Unsupervised Sentiment Clustering of Airline Tweets

Zhao Nailin, Chan Meow Ling Jacqelene, Chng Hui Mei, Xiao Lin
National University of Singapore
MSI5001 – Introduction to AI: Concepts, Applications, and Evaluation
Semester 1 (2025/26)

Introduction

Airlines face constant public scrutiny, especially on social media platforms where passengers frequently express dissatisfaction related to delays, cancellations, service lapses and other travel disruptions. Negative sentiments can escalate quickly, shaping brand perception and potentially triggering public relations challenges. Understanding sentiments in a timely manner is crucial for airlines to maintain customer trust and protect brand reputation.

Pain Points

The volume of customer feedback online is massive. Manually reading each tweet is highly inefficient and impractical for ongoing sentiment monitoring. Moreover, social media language is often emotional, informal or sarcastic, requiring more advanced techniques than simple keyword filtering.

Business Impact & Value

Automated sentiment clustering enables airlines to:

- Detect emerging customer concerns early
- Identify distinct clusters efficiently
- Prioritise operational improvements (e.g., baggage handling, delay management)
- Enhance responsiveness to concerned service-related issues
- Improve customer satisfaction and overall travel experience

By extracting insights from tweets, airlines can make more informed and proactive decisions. Our insights demonstrate how data-driven sentiment mining enables proactive action to reduce dissatisfaction and operational risk.

Objective

This project aims to:

- Classify tweets into distinct clusters (positive, neutral, negative sentiment groups) using unsupervised learning
- Compare multiple embedding and clustering models to determine which approach yields the most appropriate sentiment grouping
- Surface actionable insights - such as which airline receives more negative tweets and what topics dominate user dissatisfaction

These findings demonstrate how NLP-driven sentiment analysis can help airlines understand and respond to customer voices at scale.

Methodology

Unsupervised sentiment discovery on airline tweets using multiple combinations of data cleaning methodologies, text vectorization and embeddings models, clustering algorithms, and evaluation metrics were studied.

Data Source

Airline Sentiment dataset was selected because it provided a realistic, sentiment-rich text corpus with clear business relevance (customer experience monitoring) and a challenging unsupervised learning setup (discovering sentiment groups without using labels during training).

Data Preprocessing

The data were processed in the following order:

- Convert texts to lower cases;
- Remove common English stop words, blank spaces, &amp, RT;
- Give text emoji (such as T_T) a sentiment;
- Demojize emoji to get the meaning behind emoji;
- Remove URLs because we cannot do crawling;
- Remove mentions and the party behind @;
- Drop # but keep the topic mentioned after #;
- Remove numbers and duplicated rows;

Punctuation marks were not removed because of better clustering results (see analysis in below section). We arrived at a final cleaned dataset of 14,419 records (98% of the records) ([10-28] cleaned_text.xlsx) which the remaining texts are good in representing sentiments.

Vectorization and Text Embedding

TF-IDF and several BERT models to do contextual embeddings were considered.

Table 1 : Models Considered

Type	Model	Description
Statistical (sparse)	TF-IDF	Counts term frequencies adjusted by inverse document frequency.
Sentence Transformer	all-MiniLM-L6-v2	General-purpose BERT model.
	all-mpnet-base-v2	Larger BERT models with higher accuracy for semantic similarity.
Transformer LM for tweet	cardiffnlp/twitter-roberta-base-sentiment-latest (Twitter-Roberta)	Fine-tuned on 58M tweets. Handles social media text better.

Clustering Algorithms

K-Means (cosine distance via L2-normalized features), Agglomerative clustering, and Gaussian Mixture Model were considered.

Table 2 : Clustering Considered

Type	Model	Description
Centroid-based	K-Means	Assign points to the nearest centroid
Hierarchical	Agglomerative	Start with every point as its own cluster, iteratively merge closest clusters
Probabilistic	Gaussian Mixture Model	Model data as a mixture of Gaussian distribution estimated via EM algorithm

Evaluation Metrics

Cluster purity against given sentiment labels, silhouette cosine (used silhouette cosine instead of silhouette score because silhouette cosine focuses on how aligned the samples are in direction -- similarity in angle, making it better suited for high-dimensional or directional data like text embeddings.) and visual inspection were used in evaluating the models. Although sentiments labels provided in the dataset are not allowed to be used during clustering, the labels are used only for evaluation:

