AI-Based Sentiment Analysis of English Customer Service Calls Using Audio and Textual Cues

Bader Ayman Al-Hindi - Kareem Omar Yousef - Abd Alrahman Wael Ayyash

Supervised by: Duaa Mehiar

Contents

Introduction & Problem statement

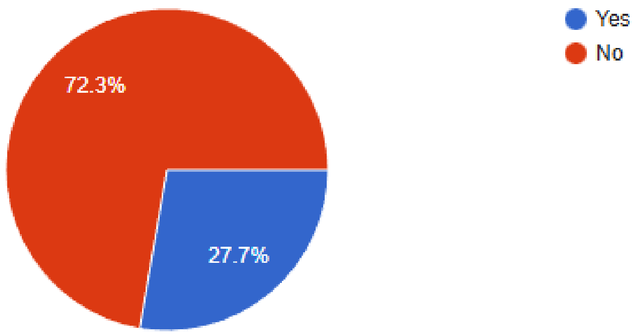
Customer service interactions are a primary communication channel between companies and customers, and customer satisfaction plays a critical role in service-oriented organizations. Customer satisfaction directly affects customer retention, brand reputation, and overall business performance. As a result, understanding customer satisfaction has become a key objective for modern organizations.

Traditional methods for evaluating customer satisfaction mainly rely on post-call surveys and manual review of recorded calls. However, these approaches suffer from several limitations. Surveys often receive low response rates and provide delayed feedback, while manual call analysis is time-consuming, subjective, and not scalable for large call volumes.

- Do you complete post-call satisfaction ratings? Yes / No
- Have you ever felt dissatisfied but did NOT submit a post-call rating? Yes / No
- During a customer service call, how do you usually express your feelings? Through the words I say / Through my tone of voice / Through both words and tone
- Would you prefer an automatic system that evaluates customer satisfaction without requiring you to fill a survey? Yes / No

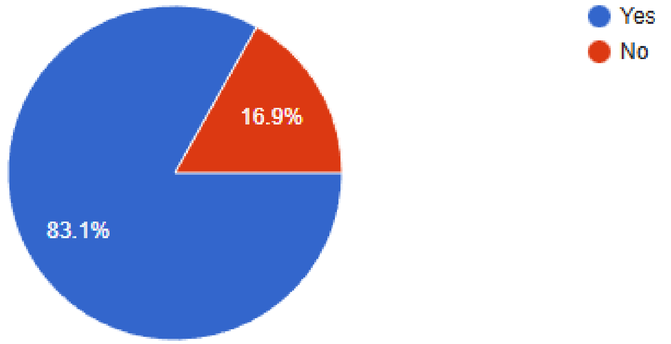
Do you complete post-call satisfaction ratings?

83 responses



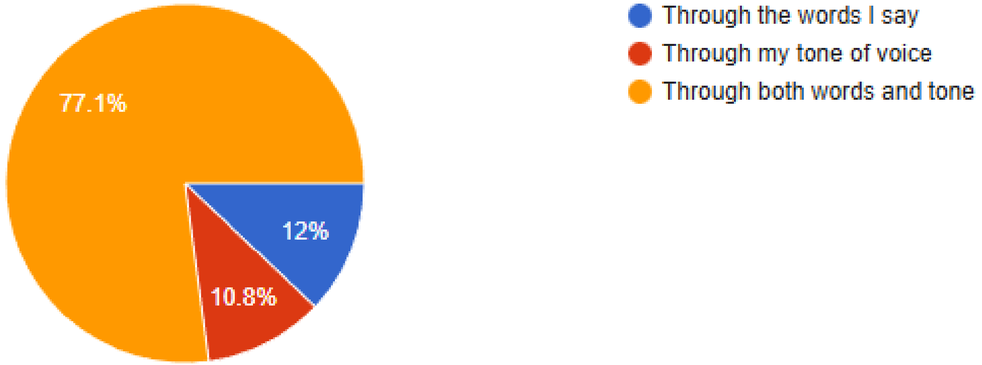
Have you ever felt dissatisfied but did NOT submit a post-call rating?

