Multi-Modal Sentiment Analysis Using Text and Audio for Customer Support Centers

Conference Paper · September 2023

DOI: 10.1007/978-3-031-37164-6_36

4 authors, including:

Hardik Srivastava

5 PUBLICATIONS **3** CITATIONS

SEE PROFILE

Shantha Kumari K.

SRM Institute of Science and Technology

14 PUBLICATIONS **17** CITATIONS

SEE PROFILE

Multi-Modal Sentiment Analysis using Text and Audio for Customer Support Centers

Hardik Srivastava[1], Sneha Sunil[2], K. Shantha Kumari[1*] and P. Kanmani[1*]

[1*]Data Science and Business Systems, SRM Institute of Science and Technology, Chennai, India.
[2]Networking and Communications, SRM Institute of Science and Technology, Chennai, India.

*Corresponding author(s). E-mail(s): shanthak@srmist.edu.in; kanmanip@srmist.edu.in;
Contributing authors: hs9644@srmist.edu.in; ss4795@srmist.edu.in;

Abstract

Customer service has become crucial for every organization looking to grow and improve its clientele in today's cutthroat marketplace. Companies cannot afford to fall short of consumer expectations. With the recent progress in Artificial Intelligence, companies have adopted AI techniques like sentiment analysis to measure customer satisfaction. Chatbots are the foundation for most AI applications in call centers trained for either question-answering tasks or calculating sentiment out of user feedback via surveys. These existing expert systems only utilize text mining techniques to classify sentiment as positive or negative. In this paper, we propose a Multimodal learning framework for tackling the sentiment classification task, employing acoustic and linguistic modalities from real-world conversations between support representatives and customers and survey feedback data using the decision-level fusion technique. Leveraging the classification abilities and feature representations from both modalities, our model achieves the best results amongst all implementations and enables us to better analyze the sentiment expressed by the user than using a single modality like text or audio.

Keywords: Sentiment Analysis, Multimodal Learning, Text Mining, BERT, Speech Representation Learning

1 Introduction

The internet is an excellent platform for people to express their views on various products and services [1]. Using feedback or communication channels, people produce a substantial amount of multimodal data each day that is rich in sentiment information. As people have the ability to comprehend and react and be transparent at the same time, it becomes a necessity for organizations to analyze human emotions regarding a product or a service. A positive opinion could increase customer retention, while a negative opinion can deteriorate possible customer relationships. By analyzing the customer emotions behind these opinions, Artificial Intelligence can help service providers achieve their desired goals, enabling them to scale up their businesses.

Existing research in sentiment analysis and opinion mining has mostly concentrated on interpreting sentiment toward movies, goods, services, acts, etc., which only includes a single entity's perspective [2]. With the introduction of Voice Assistants, Interactive Voice Response (IVR) Systems, and Chatbots, researchers have begun to develop affinity-based conversation systems to enhance the entire experience by adjusting to users' emotions [3]. However, when we consider an instance of a Human-Human Interaction like a Customer-Agent conversation in a support center, it becomes essential for the service providers to ensure customer satisfaction through quality delivery of services. Customer sentiment analysis has gained widespread importance in customer helpline sectors. Support centers are a natural resource of real-world speech data. Every day, a large number of calls are recorded in order to evaluate how well customer support agents (CSRs) and clients communicate. Sentiment analysis is carried out by analyzing the calls made between consumers and representatives.

As opposed to the artificial datasets collected in controlled environments that have traditionally been utilized in multimodal sentiment research, the real-world data—service calls—that we target and employ in this paper provide additional difficulties, such as noise and natural disruptions. The four phases of sentiment analysis are data collection, data engineering, data analysis, and interpretation [4]. However, it has been fraught with many challenges. For instance, if the opinion was neutral or contained an idiom, the trained model could have trouble identifying the sentiment. This problem arises while performing sentiment analysis using conventional techniques considering only text as an input source [5].

In this research, we provide a Multi-Modal method for tackling the sentiment analysis problem, making use of the sentiment data that is present in both the audio modality (vocal intonations) and the text modality (feedback) for a more precise analysis. Since call records as a whole are not consistent and

usually contain a mix of negative and positive utterances, sentiment analysis is addressed at the phrase level. The primary objective here is to determine the customer's emotional tone during the call, capture the sentiment from the shared feedback, and classify them as positive or negative.

