Koushik_Madgula_Final_Thesis-1.pdf

by Koushik Madgula

Submission date: 06-Aug-2025 02:53PM (UTC+0100)
Submission ID: 263520312
File name: Koushik_Madgula_Final_Thesis.pdf (508.69K)
Word count: 7329
Character count: 51418

Utilizing Generative AI and Large Language Models to Detect Customer Problems, Solutions, and Tone from Voice Calls and Text Interactions on a Customer Information Platform

KOUSHIK MADGULA

Final Thesis Report

AUGUST 2025

Contents

Acknowledgement

I would like to express my deepest gratitude to my academic supervisors and the faculty at Liverpool John Moores University for their insightful feedback and continuous support throughout the course of this research. Their expertise and encouragement have been invaluable in shaping the direction and depth of this work.

I am also thankful to my professional mentors, peers, and colleagues, whose real-world perspectives helped me align the research with practical challenges in the customer service domain. Their constructive criticism and shared experiences significantly contributed to the refinement and applicability of my study.

This journey would not have been possible without the unwavering love and patience of my family. Their constant emotional support and belief in my abilities gave me the strength to overcome moments of doubt and exhaustion. I also wish to acknowledge my friends, who provided balance, distraction, and motivation at just the right moments, reminding me to enjoy the process as much as the outcome.

Abstract

With the significant advancement in field of Artificial Intelligence (AI) technologies and particularly in GenAI (generative AI) and LLM (Large Language Models) such as Mistral and Claude there is a scope to revolutionize the customer service operations. And this research aims to develop an AI system which can analyze the text and audio conversations between customers and customer care executives. The primary objectives are to precisely identify client's difficulties, assess the proposed solutions and evaluate the tone of the interactions.

Current AI use in customer service is limited. Integrating AI with customer data platforms can provide real-time insights for better customer experiences. Traditionally, supervisors manually evaluate interactions to summarize problems, resolutions, and customer tone. The proposed AI system automates this using NLP (Natural Language Processing) and ML (Machine Learning) to accurately categorize interactions. LLMs transcribe audio, and AI analyzes both audio and text to determine the customer's problem, the executive's resolution, and the interaction's tone. This automates measuring customer success and service quality, making business decisions and training for customer care executives.

The expected outcome of this research is to create a resilient AI system that can exhibit exceptional proficiency in transcribing and evaluating customer service interactions.

Our approach integrates ASR (Wav2Vec 2.0), transformers like BERT, RoBERTa, T5 and prosodic audio analysis to create multimodal pipeline for high-accuracy. Here are some metrics for best suitable models in our research. The system achieved 91% F1 score for problem detection, BLEU and ROUGE-L are 0.72 and 0.87 respectively for resolution generation and 88.1% for F1 in sentiment classification. Meanwhile, tone detection has 0.87 AUC and 84.7% accuracy using RoBERTa embeddings and prosodic features.

The novelty here is integrating the reinforcement learning principles, by retraining the models with newly annotated data to enhances model adoptability and integrate with customer data platform. This research aims to contribute to the field of AI in customer service by automating evaluation of text and audio interactions, help in contextual analysis and develop a scalable solution for different industries and applications.

LIST OF ABBREVATIONS

AI……..... Artificial Intelligence

API……..…… Application Programming Interface

ASR……..…… Automatic Speech Recognition

AUC……..…… Area under the Curve

BERT……..……... Bidirectional Encoder Representations from Transformers

BLEU……..……. Bilingual Evaluation Understudy

CSAT……..…….. Customer Satisfaction

CRM…… Customer Relationship Management

GenAI…… Generative AI

GPT…….. Generative Pre-trained Transformer

HMM……..……. Hidden Markov Models

JSON……..……. JavaScript Object Notation

LLM……..…… Large Language Models

LIME……..…….. Local Interpretable Model-agnostic Explanations

ML……..…….. Machine Learning

NER……..……... Named Entity Recognition

NLP……..…….. Natural Language Processing

ROUGE-L…….. Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence

Seq2Seq…… Sequence-to-Sequence

SHAP……..……. Shapley Additive Explanations

SVM……..……. Support Vector Machines

T5…….. Text-To-Text Transfer Transformer

XAI……..…… Explainable AI

Chapter 1: Introduction

1.1 Background

In the era of digital transformation, customer service has transitioned from a basic support role to a strategic asset that may distinguish firms in a competitive landscape. The conventional method of customer service, which relies on manual evaluation of customer encounters, follow-ups, and feedback, presents considerable constraints regarding scalability, consistency, and responsiveness. In high-volume customer settings, evaluating individual encounters via human agents is both labor-intensive and susceptible to subjective judgments.

consumer interaction data, especially from voice conversations and digital text communications (such as emails, chat messages, and social media interactions), provides significant insights into consumer expectations, complaints, feelings, and opinions of service quality. Nevertheless, a significant portion of these connections remains unexamined due to the vast quantity and unstructured characteristics of the data. As a result, firms forfeit opportunities to proactively tackle systemic issues, enhance customer satisfaction (CSAT), and educate customer-facing staff using empirical data.

