Airline Tweet Analysis Results

Subject: Airline Sentiment Analysis & Entity Extraction – Final Results & Recommendations

Hello,

I've completed the sentiment analysis and entity extraction project, achieving **93% accuracy** on the test dataset with an average runtime of 1s per tweet, making it production-ready.

Executive Summary

The system uses GPT-4o-mini to extract airlines and customer sentiment from 1,200 tweets. Instead of expensive fine-tuning, I achieved target accuracy through prompt engineering:

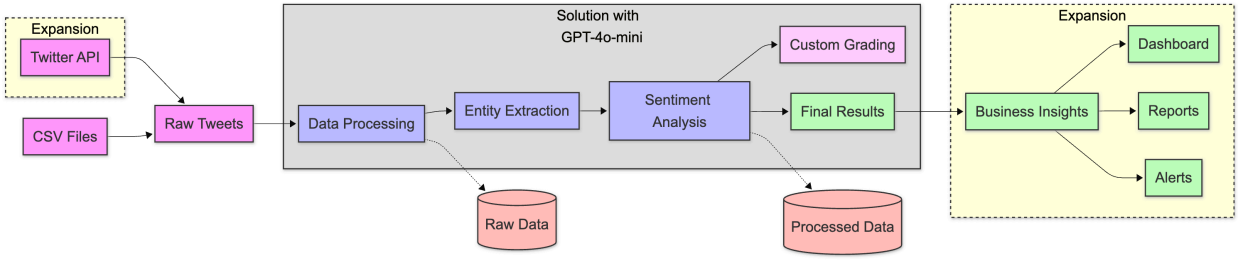
- 93% test dataset accuracy (94% on training data), exceeding the 90% target
- Handles 8 major airlines with name variations (abbreviations, handles, typos)
- Detects nuanced language such as sarcasm and subtext
- Production-ready with flexible processing options (e.g., batch size, dataset sampling, prompt selection)

Experiment	Accuracy
Entity Extraction, v1: Simple extraction	20%
Entity Extraction, v2: Predefined airline list	97%
Entity Extraction, v3: Few-shot examples	93%
Entity Extraction, v4: Custom grader	100%
Sentiment Analysis, v1: Simple analysis	90-93%
Sentiment Analysis, v2: Subtext and context	92-94%
Sentiment Analysis, v3: Custom grader	96%
Combined Analysis, v1: Simple	76%
Combined Analysis, v2: Refined prompt	90%
Combined Analysis, v3: Further refinement	96%
Full dataset (train data)	**94%**
Full dataset (test data)	**93%**

*See attached **OpenAI Evaluation Platform – Experiment Results** for further details.*

Technical Approach and Validation

I used OpenAI's Evaluations Platform with custom graders to address two challenges: sentiment subjectivity (especially distinguishing neutral from negative) and ground truth gaps (where ground truth labels missed secondary airline mentions). We set a target accuracy of 90% to account for inherent sentiment subjectivity.



The development process involved tackling entity extraction and sentiment analysis separately before combining them.

- Entity extraction handled airline name variation (abbreviations, Twitter handles, typos).
- Sentiment analysis handled subtle language like sarcasm and subtext. This was crucial, as 89% of tweets express negative sentiment, many with nuanced or implied negativity.

Business Insights and Next Steps

The dataset revealed clear patterns: United Airlines received the most negative sentiment, while Virgin America and JetBlue performed best. The 89% negative sentiment rate matches typical social media behavior where customers tweet more when experiencing problems. Notably, 97% of tweets mention a single airline.

The current solution handles individual tweets for maximum accuracy. For higher throughput in production (up to 50% faster), batching 5-10 tweets offers minimal accuracy loss.

To further improve accuracy and reduce risk of subjective evaluation criteria, consider refining sentiment classification guidelines to better distinguish between positive, neutral, and negative. This would also strengthen the foundation for future fine-tuning. Cost permitting, also consider experimenting with other models such as gpt-4o or o3-mini.

From a business perspective, this solution enables:

- Automatic categorization of customer feedback
- Early warning for customer service issues
- Competitive intelligence through customer comparisons

Airlines can leverage this to track and validate service improvements (e.g., response times, proactive issue detection, competitor benchmarks).

Expected impact: faster issue resolution, reduced churn through early intervention, and greater efficiency via automated routing and prioritization of complaints.

A potential expansion opportunity includes a full customer intelligence platform with expanded dataset (e.g., customer complaint data), real-time monitoring, sentiment alerts, and customer service system integration (e.g., ServiceNow).

Bottom Line

Prompt engineering proved effective for this dataset and lays a strong foundation for scalable customer intelligence.

I recommend moving forward with production deployment to capture immediate value and planning for a more robust platform to scale analysis.

I look forward to your thoughts.

Best regards,
Alexander Sergian