The paper is structured as follows. The introduction is followed by Section 2 which describes the various Sentiment Analysis methods that have been studied in the previous papers, including text-based, speech-based, and multimodal approaches, their shortcomings, and the novel method we have proposed for our area of application. Section 3 provides a detailed description of our proposed methodology and system implementation, along with the various speaker recognition and speech processing techniques we have employed. Later in this section, we talk about the various experiments that were carried out. Section 4 talks about our results and discussions. The potential research directions to advance the field are covered in Section 5.

2 Related Work

2.1 Text-based Sentiment Analysis

Text Sentiment Analysis is the process of identifying and analyzing the polarity expressed in a text sequence to find whether it has a positive or negative sentiment. There has been a lot of research in this domain focusing on using algorithms like Naive Bayesian, Decision Trees, and Maximum Entropy [6] [7] [8]. Vimali et al. [9] discussed a methodology for text sentiment classification using a BiLSTM RNN addressing text categorization and cross-lingual issues. They depicted how social media platforms like Twitter allow people to pool many positive and negative opinions, and a deep learning model with multiple layers is required for sentimental analysis. They took the input data from an E-Commerce platform for their research and displayed a visual representation of the positive and negative data obtained after classification. Haifeng et al. [10] achieved better accuracy in a similar task by utilizing the pre-trained BERT model for text classification from the text comments obtained from social media and obtained an accuracy rate of 85.83%. They even considered categories such as uncivilized and questionable comments while classifying the data, which was a drawback in the latter. Movie reviews are a real-world example and a good source for sentiment analysis to predict if people liked or disliked the movie. Zhongxiang et al. [11] showed how it is manually impossible to go through each review to guess the audience's emotions as the review data scales. They built an Albert model classifier [12] and used the dataset provided by Stanford University that contained a tremendous amount of movie reviews for training. Compared to the analysis performed earlier using traditional LSTMs, the accuracy of the Albert model was improved by 3% over the 89.05% accuracy rate.

2.2 Audio-based Sentiment Analysis

In text-based classification systems, we cannot model acoustic features like emotions. Voice's tonal qualities need to be extracted and represented in a form to orderly classify emotions and sentiments when dealing with audio data. For feature extraction from speech data, several researchers have used MFCCs (Mel-Frequency Cepstrum Coefficients) [13] and spectral centroid features. Christine et al. [14] developed a deep-learning model to analyze emotions in a human voice. They developed a CNN deep learning model for emotion analysis and classified it into four categories: happiness, sorrow, neutral, and anger. Their approach showed how sentiment analysis is needed in today's world in developing smart cities that can accelerate economic growth. Idioms and proverbs used in social media communication are figurative and can be very hard to detect the polarity. Bashar et al. [15] proposed a method that uses the automated annotation of idioms that do not require human intervention to enrich the performance of the sentiment classifier. They depicted how their model did not require any fine-tuning for BERT.

2.3 Multimodal Sentiment Analysis

Multi-Modal machine learning is a very recent research topic in the industry [16]. It is gaining popularity among several researchers because of the unlimited multimodality source of information on the internet, like text, audio, and video. Multimodal sentiment analysis systems use the multimodal fusion technique that merges the modalities together. In order to get emotional features via deep multimodal learning, Saumya Roy et al. [17] used a Deep 2D-CNN with multi-scale coupled to dense DNN. Their techniques to model the Speech Emotion Recognition [18] task also solve the problem of data sparsity. They discussed several methods to improve the accuracy of existing models by performing data augmentation, like pitch modification and noise insertion. A more effective multimodal strategy for sentiment analysis is suggested by Roland Goecke et al. [19]. The paper's main objective was to categorize emotions into two categories: positive and negative. The voice mood, age, or gender of the speaker was identified for a better user experience. Sungmoon Cheong et al. [20] proposed using 2-D CNNs and DNNs for encoding each segment into a vector of fixed length by merging the activations together.

The variability and noise that occur in the actual world must be handled by a strong expert system performing sentiment analysis. Our proposed system implementation collects real-world recorded calls that contain ambient noise as well as a variety of person-to-person communication patterns, in contrast to earlier proposed methodologies on multimodal sentiment analysis that used audio data recorded in laboratory settings. To properly extract usable data from various sources, these two circumstances create challenges that must be resolved.

We aim to train a Multi-Modal Sentiment Analysis model to leverage this information from speech data and text to measure customer satisfaction by

directly analyzing the customer-agent interactions recorded in an uncontrolled environment and content extracted from survey and feedback forms.