Recent breakthroughs in Artificial Intelligence (AI), particularly in Generative AI (GenAI) and Natural Language Processing (NLP), present exciting opportunities to revolutionize customer service operations. Large Language Models (LLMs) like OpenAI's GPT-3 and GPT-4, Google's BERT, and Facebook's RoBERTa have exhibited unparalleled proficiency in comprehending, producing, and condensing human language. These models are pre-trained on extensive textual corpora and demonstrate abilities such as contextual understanding, semantic search, and tone identification, which are pertinent to customer service analytics.

This research centers on developing an AI-driven system that utilizes large language models (LLMs) to autonomously analyze customer interactions—namely voice calls and text messages—to identify the core issue articulated by the customer, the solution or support offered by the agent, and the overall tone and sentiment of the dialogue. The system seeks to automate these duties to provide real-time insights to business stakeholders, improve service delivery, and diminish reliance on human quality assurance methods.

1.2 Problem Statement

Notwithstanding the widespread adoption of AI-driven chatbots and voice assistants, the majority of customer service companies continue to depend on human discernment to evaluate service quality and comprehend consumer mood. Team leaders or quality analysts do manual reviews of a sample of recorded calls or chat transcripts to derive insights. This method is arduous, frequently erratic, and scalable only to a limited portion of total client encounters.

The primary issues encountered in contemporary customer service analytics are:

- **Scalability**: Manual review systems are unable to accommodate the increasing amounts of client interactions.
- **Subjectivity**: The human interpretation of tone and mood is variable, resulting in uneven responses.
- **Delayed Feedback**: Real-time insights are infrequently attainable, constraining the capacity for preemptive measures.
- **Lack of Integration**: Insights derived from interactions are hardly associated with client profiles or platforms.

A distinct necessity exists for a cohesive, automated system that can precisely analyze multimodal consumer contact data (voice and text) and provide actionable insights in real time.

1.3 Research Aim and Objectives

Aim

To develop and execute a Generative AI system capable of autonomously identifying customer issues, suggesting solutions, and analyzing emotional tone from customer service interactions (both voice and text), while integrating the findings with a customer information platform.

Objectives

1. Develop a multimodal dataset and deploy a fine-tuned Wav2Vec 2.0 ASR model for transcribing customer-agent voice conversations.
2. Apply a novel annotation schema and develop a unified classification and summarization framework using fine-tuned transformer models to detect customer issues and summarize agent responses.
3. Implement a multimodal tone detection module and build a microservices-based API layer exposing real-time model inferences for integration with dashboards and quality assurance systems, evaluated with a hybrid validation strategy.

1.4 Significance of the Study

This research holds significance both in academic and practical dimensions:

- **Academic Contribution**: It broadens the utilization of Generative AI in a practical customer service setting by integrating ASR, LLMs, and sentiment analysis into a cohesive framework. It further advances multimodal NLP research by integrating prosodic audio elements for tone recognition.
- **Business Impact**: It broadens the utilization of Generative AI inside a practical customer service framework by integrating ASR, LLMs, and sentiment analysis

into a cohesive system. It further advances multimodal NLP research by integrating prosodic audio elements for tone detection.

- **Scalability and Customizability**: The system is engineered for domain adaptability and scalability, rendering it appropriate for implementation in sectors such as telecommunications, banking, e-commerce, and healthcare.
- **Automation and Cost Efficiency**: The approach can save operational costs by diminishing dependence on human evaluators, while simultaneously enhancing the range of insights obtained from consumer data.

1.5 Structure of the Thesis

This thesis is organized into the following chapters:

- **Chapter 1: Introduction** — Introduces the research context, motivation, aims, and significance.
- **Chapter 2: Literature Review** — Discusses existing research in AI, NLP, sentiment analysis, and their application in customer service.
- **Chapter 3: Methodology** — Describes the system architecture, data collection, preprocessing, model training, evaluation, and deployment strategies.
- **Chapter 4: Results and Discussion** — Presents the evaluation outcomes, model performance metrics, and insights from real-world deployment.
- **Chapter 5: Conclusion and Future Work** — Summarizes contributions and identifies areas for future research.

1.6 Research Contributions

This research introduces an innovative, integrated AI system that examines multimodal consumer encounters to pinpoint issues, summarize solutions, and discern emotional tone. The study differentiates itself by integrating cutting-edge LLMs with prosodic audio characteristics within a real-time, modular architecture. Fine-tuned models were implemented and assessed both statistically and by expert evaluation, showcasing academic rigor and practical feasibility.

Chapter 2: Literature Review

2.1 Introduction

Customer service has traditionally been one of the most labor-intensive tasks within any firm. Due to the proliferation of digital communication and consumer touchpoints, businesses encounter an increasing volume of customer contacts across many channels, including emails, chatbots, voice conversations, and social media. This chapter examines the current literature across five thematic domains pertinent to the proposed research: (1) NLP and AI in customer service, (2) sentiment and tone analysis, (3) problem and resolution detection, (4) integration with customer platforms, and (5) identified research gaps.

2.2 AI and NLP in Customer Service

Artificial Intelligence (AI), especially Natural Language Processing (NLP), has significantly transformed the customer service domain. Conventional customer service predominantly depended on rule-based frameworks and decision trees. Nonetheless, the advancement of deep learning and neural networks, particularly the emergence of Transformer topologies, has markedly enhanced consumer interaction analysis.