Table 3 : Evaluation Metrics Considered

Metrics	Measures	Equation
Cluster Purity	Alignment of clusters with true class labels	$\text{Purity} = \frac{1}{N}\sum_{k=1}^{K}\max_{j}\lvert C_k \cap L_j\rvert$
Silhouette (Cosine)	Compactness + separation of clusters using cosine distance	$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$
Confidence Correlation	Evaluate cluster certainty	Low confidence → misclustered
Visual Inspection	Clear boundaries, compact groupings	t-SNE 2D and 3D, PCA 3D

Results and Analysis

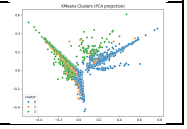
Comparison Summary

All nine modes were done using Airline_sentiment_3_(final).ipynb on the final dataset (*cleaned_tweets.xlsx*).

Data Cleaning Methodology Analysis

Default: Convert texts to lower cases, remove common English stop words, remove URLs because we cannot do crawling in this project; remove mentions and the party behind @, drop # but keep the topic mentioned after #, remove numbers, remove duplicate.

Table 4: Results from Data Cleaning

Data Cleaning Method	Silhouette Cosine	2D Visualization
Remove punctuation	0.053	
Retain punctuation	0.067	
Retain punctuation + With demoji	0.132	

Visualization of Cluster Separation

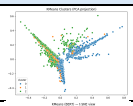
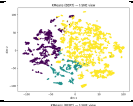
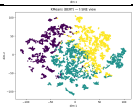
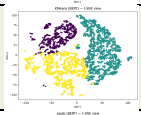
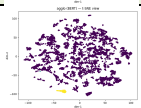
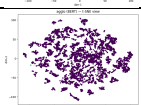
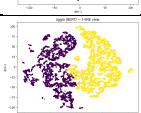
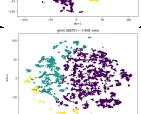
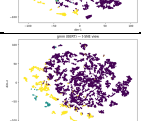
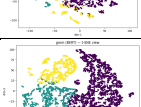
To visually assess whether the clusters exhibit meaningful separation, we applied t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the high-dimensional BERT embeddings into a lower-dimensional space.

Model results analysis

- Contextual embeddings significantly outperform TF-IDF

- K-Means clusters sentiment more clearly than hierarchical methods.
- Agglomerative clustering showed better silhouette cosine but weaker separation and more mixed sentiment clusters.
- cardiffnlp/twitter-roberta-base-sentiment-latest with K-Means delivered the best separation and purity across all models.

Table 5 : Model Results

Model	Silhouette Cosine	Cluster Purity	2D Visualization
TF-IDF with K-Means	0.132	0.514	
all-MiniLM-L6-v2 with K-Means	0.480	0.657	
all-mpnet-base-v2 with K-Means	0.310	0.630	
Twitter-RoBERTa with K-Means	**0.709**	**0.724**	
all-MiniLM-L6-v2 with Agglomerative	0.687	0.636	
all-mpnet-base-v2 with Agglomerative	0.898	0.630	
Twitter-RoBERTa with Agglomerative	0.749	0.636	
all-MiniLM-L6-v2 with GMM	0.434	0.630	
all-mpnet-base-v2 with GMM	0.383	0.630	
Twitter-RoBERTa with GMM	0.673	0.630	

Cluster vs Provided Sentiment Comparison

True positive counts are bold. Negative tweets dominate Twitter aviation discussions and confuse the clustering model to understand neutral sentiment.

Table 6 : Confusion Metrics

Cluster	Positive	Negative	Neutral
0	531	516	**1,898**
1	**5,818**	385	59
2	2,729	**2,151**	332

Sentiment Confidence Analysis

Misclustered tweets have lower sentiment confidence scores as compared to correctly clustered tweets. This shows that tweets with lower sentiment confidence scores contribute to more wrong clusters.

Table 7 : Confidence Analysis

Tweets	Mean Confidence
Correctly clustered	**0.932**
Misclustered	**0.816**

Visualization of Cluster Separation

- Three major clusters are visible and clean, indicating clear polarity distinctions. The negative and positive sentiment clusters form well-defined groups.
- Neutral tweets are more dispersed, consistent with their lower cohesion and higher ambiguity.
- The results visually support stronger clustering performance from contextual embeddings.