3 Methodology

3.1 Dataset

We have used two datasets for our experiments, one for audio and another for text. The speech dataset is collected from Summa Linguae Technologies, which provided us with actual Call Center calls in the US English language. The dataset contains 150 audio files of call center conversations sampled at 8000 Hz with a bit rate of 64 kb/s. The speech data between support agents and customers is recorded in noisy real-world conditions and not in a controlled environment like a laboratory, giving our model the benefit of generalizing to real-world data. The audio dataset contains male and female speakers selected randomly. All calls are in wave format and recorded for an average of 13 minutes. We propose leveraging audio labeling techniques to effectively annotate the speech data in a format that is subsequently compatible with our deep learning model. Our entire dataset, on average, amounts to approximately 117000 seconds of speech data, which due to infrastructure constraints, is challenging to process. Hence we consider taking a sample size of 60 audio files from the entire dataset, out of which we prepared 11700 chunks of audio samples which were, on an average of 4 seconds duration each. These utterances are labeled manually with Positive (value of 1) and Negative (value of 0) based on the sentiment of the dialogue in that particular audio sample. All the audio segments are transcribed into text dynamically using Python scripting. The annotations are then exported into JSON format and saved for our model to ingest and perform further analysis. Table 1 shows an example annotation of the first five utterances of one of the calls from our audio dataset. We labeled the utterances with either positive or negative sentiment according to the emotion expressed in the speech. The transcription for the utterance with Segment ID *4* is shown in Table 2.

The data pipeline also consists of speaker diarization [21] and automatic tagging which is discussed later in this section.

Table 1 Representation of a sample annotation of the first five utterances of one of the calls from our audio dataset

Segment ID	Start	End	Total Duration	Sentiment
1	0:00	0:04	0:04	1
2	0:04	0:06	0:02	1
3	0:06	0:08	0:02	1
4	0:08	0:22	0:14	0
5	0:22	0:36	0:14	0

For our text modality, we utilize the Multi-Domain Sentiment Dataset [22], which contains product reviews and customer feedback on multiple product and service types collected from an e-commerce website. The dataset was categorized into 4 product types, with each review consisting of a customer rating (0-5) and some meta-data. Reviews with ratings ≥ 3 were labeled as positive, those with ratings < 3 as negative, and the other reviews were disregarded because of their unclear polarity. Following this conversion, we got a balanced number of 1000 positive and 1000 negative reviews for each domain as in the polarity dataset [23]. We made sure that there were equal amounts of good and negative examples because we were able to collect labels for all of the reviews. After applying various pre-processing techniques like removing null data, filtering out feedback in languages other than English, and removing special characters and punctuation, we obtained 2000 rows of processed data that were suitable for fine-tuning the BERT model for our Sentiment Analysis task.

Table 2 Representation of the transcription for utterance with Segment ID 4, as discussed in subsection 3.1

Segment ID	Transcription
4	Oh not so good. I found out that #ah I'm not able to find any locations that are offering a rental car to throw in my One hundred dollars as bank balance I have ((Pennzoil)) I keep searching every single time What is it? But there's no results.

3.2 Problem Formulation

Now we formally propose our formulation of the problem. Our problem consists of three parts. The first part deals with extracting features from our speech data. Given a speech sequence X_S that is sampled from a conversation between a user and a support agent, its audio features are denoted as $X_S = [S_1, S_2, S_3, \ldots S_n]$. Similarly, for our second part, consider a text sequence of word-piece tokens generated from the feedback shared by a user after his interaction with the agent. Let us assume the sequence to be $T = [T_1, T_2, T_4, \ldots T_n]$, where n is the maximum length of the sequence. Since we are using BERT to generate the text embeddings, the embedding layer will add an additional embedding for the special [CLS] token. Hence, the output of the last encoder layer becomes $X_T = [E_{[CLS]}, E_1, E_2, \ldots E_n]$. Now that we have X_S features and X_T embeddings representing our audio and text modalities, we utilize the decision-level fusion technique to merge these modalities for our Multi-Modal Sentiment Analysis model. We utilize the interaction between X_S and X_T and pass the merged embedding to the Cross-Modal BERT [24] network to perform the sentiment analysis task. Our goal is to create a model that can learn from these multi-modal training data in a way that, given an audio sample

$x_1 \in X_S$ and a text sequence $x_2 \in X_T$, that has not yet been encountered, we can correctly estimate a sentiment value based on the ongoing conversation between a user and a support center agent and the feedback shared by him to efficiently and effectively analyze customer satisfaction.