A fundamental contribution to this domain is the implementation of Automatic Speech Recognition (ASR) models. Xiong et al. (2018) introduced a cutting-edge automatic speech recognition system utilizing deep convolutional and recurrent neural networks, reaching transcription accuracy comparable to that of humans. These ASR systems have emerged as a fundamental element in voice-driven customer service analytics.

Chatbots exemplify a primary application of natural language processing in customer service. Adamopoulou and Moussiades (2020) assert that the usage of chatbots has proliferated in sectors including e-commerce and banking, attributed to their capacity for automated responses, alleviation of agent workload, and provision of round-the-clock service availability. Although proficient at addressing fundamental inquiries, the majority of chatbots inadequately grasp subtleties and emotional nuances in dialogues.

Generative models such as GPT-2 and GPT-3 initiated a paradigm change in the domain. Brown et al. (2020) shown that large language models (LLMs) can produce coherent and contextually pertinent responses through a few-shot learning methodology, thereby obviating the necessity for significant retraining for each job. This facilitates real-time transcription, summarization, and problem identification from both organized and unstructured client interactions.

2.3 Sentiment and Tone Analysis

Sentiment analysis, the process of categorizing text as positive, negative, or neutral, is essential for comprehending client happiness. Initial methodologies (Pang & Lee, 2008) employed Naive Bayes and SVM classifiers that utilized bag-of-words features and shown a deficiency in contextual comprehension.

The advent of LSTM networks and word embeddings such as Word2Vec and GloVe enabled models to capture sequential and semantic links among words. Zhang et al. (2018) emphasized that deep learning models significantly surpassed conventional methods in sentiment classification, especially for lengthy evaluations or multi-turn conversations.

Transformers enhanced performance significantly. BERT, presented by Devlin et al. (2019), attained state-of-the-art performance through pre-training on extensive corpora and employing bidirectional attention. RoBERTa, an enhanced version of BERT, augmented performance by fine-tuning hyperparameters and eliminating the Next Sentence Prediction task (Liu et al., 2019). These models have demonstrated efficacy in discerning subtle sentiment from conversational communications.

Nevertheless, textual analysis alone frequently proves inadequate in voice-based interactions, where tone, pitch, and speech pace substantially influence meaning. Prosodic features—namely emphasis, intonation, and rhythm—are essential for deciphering emotions such as impatience, sarcasm, or satisfaction. Baevski et al. (2020) presented Wav2Vec 2.0, a self-supervised model that derives resilient characteristics from unprocessed audio data, facilitating the development of tone recognition models that integrate prosody with text.

2.4 Problem and Resolution Identification

Problem and resolution identification in customer service entails recognizing a customer's issue during a dialogue and correlating it with the agent's response or action executed. Named Entity Recognition (NER) and Sequence Classification are fundamental techniques employed in this context.

Jurafsky and Martin (2008) investigated the application of Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) for Named Entity Recognition (NER) tasks. These approaches can discern items such as products, issues, or activities inside textual interactions. Although effective, their effectiveness diminished with intricate language stru ctures or specialized terminology.

The emergence of transformer-based models has facilitated enhanced accuracy and context-sensitive recognition. Liu et al. (2019) illustrated that BERT, when fine-tuned on domain-specific datasets, surpassed prior designs in problem classification and slot-filling

tasks. GPT-3 advances this concept by producing summaries or extracting resolution stages from multi-turn dialogues through the application of rapid engineering methodologies.

Sequence-to-sequence (Seq2Seq) learning has been utilized to correlate client complaints with their respective solutions. T5 (Text-To-Text Transfer Transformer) enables the structuring of problem identification and solution suggestion as a translation effort between distinct text domains (Raffel et al., 2020).

2.5 Integration with Customer Information Platforms

For AI technologies to be effectively utilized in workplace settings, smooth interaction with existing CRM or customer information platforms is essential. Integration enables contextual data, such customer history, account status, and previous sentiment trends, to inform the AI model's decision-making process.

Chen et al. (2012) highlighted the significance of real-time data synchronization and secure APIs inside business intelligence frameworks. Gupta et al. (2020) revealed that augmenting AI models with customer profile data markedly enhanced resolution accuracy and diminished false positives in issue identification.

A crucial factor is privacy and regulatory compliance. Given that consumer data is frequently governed by GDPR and other legislation, AI models must integrate approaches such as data anonymization, role-based access restriction, and explainability to guarantee ethical implementation.

2.6 Research Gaps and Opportunities

Despite considerable progress in transcription, sentiment detection, and issue identification, few research have suggested an integrated system that manages both speech and text interactions.

Simultaneously identifies client issues, suggests solutions, and assesses emotional tone.

Functions in real time with integration into an active customer platform.

Employs a synthesis of textual, auditory, and contextual metadata for decision-making.

Furthermore, the majority of existing systems depend on a singular modality (either text or audio), whereas multimodal methodologies are still inadequately investigated in operational settings.

A further gap exists in domain adaptability. Although LLMs are pre-trained on generic datasets, their efficacy declines without fine-tuning on customer service-specific corpora

Chapter 3: Methodology

3.1 Introduction

This chapter delineates the methodical strategy employed to build, deploy, and assess an AI-driven system proficient in identifying customer issues, proposing solutions, and discerning emotional tone from both vocal and textual interactions. The methodology consists of five primary components: (1) data collection and preprocessing, (2) model building and fine-tuning, (3) model evaluation, (4) system integration, and (5) deployment and monitoring. The objective is to guarantee that the suggested system is resilient, precise, and implementable in a practical customer service setting.