83 responses



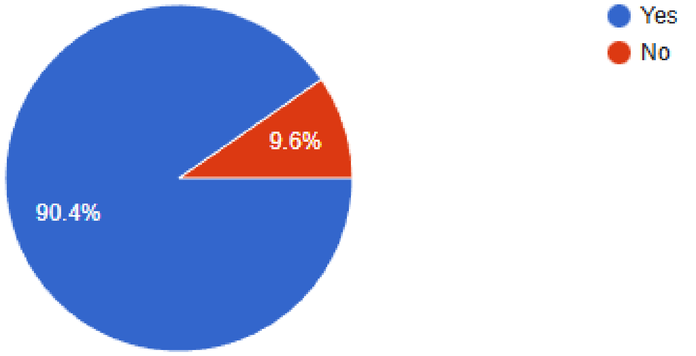
During a customer service call, how do you usually express your feelings?

83 responses



Would you prefer an automatic system that evaluates customer satisfaction without requiring you to fill a survey?

83 responses



Project Objectives

- Analyze customer satisfaction using speech signals by capturing emotional and paralinguistic cues
- Analyze customer satisfaction using transcribed text by extracting semantic and contextual information
- Design independent audio-based and text-based satisfaction analysis pipelines
- Combine both modalities using a multimodal fusion approach

Background

Sentiment Analysis in Customer Service

Sentiment analysis aims to identify opinions and emotional states from data and allows companies to understand how customers feel.

Customer satisfaction is often reflected indirectly, customers may not explicitly say they are unhappy, but their emotional tone and choice of words can indicate dissatisfaction.

Background

Sentiment Analysis in Customer Service

Because customer service centers handle a very large number of interactions every day, it is practically impossible to analyze all calls manually. Automated sentiment analysis provides a solution to this problem.

Unlike surveys, sentiment analysis allows organizations to observe customer reactions during the interaction itself rather than relying on delayed feedback.

Background

Speech Emotion Recognition (SER)

- Speech is a rich source of emotional information in human communication
- Emotional states are reflected through vocal characteristics such as:
  - Tone of voice
  - Pitch variations
  - Speaking intensity and rate
- Speech emotion recognition focuses on detecting these emotional cues directly from audio signals
- In customer service calls, speech can help identify dissatisfaction

Background

Text-Based Sentiment Analysis

- Text-based sentiment analysis focuses on understanding the meaning of customer utterances
- It is commonly applied to transcriptions of customer service calls or to text customer assistants
- Text analysis helps identify:
  - Customer intent
  - Complaints and requests
  - Explicit expressions of satisfaction or dissatisfaction
- Modern text-based approaches aim to capture contextual meaning rather than relying on keywords alone

Background

Challenges of Using a Single Modality

- Speech-only analysis:
  - Emotional expression differs across speakers, leading to inconsistent interpretation.
  - Lacks clear semantic explanation

- Text-only analysis:
  - Loss of vocal emotion, stress, and emphasis
  - Difficulty detecting sarcasm and indirect dissatisfaction
  - Neutral wording may hide negative customer experience

Multimodal Approach

Speech + Text

Multimodal analysis means using more than one type of information to understand a problem better. In the context of customer service interactions, satisfaction is expressed through both spoken language and vocal delivery. A multimodal approach leverages these different sources jointly, enabling a more comprehensive representation of customer experience compared to relying on a single modality.