3.3 System Implementation

In this paper, we propose a Multi-Modal Sentiment Analysis technique that aims to utilize the information available in different modalities to predict customer satisfaction during the customer-agent interaction using sentiment prediction. In our model, we use features extracted from speech signals from conversations between support agents and customers and text data from feedback and survey forms to build a learning framework that leverages the classification ability of each modality. Figure 1 represents our overall system implementation.

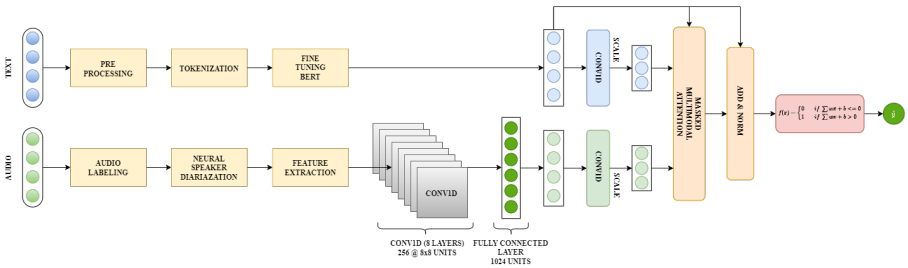

Fig. 1 Representation of our overall System Implementation

We begin with our modeling phase by initially annotating our audio files to extract utterances from the speech samples separated by speaker identifiers. These segments are manually labeled by us in accordance with the output type of our binary classification model. We perform Neural Speaker Diarization [21] or the Speaker Recognition process to fetch the speakers involved in the conversation using PyTorch's API implementation of the Pyannote Audio Toolkit [25]. After generating the chunks of data, we automatically tag them with unique speaker identifiers so that it becomes easy to transcribe their respective data. The tagging algorithm works in an unsupervised fashion, i.e., it first finds whether the chunks are from the same speaker or a different speaker. In addition to transcribing the chunks to text, we also tag them with necessary meta-data like the total running time, the start and end time, the primary language used, etc. This annotated data in JSON format now serves as a segmented and well-labeled representation of the entire audio file with separate information for each speaker's speech output as they converse in the call. Table 3 shows audio annotations for five utterances of an audio file in a tabular format.

After generating our audio annotations, we generate their speech signals using the Librosa framework [26] in Python. Since humans can hear sounds

Table 3 Representation of a sample annotation of the first five utterances of one of the calls from our audio dataset

Segment ID	Primary Type	Locale	Speaker ID	Start	End	Total Duration	Sentiment
1	Speech	*en_US*	*en_US*_1000	0:00	0:04	0:04	1
2	Speech	*en_US*	*en_US*_1001	0:04	0:06	0:02	1
3	Speech	*en_US*	*en_US*_1000	0:06	0:08	0:02	1
4	Speech	*en_US*	*en_US*_1001	0:08	0:22	0:14	0
5	Speech	*en_US*	*en_US*_1000	0:22	0:36	0:14	0

with a comparatively small range of frequencies and loudness, we also produced an identical signal for time-frequency analysis of the various frequencies and amplitudes of our signal over time. The frequency and amplitude are calculated which helps us plot the spectrogram for analysis, showing decibels of the various frequencies over time.

Following that, we extract the Mel Frequency Cepstral Coefficients [13] from each signal wave. These coefficients are used for the feature extraction process when dealing with speech data. They represent a sound's power spectrum across a short time span [27]. These MFCC values help in further making our speaker diarization system more robust. Figure 2 shows an MFCC spectrogram plotted for one of the audio files.

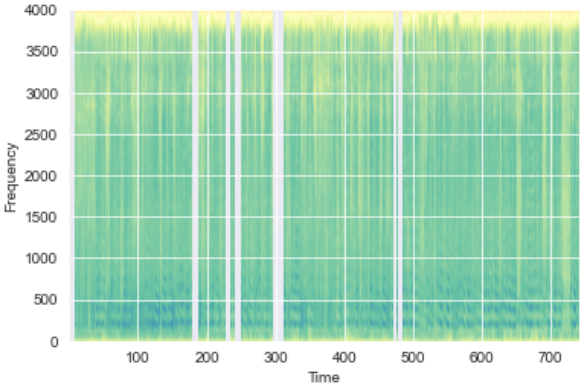

Fig. 2 Representation of an MFCC Spectrogram of an audio file in the Time-Frequency domain