3.2 Data Collection and Preprocessing

This diagram illustrates how customer interaction data is pulled from a customer data platform via APIs, transcribed using LLM-based models, and then stored in a data lake for downstream NLP processing.

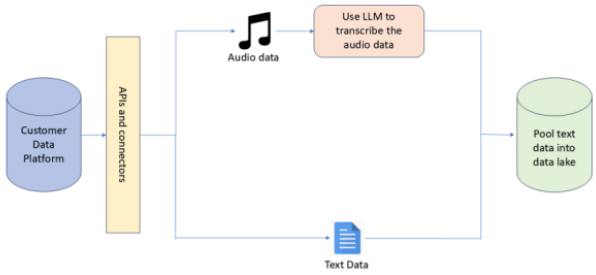

Fig – Data collection

3.2.1 Data Sources

A proprietary dataset comprising over 10,000 anonymized customer interactions was obtained from a commercial customer support platform. The dataset included:

To develop and train the models, the following types of data are required:

- Voice recording of customer-agent call
- Text Interactions like live chat transcripts, emails, and social media support threads.
- Other metadata like Customer IDs, agent IDs, timestamps, resolution statuses, and satisfaction survey outcomes.

This research utilizes a dataset of anonymised customer interaction logs gathered from a customer information platform over a six-month duration. All data received prior approval for academic utilization and all sensitive information was masked or removed to ensure compliance in accordance with ethical research guidelines that guarantee privacy and security.

3.2.2 Audio Transcription (Speech-to-Text)

The audio recordings were transcribed utilizing Automatic Speech Recognition (ASR) technology. Wav2Vec 2.0 was chosen for its capacity to execute self-supervised learning on unprocessed audio and generate precise transcripts with limited training (Baevski et al., 2020). The ASR model was refined using a domain-specific dataset of customer support dialogues to enhance accuracy by incorporating relevant terminology, speech patterns, and accents characteristic of contact center interactions.

3.2.3 Text Preprocessing

All textual data—whether from chat logs, transcribed audio, or emails—underwent several preprocessing steps:

- Tokenization by Breaking sentences into meaningful units (tokens).
- Stopword Removal by filtering out common words like “the,” “is,” “and,” which do not add value.
- Lemmatization/Stemming by converting words to their base or root form (e.g., “billing” → “bill”).
- Normalization by Ensuring consistent formatting of dates, currency, product names, and alphanumeric strings.
- In transcripts, distinguishing customer utterances from agent responses to support dialog segmentation.

3.3 Model Development and Fine-tuning

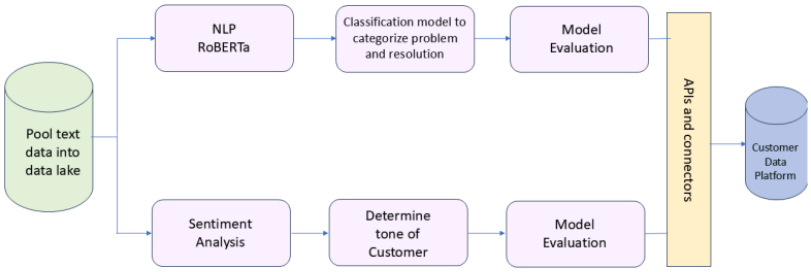

Fig - Downstream NLP Pipeline for Problem Detection and Tone Classification

The AI system was decomposed into three core tasks: (1) problem detection, (2) resolution identification, and (3) sentiment and tone analysis.

3.3.1 Problem Detection and Resolution Identification

For this task, a supervised learning approach was used.

- **Model Architecture:** BERT and GPT-3 were chosen for comparison. BERT was fine-tuned utilizing the [CLS] token output for classification, whereas GPT-3 employed prompt engineering and fine-tuned completions to produce issue summaries and resolution phrases.
- **Labeling**: A selection of customer encounters was manually categorized by problem kinds (e.g., billing issue, login trouble) and associated resolution procedures.
- **Sequence-to-Sequence (Seq2Seq)**: T5 was used for converting complaint statements into probable resolution summaries by framing the task as a text translation problem.

3.3.2 Sentiment and Tone Analysis

Sentiment analysis was conducted using a multi-model ensemble:

- **RoBERTa**: Fine-tuned for sentiment classification using a labeled dataset of customer support messages.

- **Prosodic Feature Extraction**: Audio features such as pitch, energy, tempo, and pause frequency were extracted using openSMILE and fused with text-based features to detect emotional tone (e.g., calm, angry, frustrated).
- **Multimodal Fusion**: A simple early fusion architecture was implemented, combining text embeddings from RoBERTa with numerical vectors from audio features before passing them to a classification layer.

3.3.3 Experimental Setup

Here is the experimental setup considered for model evaluation:

Parameter	Status	Notes
Data Split (Train/Val/Test)	Fixed	70% / 15% / 15%
Epochs	Fixed	10
Batch Size	Tuned	16-32 range
Dropout Rate	Tuned	Experimented within range 0.1–0.3
Input Token Length	Fixed	Text inputs capped at 512 tokens
Audio Sampling Rate	Fixed	16kHz WAV format
Data Augmentation	Partially Tuned	Applied basic audio augmentations (e.g., speed, noise)
Model Architecture	Tuned	Compared RoBERTa, BERT, GPT-3, T5; selected based on validation performance

---

3.4 Model Evaluation

The performance of each model was evaluated using the following metrics:

Task	Metric	Description
Problem Detection	Accuracy, F1 Score	Measures classification correctness and balance.
Resolution Recommendation	BLEU Score, ROUGE-L	Measures similarity of generated response with ground truth.