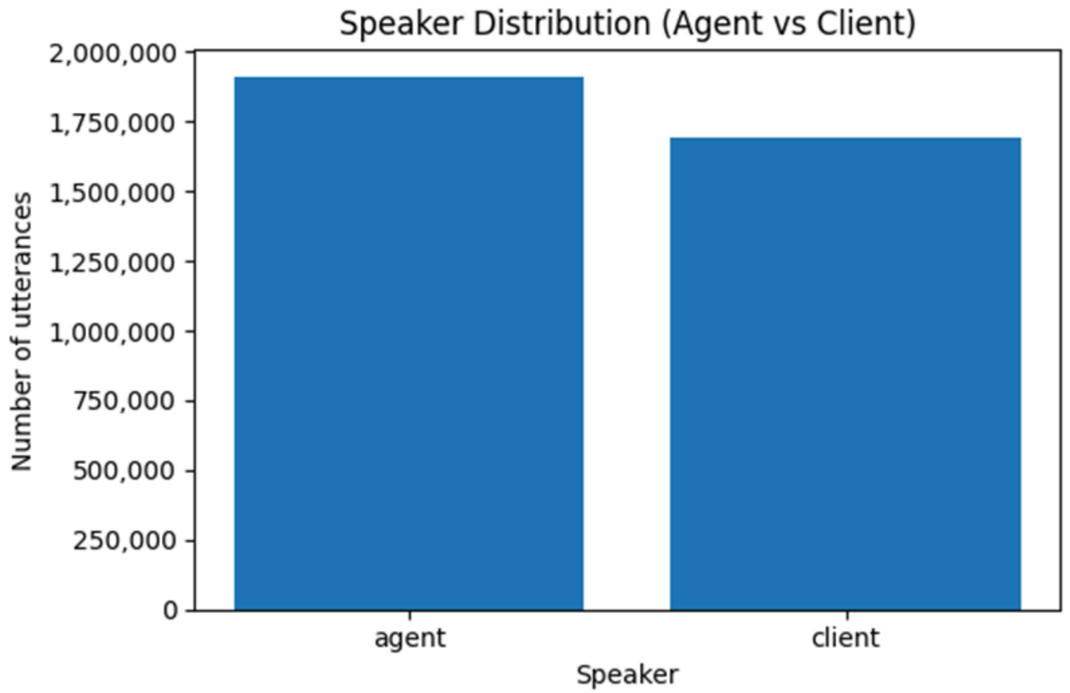
Multimodal Approach

Speech + Text

In a multimodal satisfaction analysis system, speech and text contribute complementary information. Speech captures emotional and paralinguistic cues that reflect how the customer feels during the interaction, while text provides semantic and contextual information that explains the content and intent of the customer’s message. By integrating these two perspectives, the system gains a more complete understanding of customer satisfaction.

