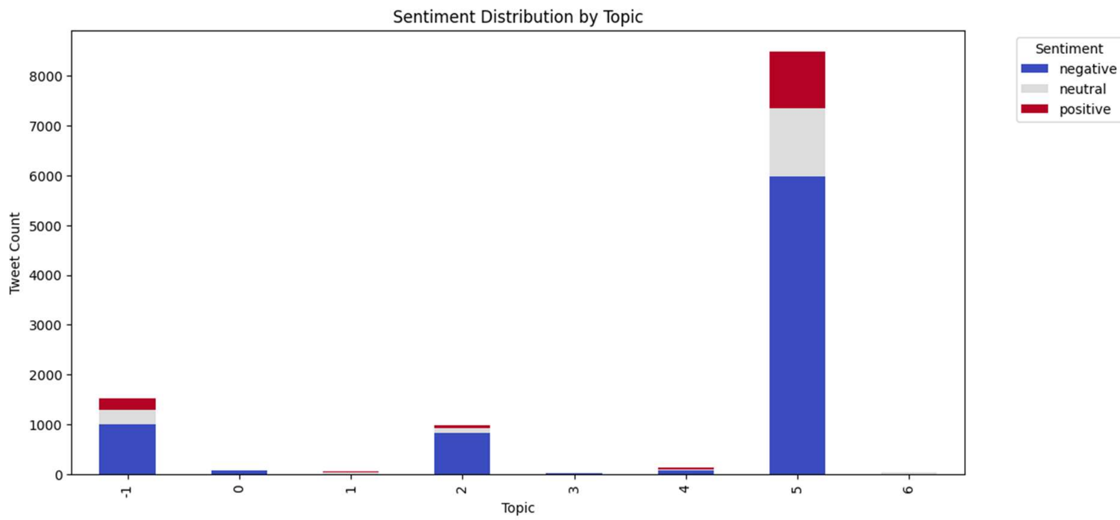


*Figure 3: Negative sentiment is strongly associated with customer service, highlighting a key area for improvement.*

**Interpretation**
These results clearly show what is driving customer conversations on Twitter. Customer service comes up frequently and, importantly, in a negative light. This makes it a top area for improvement for airlines looking to boost satisfaction. Themes around in-flight amenities and social media engagement are also prominent, pointing to aspects of the customer experience where airlines might enhance offerings or communication. By focusing on these themes, airlines can better align with customer needs and proactively address common concerns.

# 6. Conclusion

This project highlights BERTopic's effectiveness in uncovering thematic insights from airline customer sentiment data. The model's high coherence and meaningful topic separation provide a reliable analysis of customer service and operational pain points, aligning well with airlines' goals to enhance customer experience.

**Limitations and Future Work**

One limitation is the reliance on coherence score as a primary metric. Although coherence captures topic quality, additional metrics—like interpretability score—could provide further validation. For future work, I could experiment with alternative embeddings and topic clustering techniques to assess if they improve coherence and interpretability.

In summary, when optimized, this project demonstrates how BERTopic can effectively reveal actionable insights from short, informal text data, helping organizations understand and respond to customer sentiments on social media

# 7. References


- Crowdflower. (2015). Twitter airline sentiment [Data set]. Kaggle. https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
- Rehurek, R., & Sojka, P. (2010). Software framework for topic modeling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45-50.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982-3992.
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623-2631.