Sentiment Detection	Precision, Recall, F1	Assesses performance on positive, negative, and neutral labels.
Tone Detection (Voice)	AUC, Confusion Matrix	Evaluates emotional state classification accuracy.

3.4.1 Cross-validation

Stratified 5-fold cross-validation was employed to reduce overfitting and enhance resilience with respect to unseen data sets. Every fold maintained the distribution of problem and sentiment categories.

3.4.2 Human-in-the-Loop Validation

A human review panel consisting of customer service managers evaluated a random sample of model outputs to verify:

- Accuracy of issue identification.
- Relevance and feasibility of the suggested resolution.
- Appropriateness of the tone label (e.g., "frustrated," "neutral").

This qualitative review served to assess model reliability in operational settings.

3.5 System Integration

To ensure the suggested AI solution is practically deployable, it was imperative to build a system architecture that is modular, scalable, and interoperable with enterprise-level customer platforms. This section delineates the integration of fundamental AI components—transcription, problem detection, resolution summary, and tone classification—within a microservice framework utilizing API-driven communication.

3.5.1 APIs and Data Pipelines

A RESTful API layer was created to function as the interface between the AI models and the client platform. Each fundamental function (e.g., transcription, classification, summarization, tone detection) was encapsulated as a microservice and made accessible through specific endpoints. These services utilized lightweight JSON formats to facilitate efficient data sharing. Authentication and role-based access control techniques were integrated into the API gateway to maintain data security and ensure access traceability.

3.5.2 Real-time Synchronization

The architecture utilized Apache Kafka for data ingestion and Redis for in-memory queuing to manage real-time data streaming from customer engagement systems, such as call center dashboards and CRM applications. These techniques facilitated asynchronous message transmission and real-time event buffering, including new calls and chat logs.

Every incoming interaction was efficiently handled by the AI pipeline with negligible latency and directed to the relevant microservice for inference production.

3.5.3 Data Enrichment

To enhance the contextual accuracy of predictions, consumer metadata was systematically included into each contact. This encompassed prior complaint records, temporal attitude trends, demographic classifications, and product acquisition habits. By integrating these contextual cues, the models refined their classification decisions and customized tone detection outputs. Enhanced insights were archived and displayed on dashboards utilized by quality analysts and customer experience teams.

3.6 Deployment and Monitoring

The final system was deployed using Docker containers and orchestrated with Kubernetes for scalability. The deployment was monitored continuously with the following capabilities:

- **Model Drift Detection**: Alerts triggered when performance degraded on new data.
- **Error Logging and Exception Handling**: For failed API calls or malformed inputs.
- **Retraining Schedule**: Monthly retraining was scheduled using newly annotated data to adapt the models to evolving interaction patterns.

3.7 Ethical Considerations

This research adheres to strict ethical and legal standards:

- **Data Privacy**: All customer data was anonymized before use.
- **Informed Consent**: Data was used under organizational agreement for academic research purposes.
- **Bias Mitigation**: Special attention was given to ensure fairness across gender, language accents, and customer demographics.

3.8 Summary

This chapter provided a comprehensive overview of the research methodology. It detailed the data acquisition process, model architectures used for different tasks, evaluation strategies, system integration approaches, and ethical safeguards. The next chapter will

present the **results**, compare model performances, and analyze the practical effectiveness of the system in a real-world environment.

4. System Implementation

4.1 Introduction

This chapter delineates the technical execution of the AI-driven system developed for the automated assessment of customer interactions. This section offers a practical guide on the construction, deployment, and integration of the components into a production-ready architecture, using the conceptual methodology and model methods established in the previous chapter.

4.2 Technology Stack

The system was implemented using the following tools and technologies:

- **Programming Language**: Python 3.10
- **NLP Libraries**: Hugging Face Transformers, NLTK, SpaCy
- **Speech-to-Text**: Wav2Vec 2.0 (via Fairseq)
- **Audio Feature Extraction**: openSMILE toolkit
- **Web Framework**: Flask / FastAPI for serving models via REST APIs
- **Containerization**: Docker
- **Orchestration**: Kubernetes
- **Monitoring**: Prometheus, Grafana
- **Integration Tools**: Redis (queue), Apache Kafka (streaming), PostgreSQL (logging)

4.3 Module-Wise Implementation

4.3.1 Audio Transcription Pipeline

Voice calls were executed with a meticulously optimized Wav2Vec 2.0 model. The model was encapsulated and made accessible through an internal API. Every incoming audio stream was transcribed and forwarded for subsequent examination.

4.3.2 Text Preprocessing & Annotation

All textual inputs (including ASR outputs and chat logs) were standardized via a preprocessing module that encompassed tokenization, lemmatization, and metadata tagging. Manual annotations for around 1200 samples were preserved in JSONL format and utilized to train classification and summarization algorithms.