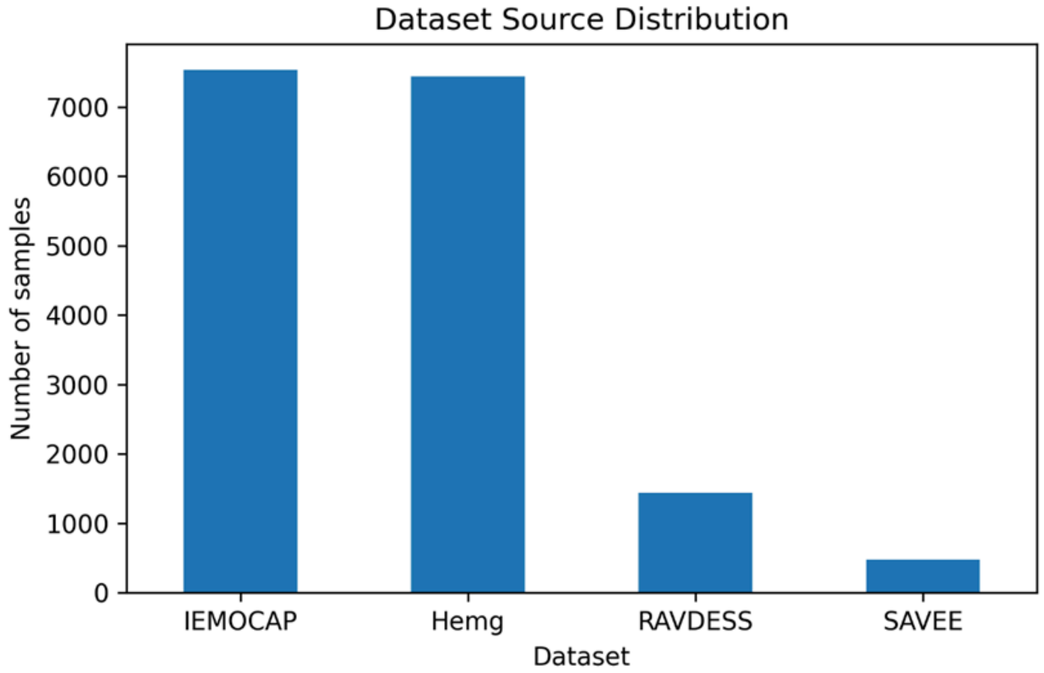
Datasets

For text, a large-scale telecom conversation dataset was used. This dataset consists of customer service interactions and includes satisfaction labels.



Datasets

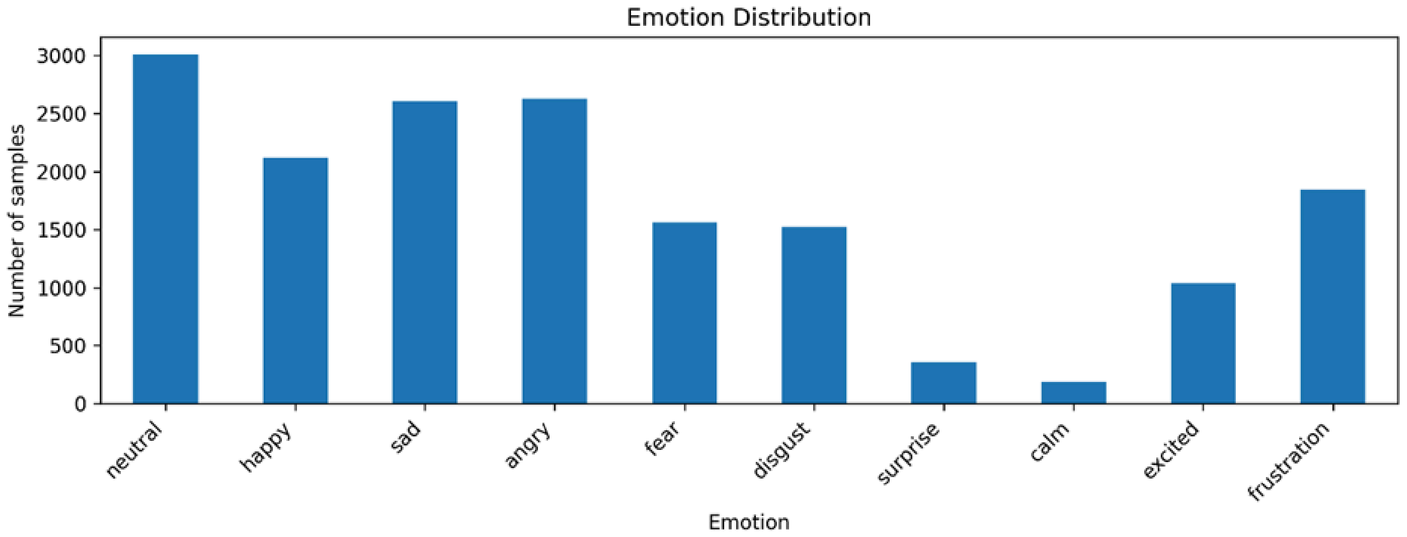
For audio, Multiple publicly available speech emotion datasets were used.
These datasets contain English speech samples with different emotional states.



Datasets

Audio Dataset Unification and Emotion Selection

To ensure consistency across datasets, all audio data was merged into a single unified corpus. Emotion labels from different datasets were mapped into a common set of ten emotion categories.



Datasets

Audio Preprocessing and Standardization

Before training, all audio samples were standardized to ensure uniform input conditions. Preprocessing steps included resampling audio to a common sampling rate, converting recordings to a single channel, and segmenting speech into shorter, consistent segments. These steps help reduce variability across datasets and improve model robustness during training.