Since speech signals are sequence data, it is recommended to use an Encoder-Decoder framework, as RNNs are noted for being effective in audio modeling tasks. However, we propose using a Feed-Forward Convolutional Neural Network to model the speech data in our system implementation, as recent studies prove that CNNs tend to perform better than RNNs in various scenarios. The network architecture consists of an input layer, eight convolutional layers (Conv1D) for feature extraction, and one fully connected

layer with 1024 hidden neurons. A Max Pooling or an Average Pooling Layer, and a Dropout layer with a dropout value of 0.5 was placed after every Conv1D Layer. Convolutional and Fully connected layers employ Rectified Linear Units (ReLU) [28] as activation functions to introduce non-linearity. The number of kernels is set to eight for all convolutional layers, respectively. Cross-entropy was used to calculate the loss while training and validation, and the Stochastic Gradient Descent [29] algorithm was employed as the optimizer over RMSProp to minimize the loss function over the mini-batches of the training data. We use a learning rate, best suited for our task, of 1e-4 and a batch size of 32 to train our model for 100 epochs over our call center audio data for the sentiment analysis task.

For our text modeling task, we leveraged the Transformers API from Hugging Face to implement the BERT Language Model [30] in TensorFlow and tweaked it with additional training data to make it perform our Sentiment Analysis task.

The multimodal machine learning technique with decision-level fusion (late fusion) was utilized to merge the audio data from our tests with the BERT results in the last phase of our research. We mixed the aforementioned models by fusing feature representations from both our CNN network and the BERT model using the Cross-Modal BERT (CM-BERT) [24] architecture into a single output representation.

3.3.1 Neural Speaker Diarization

Speaker diarization [21] is a method of segmenting a human voice audio stream into homogenous parts based on the identification of each speaker. The readability of the speech transcription is improved by annotating the speech signal by segmenting the audio stream into speaker turns and identifying the speaker's real identity when used in conjunction with speaker identification technologies. NSD answers the question of "Who spoke when?" Speaker segmentation and speaker clustering are used to create speaker diarization. One of the most popular techniques to implement NSD is Gaussian mixture models. We leverage Pyannote's Audio API, an open-source toolkit in Python, to perform our diarization task. After all the speakers participating in the conversation are identified, we assign them unique identifiers and tag them with their respective transcribed text. This helps us generate our annotated call center audio data. Table 3 represents an example annotation showing the first five utterances of one of our audio files.

3.3.2 Mel Frequency Cepstral Coefficients (MFCCs)

Any automatic speech recognition system starts with the process of extracting features, which involves recognizing the parts of the audio signal that are good for detecting the linguistic content and ignoring anything else that contains information such as background noise, emotion, etc. We use MFCCs [13] to divide the frequency band of the signal into sub-bands using the MEL scale and then extract the Cepstral Coefficients using a linear cosine transform. A

pure tone's perceived frequency or pitch, and its actual measured frequency is related using the Mel scale. Using this scale makes our features more similar to human hearing. A spectrogram showing the distribution of frequency v/s time of several pronounced words is shown in Figure 2. Equation 1 shows the relation between frequency and the Mel scale.

$$M(f) = 1125 \ln (1 + f/700) \tag{1}$$

3.3.3 Audio Feature Extraction

Raw audio signals contain a lot of noise and disturbance, even when recorded in a controlled environment. However, our call center audio dataset contains recordings of actual conversations from an organization's support center that took place in an uncontrolled environment. Using noisy disturbances and real-world data helps us model the actual speech features which our algorithm will be evaluated on when deployed in a real-world scenario. Because of the noise and disturbance, these signals cannot be fed directly as input to our network. It is observed that for using the base model's input, features extracted from the audio signal rather than the raw audio signal alone would result in significantly improved performance. Hence, MFCCs are used to extract characteristics from audio signals, and therefore we have used them to be our input feature. Loading audio data and converting it to MFCC format can be quickly done by the Python package Librosa.

3.3.4 Text Encoder

The text input X^t for our model, extracted from customer feedback, is first tokenized by the BERT Tokenizer using WordPiece [31]. We mark the beginning of sentences using the [CLS] special tokens respectively. Adding the [CLS] token is necessary for BERT to generate our word embeddings. These tokens are then converted to token ids using BERTTokenizer's fixed-size vocabulary. Assuming the token ids to be X^{tk}, the encoded feature representation in the form of word embeddings is generated and given as output by BERT. This computation is formally represented as $E^t = BERT(X^{tk})$

3.3.5 Fusion

Fusion is the technique to merge multiple representations from different classifiers into a single representation. There are two ways to perform this task: decision-level fusion (late fusion) and feature-level fusion (early fusion). We used decision-level fusion in our experiments to merge both our modalities into one single representation. The benefits of decision-level fusion are numerous [32]. We employed classifiers for text and audio characteristics independently, which is one of the major advantages when using late fusion. A further advantage is the method's processing speed and high-performance capabilities.