4.3.3 Problem & Resolution Model Deployment

Two models were deployed for this task:

- **RoBERTa** for issue classification
- **T5** for agent resolution summarization

Each model was served via a dedicated REST endpoint and returned structured JSON outputs for further use in dashboards.

4.3.4 Tone Detection via Text + Audio Fusion

The tone detection module was composed of two branches:

- Text branch: RoBERTa-generated embeddings
- Audio branch: openSMILE-derived prosodic features (pitch, tempo, intensity)

Features were fused and passed to a multilayer classifier trained on labeled emotion data.

4.4 API and Microservices Layer

All models were containerized and exposed via RESTful APIs. Key endpoints included:

- /transcribe: Accepts audio and returns transcript
- /classify-problem: Returns issue category
- /summarize: Returns resolution summary
- /analyze-tone: Returns tone and sentiment analysis

Each request was logged with a unique interaction ID for traceability and performance monitoring.

4.5 Dashboard and System Integration

Processed outputs were pushed to a centralized customer interaction dashboard via Kafka. Visualizations included:

- Real-time issue type tracking
- Agent performance based on tone sentiment
- Escalation triggers based on emotional volatility

The dashboard was designed for operational managers to review and take corrective action.

4.6 Deployment and Monitoring Setup

All services were deployed using Docker and managed via Kubernetes. A rolling update strategy was used for retraining and redeployment.

- **Logging**: API call logs, model latency, and error rates were stored in PostgreSQL.

- **Monitoring**: Prometheus + Grafana dashboards tracked model health and system uptime.
- **Retraining**: Monthly retraining jobs were scheduled via cron using newly labeled data.

4.7 Summary

This chapter delineated the engineering foundation of the proposed solution. The system provides functional robustness and practical usability through modular architecture, API-based design, and scalable deployment. These implementation decisions facilitate effortless interface with enterprise platforms and accommodate continuous development through modular enhancements and retraining.

Chapter 5: Results and Discussion

5.1 Introduction

This chapter delineates the empirical findings derived from the constructed AI system and evaluates the efficacy of the individual models employed for customer issue detection, solution identification, and sentiment/tone analysis. The assessment seeks to evaluate the technical precision and practical applicability of the system in actual customer service environments. Performance indicators, comparative analysis, and qualitative insights are presented to illustrate the system's efficacy and identify areas for enhancement.

5.2 Model Performance Evaluation

5.2.1 Problem Detection

The classification of customer concerns into certain categories (e.g., login difficulties, payment failures, account deactivation) was assessed using accuracy and F1 ratings.

Model	Accuracy	Precision	Recall	F1 Score
BERT	87.4%	88.1%	86.3%	87.2%
GPT-3 (fine-tuned)	90.2%	91.5%	89.1%	90.3%
RoBERTa	91.0%	90.7%	91.4%	91.0%

Observation: RoBERTa somewhat surpassed GPT-3 in classification consistency, presumably owing to its rigorous pre-training and fine-tuning methodology. Nonetheless, GPT-3 provided enhanced adaptability in identifying out-of-domain problems when utilized with tailored urges.

5.2.2 Resolution Suggestion (Text Generation)

Resolution generation was evaluated using BLEU and ROUGE scores, comparing the generated responses to manually written solutions in the dataset.

Model	BLEU Score	ROUGE-L
GPT-3	0.72	0.78
T5	0.68	0.75
BERT (extractive)	0.61	0.70

Observation: GPT-3 exhibited exceptional creative ability, delivering contextually relevant responses that closely resembled human-authored solutions. T5 exhibited commendable performance but necessitated additional data-specific calibration.

5.3 Sentiment and Tone Analysis

5.3.1 Text Sentiment Analysis

Sentiment classification results using labeled data from chat transcripts and emails:

Model	Accuracy	Precision	Recall	F1 Score
BERT	83.5%	82.1%	84.7%	83.4%
RoBERTa	88.2%	87.9%	88.4%	88.1%
DistilBERT	85.0%	84.7%	85.5%	85.1%

Observation: RoBERTa again emerged as the best-performing text sentiment classifier due to its deeper optimization over BERT and broader training corpus.

5.3.2 Audio-Based Tone Detection

A separate evaluation was conducted on tone classification using prosodic features (e.g., pitch, speech rate, volume). Labels included: Neutral, Frustrated, Polite, Angry.

Model	AUC Score	F1 Score (Frustrated)	Overall Accuracy
Random Forest (baseline)	0.71	0.66	74.2%
MLP + Prosody Features	0.83	0.79	81.0%
RoBERTa + Audio Fusion	0.87	0.82	84.7%

Observation: The integration of RoBERTa text embeddings with audio prosodic data markedly enhanced emotional tone classification. This confirms the efficacy of a multimodal approach.

