Advancing Multimodal Sentiment Analysis through Enhanced Sarcasm Detection

Evelyn Huang
UC San Diego
xih037@ucsd.edu

Feiyang Jiang
UC San Diego
fejiang@ucsd.edu

Vivian Zhao
UC San Diego
vxzhao@ucsd.edu

Zhiqing Wang
UC San Diego
zhw055@ucsd.edu

1 Introduction

The proposed project aims to develop an advanced multimodal sentiment analysis model that incorporates both audio and text data inputs to detect underlying sentiments, specifically focusing on enhancing the pre-existing models' ability to identify sarcasm. Conventional sentiment analysis models often misclassify sarcastic text as positive without reading under the lines. By integrating audio cues, which may convey negativity or dissonance with the textual sentiment, our model seeks to accurately detect sarcasm and thereby improve the overall accuracy of sentiment classification.

We propose to train **an additional model** that focuses on identifying sarcasm. This sarcasm model will be used in the classification step of conventional multimodal sentiment models in order to enhance their accuracy in detecting sarcasm.

2 Background & Related Work

Current sentiment analysis research in machine learning has led to the development of various text-based feature extraction techniques, such as Bag of Words (BoW), Word2Vec, N-gram, TF-IDF, Hashing Vectorizer (HV), and Global Vectors for Word Representation (GloVe). These methods are frequently utilized alongside algorithms such as Naive Bayes, Support Vector Machines (SVM), and decision trees (Nandwani and Verma, 2021). Text sentiment analysis, particularly in binary classification (positive/negative), has demonstrated high levels of accuracy. For example, Soumya and Pramod classified 3,184 Malayalam tweets with an impressive accuracy of 95.6% (Soumya and Pramod, 2020). Beyond binary classification, text analysis has been extended to multiclass emotion detection. For instance, Mondher Bouazizi and Tomoaki Ohtsuki attempted to classify tweets into seven distinct emotion categories but achieved a significantly lower accuracy rate of 56.9% (Bouazizi and Ohtsuki, 2016). This reduction in performance highlights the challenge of text-based sentiment analysis, particularly in recognizing complex phenomena like irony, sarcasm, or subtle emotional cues.

Recent research has increasingly focused on audio-based sentiment analysis, which identifies emotions through acoustic features such as fundamental frequency, energy profile, silence duration, and speech quality. Luitel and Anwar converted audio signals into spectrograms and utilized a visual bag-of-words (BoVW) method to transform the descriptors from each column of the spectrogram into histograms, allowing for the classification of emotions in a multilingual dataset (Luitel and Anwar, 2022). Meanwhile, Wu and Liang implemented a hybrid approach that combines acoustic-prosodic (AP) recognition with semantic label (SL)-based recognition through a maximum entropy model, resulting in improved accuracy in emotion classification (Wu and Liang, 2011). These approaches highlight the benefits of leveraging audio data in sentiment analysis, particularly for detecting emotions that may be challenging to discern through text alone.

Faced with the challenges of text-based sentiment analysis, particularly in addressing irony and complex emotional states, and recognizing the effectiveness of audio-based sentiment analysis in emotional classification, this paper proposes a multimodal approach that integrates both text and audio

Github link: https://github.com/VivianZhao12/Multimodal-Sentiment-Analysis-with-Sarcasm-Detection data. By combining textual analysis with audio features, this study aims to improve the detection of irony and sarcasm and increase overall sentiment classification accuracy (Alyzehkazmi, 2024).

3 Data Collection

Our primary motivation for this data collection is the development of a multimodal sentiment analysis model that includes sarcasm detection capabilities. Traditional sentiment analysis models often rely heavily on textual data, but the nuanced nature of human emotions, especially sarcasm, requires a more robust approach that incorporates audio cues. This is because audio can convey subtle variations in tone and inflection that text alone cannot capture. As a result, our model not only trains on sentiment through textual analysis but also learns from audio inputs to better understand and interpret emotions. We utilized the Crema-D dataset as we found that its focus was predominantly on audio quality, with less emphasis on textual content, allowing it to train the audio features better. Additionally, our project had specific requirements for sarcasm detection that were not sufficiently met by any existing datasets. To address this gap, we took the initiative to collect our own dataset from a variety of real-world sources, ensuring that our model could accurately detect sarcasm and other complex emotional states, thereby enhancing its effectiveness and applicability in practical scenarios.

Existing Dataset: Crema-D

The dataset consists of 7,442 audio clips from 91 actors (48 male and 43 female) aged 20-74 acting out a set of sentences with different assigned emotions. These actors represent diverse ethnicities, including African American, Asian, Caucasian, and Hispanic individuals.

Sources for Manually Collected Data

We also collected video clips from a wide range of sources, including movies such as Inside Out, shows like Big Bang Theory, stand-up comedy performances, and talk shows like the Jonathan Ross Talk Show. These clips provide accurate presentations of various emotions, including sarcasm and sentiment, ensuring correct labeling for each data point.

Table 1: Dataset Labeling and Features Description

Label/Feature	Description	Type
emotion_label	A string recording the emotion of each clip	String
sentiment_label	A binary for positive and negative	Binary
sarcasm_label	A binary indicating if the clip belongs to sarcasm emotion	Binary
text	The content of the audio transcribed using the OpenAI-Whisper model	String
wav audio files	The unprocessed audio clips	Audio File

4 Methods

Our aim in this pipeline is to train two models: Sarcasm Detection Model and Sentiment Model. We first train the Sarcasm Detection Model and incorporate its results into the final Sentiment Model in hopes of enhancing the performance of the sentiment model.