3.4 Experiments

3.4.1 Data Pre-Processing

The methods of pre-processing while extracting features from our audio and text modalities are as below:
Text: The reviews are tokenized using WordPiece, and the encoded text representations are generated in accordance with the BERT's format of input embeddings.
Audio: We annotate our call center audio data to prepare chunks of shorter audio clips segregated by respective speaker ids. Annotating the calls helps us to label them with respect to the sentiment expressed in the dialogue in accordance with our Sentiment Analysis task. We use MFCCs to extract the audio features for our model and further analyze the respective speech signals.

After several minor changes and tweaks in hyperparameter values like the optimizer, learning rate schedule, and even our network architecture, we discovered that our CNN model performs better generalization if we introduced augmented audio samples to our network. Hence, we explored some audio augmentation methods to increase the variability in our audio data. We explored techniques like Pitch Tuning and Shifting, which shift the pitch of the speech signal up or down depending on the pitch interval that is chosen beforehand.

3.4.2 Fine-Tuning

We leverage the BERT Language Model [30] and tweak it with additional training data to make it perform our Sentiment Analysis task for our text modality. We express the input embeddings as the sum of token, segmentation, and position embeddings to fine-tune the sentiment analysis model on our multi-domain sentiment dataset [33]. The pre-training head of the model is removed and replaced with a classification head. When fine-tuning, only the classification layer weights $W \in R^{N \times H}$, where N represents the number of output labels, are added as additional parameters. We use a batch size of 16 and fine-tune for 100 epochs over the data. We choose the best fine-tuning learning rate of 2e-5 for each task on the validation set. During the evaluation, our model achieves an accuracy of 77% and a weighted F1-score of 0.76 on the test set.

4 Results and Discussions

We use accuracy and the weighted F1-score, which is the weighted harmonic mean of precision and recall, as evaluation metrics. The likelihood of delivering accurate values is known as precision. The likelihood that the algorithm would provide meaningful values is known as recall. Both precision and recall for our positive and negative classes are also considered for evaluation.

When we ran our experiments and recorded our results, as tabulated in Table 4, we observed that CNNs performed better than RNNs in modeling

the audio features. Our CNN model when evaluated over our data, initially achieved an accuracy of 69% and a weighted F1-score of 0.68, after applying augmentation techniques, scaled its performance by 17% to reach an accuracy of 83% and a weighted F1-score of 0.81. These findings were better than what our RNN network produced. For our text modality, we ran experiments on machine learning-based classifiers to get an analytical overview. Multinomial Naïve Bayes [34] and Support Vector Machines [35] achieved overall weighted F1-scores of 0.75 and 0.30. Long Short-Term Memory Network (LSTM) [36] performs better than these classifiers by reaching an accuracy of 68% and a weighted F1-score of 0.69. When we choose to fine-tune BERT over our multi-domain sentiment dataset, our model achieves acceptable results, like an accuracy of 77%, considering that our dataset contains some instances of transcribed examples and some grammatically incorrect data points that are insignificant and vague.

Table 4 Representation of the results from the Sentiment Classification task on our test data. Audio and Text are treated as individual and separate modalities here

	Audio			Text			
Metrics	CNN	CNN (Aug)	RNN	MNB	SVM	LSTM	BERT
Accuracy	0.69	0.83	0.77	0.74	0.58	0.68	0.77
F1 (w)	0.68	0.81	0.81	0.75	0.30	0.69	0.76
Precision (+)	0.59	0.79	0.83	0.74	0.94	0.57	0.77
Precision (-)	0.55	0.77	0.80	0.76	0.55	0.81	0.78
Recall (+)	0.78	0.85	0.79	0.77	0.18	0.88	0.76
Recall (-)	0.78	0.76	0.83	0.73	0.99	0.66	0.73

Note: F1 (w) is the macro-weighted F1-score. Precision (+) and Recall (+) denote the precision and recall values for our positive class and Precision (-) and Recall (-) denote the precision and recall values for our negative class