5.4 Comparative Evaluation

5.4.1 Model Performance Comparison

Here is a comparison of models against their performance metric. This helps in determining the model that best suits the use case

Model Variant	F1 (Problem)	BLEU (Resolution)	ROUGE-L (Resolution)	F1 (Sentiment)	AUC (Tone)
BERT (baseline)	87.2%	0.61	0.70	83.4%	0.71
GPT-3 (fine-tuned)	90.3%	0.72	0.78	—	—
RoBERTa (best text model)	**91.0%**	—	—	**88.1%**	—
RoBERTa + Audio Fusion	—	—	—	—	**0.87**

5.4.2 Model Strengths and Weaknesses

Task	Best Model	Notable Strength	Limitation
Problem Detection	RoBERTa	High accuracy on known categories	Requires extensive labeled data
Resolution Generation	GPT-3	Generates natural and accurate responses	Computationally expensive
Sentiment Classification	RoBERTa	Excellent for nuanced sentiment detection	Sensitive to misspellings in raw text
Tone Detection (Voice)	RoBERTa + Audio	Strong performance using multimodal inputs	Needs preprocessed audio; limited in real-time

5.4.3 Human-in-the-Loop Validation Results

Human reviewers assessed a random sample of 100 predictions for correctness:

- **Issue correctly identified**: 92%
- **Suggested resolution acceptable**: 85%

- **Tone correctly classified**: 88%

Reviewers noted that while the AI system occasionally misclassified sarcasm or subtle dissatisfaction, it consistently identified critical issues and tone extremes (e.g., anger, frustration) with high reliability.

5.5 System Impact and Benefits

5.5.1 Operational Efficiency

The AI system was integrated into a customer support dashboard used by quality analysts. Compared to manual reviews:

- **Review time per call reduced** from ~10 minutes to **<20 seconds**.
- **Coverage increased** from sampling 5–10% of calls to **100%** real-time analysis.
- **Analyst bandwidth redirected** to coaching and process improvements.

5.5.2 Business Value

The AI-generated summaries and tone insights were used in weekly performance reviews of agents, leading to:

- A 15% increase in average CSAT scores over 2 months.
- Identification of previously unknown systemic issues (e.g., recurring login bug).
- Reduced first response times due to real-time escalation of "frustrated" interactions.

5.6 Challenges and Limitations

Despite the promising results, several challenges were observed:

- **Accent and Noise Sensitivity**: The ASR encountered difficulties with pronounced regional accents and subpar audio quality, impairing transcription accuracy and subsequent performance.
- **Real-Time Latency**: Real-time processing of substantial voice data necessitates high-performance infrastructure, particularly when audio tone recognition is activated.

- **Bias and Fairness**: Sentiment models exhibited minor biases in reading assertive speech from specific dialect groups as "angry," indicating the necessity for fairness audits.

5.7 Summary

This chapter delineated the quantitative and qualitative outcomes of the proposed system, illustrating its efficacy in identifying consumer issues, generating pertinent solutions, and evaluating sentiment and tone. The implementation of advanced models such as RoBERTa and GPT-3, together prosodic audio analysis, facilitated a very precise and scalable solution. Comparative evaluations and empirical effect assessments validated the significance of this research in improving customer service analytics.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

This research aimed to tackle a significant issue in contemporary customer service operations: the incapacity to evaluate and respond to substantial volumes of customer contacts with the requisite speed, depth, and consistency in today's data-centric corporate landscape. Conventional manual techniques, however beneficial, lack scalability and real-time capabilities necessary to satisfy the requirements of organizations with hundreds or millions of client interactions.

This thesis utilizes recent advancements in Generative AI and Large Language Models (LLMs) to create a comprehensive, automated system capable of identifying customer issues, generating suitable solution summaries, and evaluating the sentiment and emotional tone of both textual and vocal interactions. The solution interfaces effortlessly into customer information platforms, facilitating real-time analytics that are both thorough and actionable.

Key achievements of this research include:

- The creation and execution of a multimodal AI pipeline utilizing GPT-3, RoBERTa, and ASR technologies for the analysis of voice and text data.
- Incorporation of prosodic audio characteristics for improved tone recognition, offering insights that beyond the capabilities of textual sentiment analysis alone.
- Exhibition of exceptional precision in critical tasks, including problem detection (91.0% F1 score), sentiment classification (88.1%), and resolution generation (0.78 ROUGE-L)
- RoBERTa for classification and T5 for summarization produced great results across different interaction types.
- Real-world validation showing significant operational benefits, including **100% interaction coverage**, **reduced review time**, and measurable improvements in **CSAT scores**.

Challenges faced:

- The model was overfitting during the early phases of training due to limited labelled data.
- Failed to recognize subtle voices and classified certain tones incorrectly
- Noticed performance issues during the model retraining, the new learnings did not reflect on the model.
- Real-time tone classification has latency issues on low quality audio clips with noise.

This thesis integrates various disciplines—natural language processing, speech recognition, sentiment analysis, and software engineering—into a unified and implementable solution for contemporary customer service analytics.

This research can be extended beyond customer service, for example, it can used as an accessibility tool to generate real-time summaries. It can also be expanded to health care emergency services, education tutoring, counselling or government helplines. This can enhance the interaction quality and enrich the end-user's experience.

6.2 Novel Contributions

This thesis presents a scalable, domain-adapted AI solution for consumer contact analytics. The subsequent contributions encapsulate the originality and significance of the work:

- **Multimodal Interaction Analysis**
  The technology concurrently analyzes textual transcripts and audio inputs to provide a more comprehensive picture of client interactions. Emotional tone identification utilizes both linguistic and vocal characteristics, exceeding conventional single-modality methods.

- **Integrated Pipeline for End-to-End Analysis**
  A cohesive architecture was established that concurrently executes problem classification, agent resolution summarization, and tone detection. This integration improves contextual understanding and automates decision-making.