Figure 1. t-SNE 2D projection of K-Means clusters using **Twitter-RoBERTa** embedding

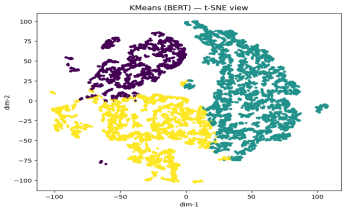
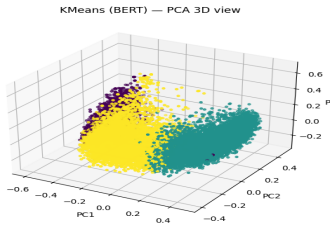


Figure 2. t-SNE 3D projection of K-Means clusters using **Twitter-RoBERTa** embeddings



Insights and Discussion

Results showed that contextual embeddings, especially Twitter-RoBERTa, capture sentiment more accurately than

statistical TF-IDF. The final clustering solution naturally reveals three distinct customer experience themes:

(1) High concentration of negative tweets indicates Strong dissatisfaction signals

Negative tweets dominate the dataset, mostly related to delays, cancellations and poor recovery. Twitter tends to be a popular channel used for customer service escalation. Topics surround unhappiness with airline communication.

Proactive Actions Recommended:
- Monitor spikes in complaints in real time
- Deploy auto-triage sentiment alerts to ground operations
- Improve communication speed during disruptions

(2) Neutral cluster includes factual updates and inquiries

Cluster 0 is largely neutral or low-emotion tweets (flight info, baggage status, questions).

Proactive Actions Recommended:
- Identify and monitor trending neutral topics
- Improve automation (chatbots) for informational tweets

(3) Positive sentiment cluster is small but meaningful

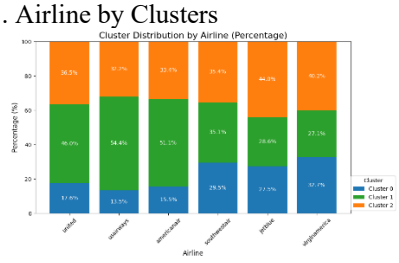
A distinct cluster captures praise, especially toward helpful staff and smooth flights. Positive feedback occurs when staff show empathy and quick service recovery succeeds or passenger expectations are exceeded

Proactive Actions Recommended:
- Scale up best practices
- Promote positive stories

(4) Airline Specific Concerns and Benchmark Against Industry

Based on the model results, the distribution of each airline by the different clusters is shown. JetBlue has the highest proportion of negative tweets and Virgin America has the highest proportion of positive tweets.

Figure 3. Airline by Clusters



Conclusion and Future Work

This project demonstrates that contextual embeddings, particularly Twitter-RoBERTa, combined with K-Means clustering are highly effective for uncovering sentiment patterns from airline tweets without using sentiment labels during training. The approach not only produced strong performance metrics (Silhouette 0.7094, Purity 0.7244) but also shows clear separation between positive, neutral, and negative sentiment groups, as supported by the t-SNE visualizations.

The imbalance of tweets with a heavy tendency towards negative sentiments (3:1:1 for negative, positive and neutral) could have impacted the model performance. Neutral tweets show higher ambiguity and can transition toward negativity if passengers do not receive timely information.

A real-time dashboard could be implemented by transforming unstructured customer feedback into meaningful sentiment clusters, where airlines can monitor customer sentiments in near real time, identify recurring pain points to improve communication and recovery actions before issues escalate. These insights reinforce the value of leveraging domain-specific NLP models for customer experience monitoring and operational decision-making.

Overall, this project shows that modern NLP and clustering techniques enable airlines to convert public social sentiment into actionable intelligence - enhancing responsiveness, protecting reputation and ultimately improving customers' end-to-end travel experience.

Appendices

ChatGPT (GPT-5) and Google Colab with AI powered by Gemini were used to support this project

References

Data source: Airline Sentiment Twitter Dataset (MSI5001 Project List)
Tools: Python, NLTK, scikit-learn, sentence-transformers, Hugging Face
Models tested: TF-IDF, all-MiniLM-L6-v2, all-mpnet-base-v2, twitter-roberta-base-sentiment-latest
Final combination: Twitter-RoBERTa + K-Means (cosine distance)
Visualization: t-SNE plots (for cluster separation), confusion matrix for sentiment alignmentDevlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL
Liu, Y.; Ott, M.; Goyal, N.; et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692