Table 5 Representation of the results from the Sentiment Classification task on both audio and text modalities

	Audio & Text
Metrics	Multi-Modal Fusion Model
Accuracy	0.89
F1 (w)	0.88
Precision (+)	0.86
Precision (-)	0.84
Recall (+)	0.90
Recall (-)	0.88

In order to achieve better results, we merge the two modalities (audio and text) together using the CM-BERT (Cross-Modal BERT) architecture. The

multimodal model shows better results than the CNN model (with augmented data) for audio and the classifiers employed for text sentiment analysis. Table 5 represents the tabulated results for our Multi-Modal model when evaluated on both modalities. Our Multi-Modal Sentiment Analysis model achieved an overall accuracy of 89.41% and a weighted F1-score of 0.88 on our call center audio and customer sentiment datasets. Our results prove that leveraging both linguistic and acoustic modalities enable us to better analyze the sentiment expressed by the user than using a single modality like text or audio. The audio features help us to capture the customer's tonality and emotion when engaged in a dialogue by analyzing the respective speech signals, which is not possible when considering only text.

5 Conclusion

Insights gained from customer sentiment analysis are rich and valuable, allowing organizations to build customer relationships and enhance loyalty. In this paper, we proposed a system that performs the sentiment analysis task using a multi-modal representation of both audio and text. We benchmark the Call Center Dataset by Summa Linguae Technologies for audio modality and Multi-Domain Sentiment Analysis Dataset for text modality with four modeling approaches, including both machine learning and deep learning techniques. Our Multi-Modal Sentiment Analysis model demonstrates that fusing the audio and text representations to measure customer sentiment in support center areas is better than adopting a single text representation, delivering the best results among all implementations.

Our future research will focus on broadening the existing network architecture, incorporating new features, and increasing the data domain for our model, especially for the audio modality. We are seeking more effective ways to analyze acoustic and text features and how they can help drive important decisions. Exploring additional modalities like video and their role in assessing customer sentiment will open new areas of research. Furthermore, we would like to improve our multi-modal fusion process to achieve better results.

References

[1] P. Mishra, R. Rajnish and P. Kumar, "Sentiment analysis of Twitter data: Case study on digital India," 2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit on the Theme - Internet of Things: Connect your Worlds, Noida, India, 2016, pp. 148-153, doi: 10.1109/INCITE.2016.7857607.

[2] M. R and S. M.R, "Opinion Mining on Movie Reviews," 2019 1st International Conference on Advances in Information Technology (ICAIT), Chikmagalur, India, 2019, pp. 282-286, doi: 10.1109/ICAIT47043.2019.8987366.

[3] R. R. Sehgal, S. Agarwal and G. Raj, "Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems," 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, France, 2018, pp. 213-218, doi: 10.1109/ICACCE.2018.8441741.

[4] A. J. Nair, V. G and A. Vinayak, "Comparative study of Twitter Sentiment On COVID - 19 Tweets," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1773-1778, doi: 10.1109/ICCMC51019.2021.9418320.

[5] Y. Chen and Z. Zhang, "Research on text sentiment analysis based on CNNs and SVM," 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, China, 2018, pp. 2731-2734, doi: 10.1109/ICIEA.2018.8398173.

[6] Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.

[7] Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 115-124). Association for Computational Linguistics.

[8] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

[9] J. S. Vimali and S. Murugan, "A Text Based Sentiment Analysis Model using Bi-directional LSTM Networks," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1652-1658, doi: 10.1109/ICCES51350.2021.9489129.

[10] H. Liang, B. Tang and S. Cao, "Sentiment Analysis of Comment Texts on Converged Media Platforms based on BERT Model," 2021 International Conference on Culture-oriented Science & Technology (ICCST), Beijing, China, 2021, pp. 552-555, doi: 10.1109/ICCST53801.2021.00120.

[11] Z. Ding, Y. Qi and D. Lin, "Albert-based sentiment analysis of movie review," 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Changsha, China, 2021, pp. 1243-1246, doi: 10.1109/AEMCSE51986.2021.00254.

[12] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. ArXiv. https://doi.org/10.48550/arXiv.1909.11942

[13] M. A. Hossan, S. Memon and M. A. Gregory, "A novel approach for MFCC feature extraction," 2010 4th International Conference on Signal Processing and Communication Systems, Gold Coast, QLD, Australia, 2010, pp. 1-5, doi: 10.1109/ICSPCS.2010.5709752.