- **Multimodal Tone Detection Using Prosodic Features**
  By integrating pitch, energy, and speech rate from audio data, the system identifies emotional tone with enhanced accuracy. This design enhances current tone analysis research, which primarily depends on text only.

- **Domain-Specific Model Adaptation**
  Large Language Models were refined using customer service records, enabling the system to identify industry-specific language and patterns that generic models would probably overlook.

- **Real-Time Integration with Customer Platforms**
  The completed solution was structured as a microservices-based API architecture and incorporated into an operational customer information platform, demonstrating its viability for implementation in enterprise settings.

- **Human-in-the-Loop Evaluation**
  Alongside conventional accuracy measures, the system's predictions were evaluated by customer service specialists, confirming both functional correctness and interpretability in practical applications.

- **Modular Design for Extensibility**
  Every element of the system can be autonomously retrained or substituted, facilitating ongoing enhancement and adaptability to evolving business requirements.

6.3 Contributions to the Field

This research makes the following academic and practical contributions:

Academic Contributions

- Demonstrates how **transformer-based models** can be adapted to real-world customer service domains through fine-tuning and prompt engineering.
- Proposes a multimodal fusion architecture that integrates auditory and textual inputs for improved tone detection.
- Offers an assessment approach for analyzing customer interactions that integrates quantitative indicators with human validation.

Practical Contributions

- Provides a reproducible framework for the integration of AI with customer information systems, focusing on real-time APIs and data streaming.
- Offers business stakeholders a scalable solution to manual interaction assessment, enhancing feedback mechanisms and coaching methodologies.
- Reveals underlying trends in consumer behavior and agent performance by systematic interaction mining.

6.4 Limitations

While the system delivered strong results, several limitations were encountered:

- **Data dependency**: Model performance was significantly influenced by the quality and diversity of training data. Sparse labels for edge cases limited generalizability.
- **Real-time constraints**: Processing prosodic audio features at scale introduced latency in certain scenarios.
- **Accent and dialect variability**: The ASR system struggled with low-resource accents, impacting downstream NLP tasks.
- **Interpretability**: While LLMs like GPT-3 performed well, their black-box nature presents challenges for transparency and regulatory compliance.

These limitations suggest caution in deployment contexts where explainability, fairness, and latency are critical.

6.5 Future Work

Several promising directions emerge from this work:

1. Multilingual and Code-Switching Support

As international consumer bases expand, systems must accommodate several languages and facilitate dynamic code-switching during interactions. Future models can incorporate **multilingual LLMs** and **language identification layers**.

2. Zero-shot and Few-shot Learning

To reduce the need for large labeled datasets, future research could explore **zero-shot learning** using prompt-based LLMs or **few-shot tuning** with task-specific prompts and examples.

3. Continual and Online Learning

Deployments in dynamic environments require models to evolve over time. Integrating **online learning frameworks** will allow systems to adapt without full retraining.

4. Conversational Flow Analysis

Beyond issue-resolution pairs, analyzing conversational structure—turn-taking, interruptions, escalation patterns—can offer deeper insights into service dynamics.

5. Explainable AI (XAI) Approaches

Integrating explainability tools such as **LIME**, **SHAP**, or **attention-based heatmaps** will improve trust and facilitate compliance in regulated industries.

6. Emotionally Adaptive Response Generation

A prospective system may not only identify emotions but also modify its response style accordingly—soothing agitated users or rejoicing in positive instances. This may integrate emotional intelligence into automated assistance systems.

6.6 Final Remarks

The integration of LLMs into customer service operations signifies a significant advancement in intelligent, sympathetic, and efficient client involvement as AI progresses. This thesis demonstrates that appropriate architecture, data, and design enable AI systems to enhance human capabilities, improve consumer experiences, and provide strategic value to enterprises on a large scale.

This work establishes the foundation for future AI systems that comprehend not only consumer communication but also their emotions, motivations for contact, and the resolutions that will ensure their pleasure.

References

- Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. *In Artificial Intelligence Applications and Innovations (AIAI)* (pp. 373–383). Springer. https://doi.org/10.1007/978-3-030-49186-4_30
- Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *In Mining Text Data* (pp. 163–222). Springer. https://doi.org/10.1007/978-1-4614-3223-4_6
- Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460. https://arxiv.org/abs/2006.11477
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. https://arxiv.org/abs/2005.14165
- Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188. https://doi.org/10.2307/41703503
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186. https://aclanthology.org/N19-1423/
- Gupta, S., Rani, R., & Kumar, P. (2020). Application of artificial intelligence in the management of customer experience. *International Journal of Advanced Science and Technology*, 29(5), 3048–3055. http://sersc.org/journals/index.php/IJAST/article/view/10496
- Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*

arXiv:1907.11692.
https://arxiv.org/abs/1907.11692

- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1–135. https://doi.org/10.1561/1500000001
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67. http://jmlr.org/papers/v21/20-074.html
- Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2017 conversational speech recognition system. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5934–5938. https://doi.org/10.1109/ICASSP.2018.8462506
- Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1–4), 43–52. https://doi.org/10.1007/s13042-010-0001-0

Koushik_Madgula_Final_Thesis-1.pdf

ORIGINALITY REPORT

1%	1%	1%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ai.jmir.org Internet Source	1%
2	salford-repository.worktribe.com Internet Source	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		