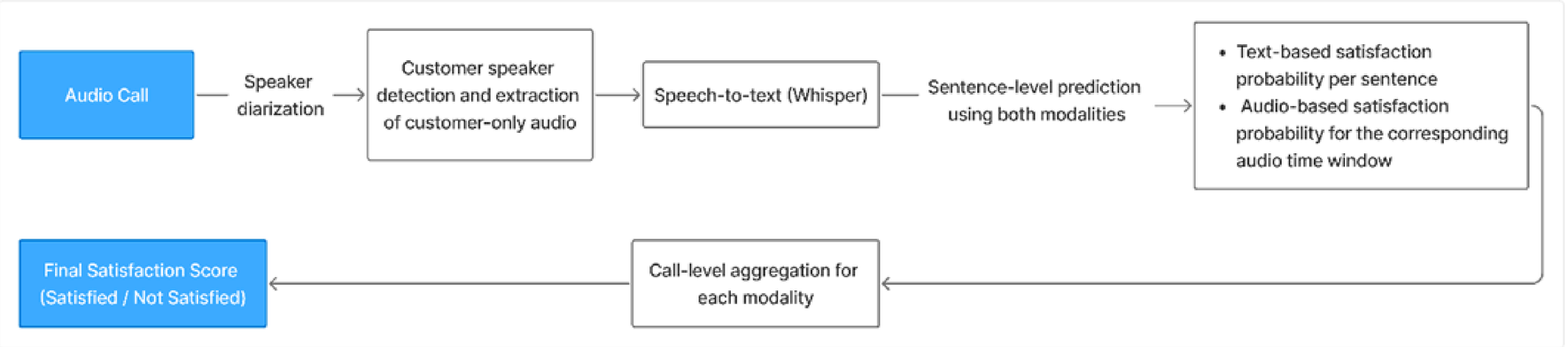
Datasets

Data Preparation

During the training phase, audio augmentation is applied to improve robustness under telephony conditions by simulating telephone bandwidth limitations (300–3400 Hz) and dynamic range compression using μ-law companding. Then the prepared customer speech is then processed through two parallel pipelines.

Pipelines & Multimodal Fusion

The proposed system is designed as a multimodal framework that processes customer service calls through two parallel pipelines: an audio pipeline and a text pipeline. Each pipeline analyzes the data independently and focuses on a different aspect of customer satisfaction. The outputs of both pipelines are later combined to produce a final multimodal satisfaction prediction.



Evaluation Metrics & Results

Evaluation Metrics

To evaluate the performance of the proposed system, standard classification metrics are used to measure how accurately customer satisfaction is predicted. These metrics provide insight into different aspects of model performance, including overall correctness and the ability to distinguish between satisfied and dissatisfied customers. Using multiple metrics ensures a fair and comprehensive evaluation of both audio-based and text-based models.