[14] C. J. Sora and M. Alkhatib, "Speech Sentiment Analysis for Citizen's Engagement in Smart Cities' Events," 2022 7th International Conference on Smart and Sustainable Technologies (SpliTech), Split / Bol, Croatia, 2022, pp. 1-5, doi: 10.23919/SpliTech55088.2022.9854309.

[15] B. M. A. Tahayna, R. K. Ayyasamy and R. Akbar, "Automatic Sentiment Annotation of Idiomatic Expressions for Sentiment Analysis Task," in IEEE Access, vol. 10, pp. 122234-122242, 2022, doi: 10.1109/ACCESS.2022.3222233.

[16] Ngiam, Jiquan & Khosla, Aditya & Kim, Mingyu & Nam, Juhan & Lee, Honglak & Ng, Andrew. (2011). Multimodal Deep Learning. Proceedings of the 28th International Conference on Machine Learning, ICML 2011. 689-696.

[17] S. Roy, S. Ghoshal, R. Basak, P. Basu and N. Roy, "Multimodal sentiment analysis of human speech using deep learning," 2022 Interdisciplinary Research in Technology and Management (IRTM), Kolkata, India, 2022, pp. 1-4, doi: 10.1109/IRTM54583.2022.9791745.

[18] Aouani, H., & Ayed, Y. B. (2020). Speech Emotion Recognition with deep learning. Procedia Computer Science, 176, 251-260. https://doi.org/10.1016/j.procs.2020.08.027

[19] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. 2013. Emotion recognition in the wild challenge 2013. In Proceedings of the 15th ACM on International conference on multimodal interaction (ICMI '13). Association for Computing Machinery, New York, NY, USA, 509–516. https://doi.org/10.1145/2522848.2531739

[20] Sungmoon Cheong, Sang Hoon Oh, Soo-Young Lee "Support Vector Machines with Binary Tree Architecture for Multi-Class Classification".

[21] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, "Speaker Diarization: A Review of Recent Research," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 356-370, Feb. 2012, doi: 10.1109/TASL.2011.2125954.

[22] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

[23] Yuan, Z., Wu, S., Wu, F., Liu, J., & Huang, Y. (2018). Domain attention model for multi-domain sentiment classification. Knowledge-Based Systems, 155, 1-10. https://doi.org/10.1016/j.knosys.2018.05.004

[24] Kaicheng Yang, Hua Xu, and Kai Gao. 2020. CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 521–528. https://doi.org/10.1145/3394171.3413690

[25] H. Bredin et al., "Pyannote.Audio: Neural Building Blocks for Speaker Diarization," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7124-7128, doi: 10.1109/ICASSP40776.2020.9052974.

[26] McFee, Brian & Raffel, Colin & Liang, Dawen & Ellis, Daniel & Mcvicar, Matt & Battenberg, Eric & Nieto, Oriol. (2015). librosa: Audio and Music Signal Analysis in Python. 18-24. 10.25080/Majora-7b98e3ed-003.

[27] Z. Wanli and L. Guoxin, "The research of feature extraction based on MFCC for speaker recognition," Proceedings of 2013 3rd International Conference on Computer Science and Network Technology, Dalian, China, 2013, pp. 1074-1077, doi: 10.1109/ICCSNT.2013.6967289.

[28] Agarap, A. F. (2018). Deep Learning using Rectified Linear Units (ReLU). ArXiv. https://doi.org/10.48550/arXiv.1803.08375

[29] Ruder, Sebastian. "An overview of gradient descent optimization algorithms." ArXiv abs/1609.04747 (2016): n. pag.

[30] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv. https://doi.org/10.48550/arXiv.1810.04805

[31] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv. https://doi.org/10.48550/arXiv.1609.08144

[32] S. Roheda, H. Krim, Z. -Q. Luo and T. Wu, "Decision Level Fusion: An Event Driven Approach," 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 2018, pp. 2598-2602, doi: 10.23919/EUSIPCO.2018.8553412.

[33] Z. A. Guven, "Comparison of BERT Models and Machine Learning Methods for Sentiment Analysis on Turkish Tweets," 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021, pp. 98-101, doi: 10.1109/UBMK52708.2021.9559014.

[34] Xu, Shuo & Li, Yan & Zheng, Wang. (2017). Bayesian Multinomial Naïve Bayes Classifier to Text Classification. 347-352. 10.1007/978-981-10-5041-1_57.

[35] Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.

[36] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.