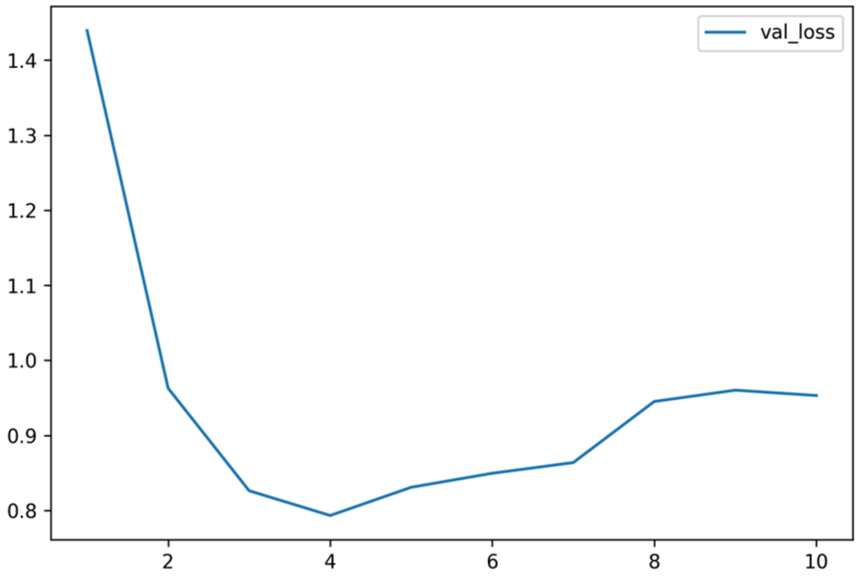
Audio-Based Results

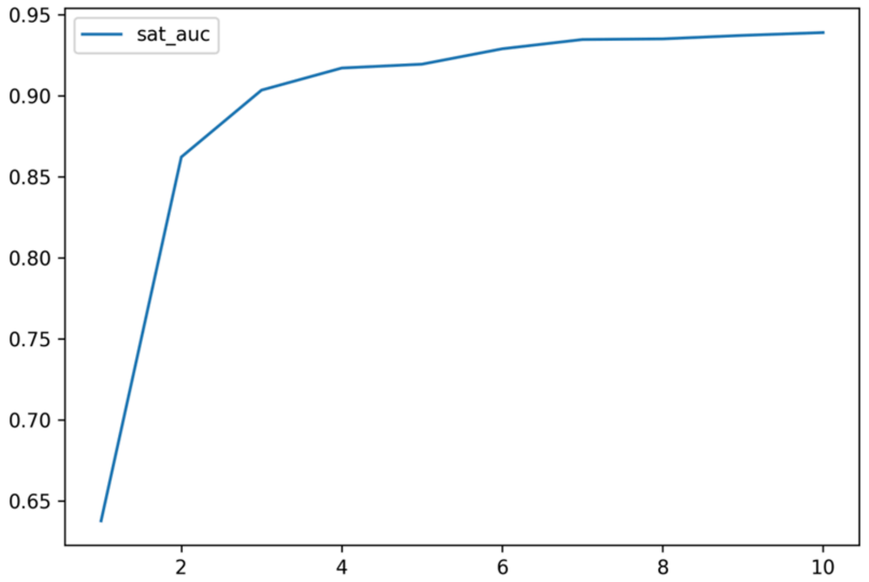
Speech emotion recognition achieved:

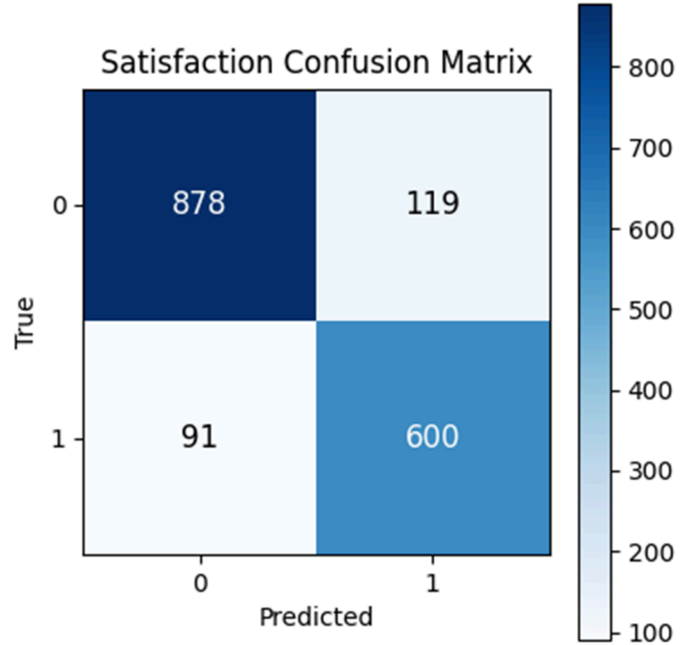
- Accuracy: 70.5%
- F1-score: 71.4% across 10 emotion classes

Speech-based customer satisfaction prediction achieved a high:

- AUC: 94%



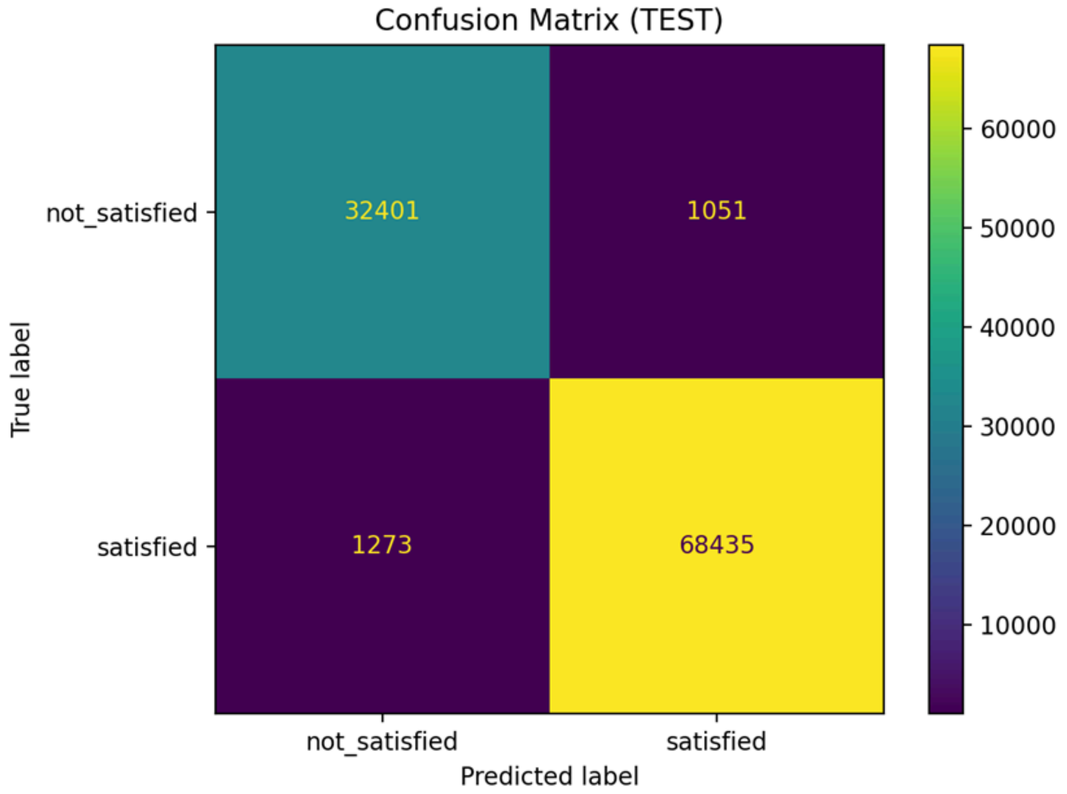




Text-Based Results

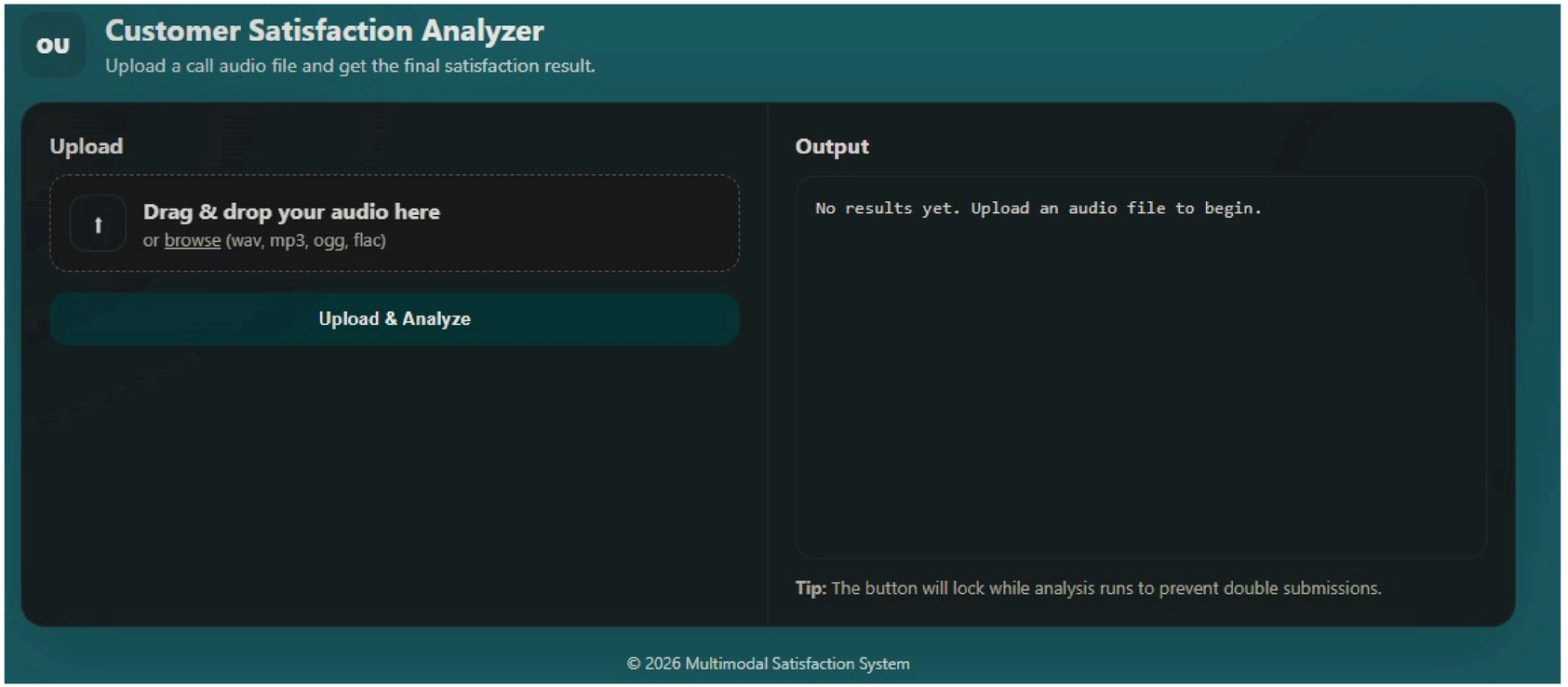
The text-based pipeline demonstrated very strong performance in predicting customer satisfaction from customer utterances
Satisfaction classification achieved:

- Accuracy: 97.7%
- Precision: 98.5%
- Recall: 98.1%
- F1-score: 98.3%



System Interface & System Overview

The system interface provides a simple way to interact with the customer satisfaction analysis system. Users can upload a customer service call audio file and initiate the analysis process. The interface then displays the predicted customer satisfaction results in a clear and understandable format.



System Overview

- The system begins when a user uploads a customer service call audio file through the interface.
- The system extracts the customer’s speech from the call to ensure that only customer-side information is analyzed.
- The audio is then resampled to 16 kHz, converted to mono, and normalized to match the requirements of pretrained speech models and ensure consistent input conditions.
- After standardization, the audio is segmented into shorter speech segments, and non-speech portions are removed to support sentence-level analysis.

- The audio pipeline analyzes speech to capture emotional cues related to satisfaction, while the text pipeline converts speech into text and extracts semantic and contextual information.
- Each pipeline produces a satisfaction prediction for each sentence segment. These predictions are aggregated and combined using equal weighting, where 50% of the final score comes from the audio pipeline and 50% from the text pipeline.
- The resulting multimodal prediction represents the overall customer satisfaction for the call and is displayed through the system interface as a clear satisfied or not satisfied output.

Conclusion

This project presented a multimodal system for analyzing customer satisfaction from customer service calls using both speech and textual information. The results indicate that integrating both modalities leads to strong and consistent performance. However, the system is currently limited by the use of acted speech datasets, dependence on speech transcription quality, and restriction to English-language data. Future work will focus on supporting the Arabic language, exploring deeper multimodal integration strategies, and evaluating the system on real-world customer service calls.

Thank You!
Any Questions?