4.1 Data Source

The pipeline begins with the collection of datasets: a General Sentiment Dataset and a Sarcasm Dataset. These datasets are likely composed of both text and audio data, providing a rich foundation for analysis. The sarcasm detection model dataset contains 257 samples and the general sentiment model dataset contains 172 samples. These dataset will be used to train and test the sarcasm model and sentiment model separately.

4.2 Method Pipeline

Our proposed pipeline consists of multiple sequential stages that integrate preprocessing, feature engineering, model design, and evaluation. The goal is to utilize multimodal data synergistically to improve sentiment analysis accuracy while addressing sarcasm as a confounding factor. A high-level schematic of the pipeline is shown in Figure 1.

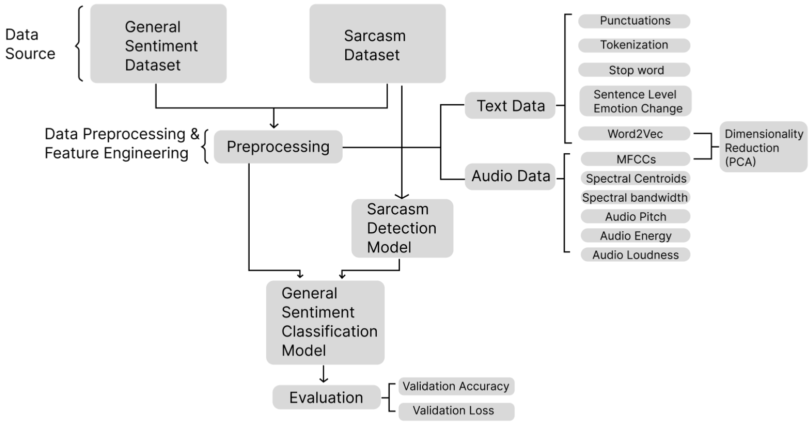


Figure 1: An example image demonstrating the pipeline.

4.3 Feature Engineering

This stage involves cleaning and standardizing the data, which is crucial for both datasets. For text data, preprocessing steps might include punctuation removal, tokenization, elimination of stop words, and capturing sentence-level emotion changes. For audio data, preprocessing likely involves noise reduction and normalization to ensure consistency across samples.

4.3.1 Audio Data

In our audio feature extraction process, we implemented several critical steps to enhance the quality and consistency of the audio data, crucial for accurate sentiment analysis. To ensure the clarity of the audio tracks and reduce background noise, we applied advanced noise reduction techniques. This step was essential to isolate the vocal elements clearly, minimizing distractions that could skew the analysis. Additionally, we standardized all audio clips into a single track format, which was vital for maintaining uniformity across the dataset, thus ensuring that all features extracted were comparable and not influenced by differences in audio format or channel distribution. The selection of specific features like **Mel Frequency Cepstral Coefficients (MFCCs), energy, and spectral centroid** was driven by their relevance and effectiveness in capturing the nuanced characteristics of sound that are indicative of human emotions. MFCCs are especially valuable as they encapsulate the short-term power spectrum of sound, providing a robust representation of the audio's timbre and texture, which are closely linked to the vocal expression of emotions. The energy feature measures the signal's power, giving insights into the intensity and dynamics of the spoken content, which can vary significantly with different emotional states. The spectral centroid quantifies the brightness or sharpness of a sound, offering cues about the emotion being expressed, with higher centroids often associated with more intense or aggressive emotions.

4.3.2 Textual Data at Sentence Level

For **Sentence Level similarity in sentiment within the text**, we first split the text into sentences, remove stop words for each sentence and combine the remaining words into sentences to compare the sentence level similarity in words within the text. Then we use the Bert model to get representation for each sentence. After getting the representation, we employ a bigram approach to compare sentences in pairs, calculating the average similarity score to derive the final representation. The resulting feature ranges from -1 to 1, where -1 indicates no similarity, and 1 indicates high similarity. For texts containing only one sentence, we assign a default feature value of 0 to avoid introducing biases. The design intuition is that sarcasm relies on a high level of similarity between words within a sentence, as they work together to convey the underlying sarcastic intent. Therefore, we anticipate that words within sarcastic sentences will exhibit a closer semantic relationship. For **Sentence Level similarity in sentiment within the same text** We process the text the same way as “Sentence Level similarity in words” without removing the stop words. We plan on using our general sentiment model but without the sarcasm detection model to average sentiment change score within the text. Sarcasm often involves exaggerated shifts between positive and negative emotions, creating stark sentiment contrasts. For example, in “That’s a good idea. Just kidding, you idiot,” the tone shifts sharply from positive to negative, emphasizing the sarcasm.

4.3.3 Textual Data at Token Level

At the token level, we calculated the average embeddings for each text using Word2Vec. This involved representing each word in the text as a dense vector in a high-dimensional space and then averaged the embeddings of all its tokens to create a unified vector representation. To explore the influence of stop words in sarcasm sentences, we performed this embedding process in two versions: one where stop words were included and another where they were excluded. Stop words are often considered less informative for tasks such as sentiment prediction, so excluding them reduce noise and highlight more meaningful tokens. On the other hand, including stop words might preserve the linguistic context which is essential in detecting sarcasm, potentially providing insights into subtle shifts in tone. This feature is particularly useful in capturing complex emotion as it encapsulates the overall semantic content of the text, allowing the model to better capture nuanced patterns. We also using exclamation marks as a feature. By identifying and removing these, we aimed to focus the analysis on the core textual content without distraction.

4.3.4 Dimensionality Reduction

For high-dimensional representations like Word2Vec embeddings and MFCCs, PCA was applied to manage the dimensionality of the input features. This method reduced computational overhead while retaining the most significant features, enabling efficient and effective model training.

4.4 Models

4.4.1 Sarcasm Detection Model

Given the size of our dataset, we began with simpler models for their interpretability and resistance to overfitting, using them as a baseline for comparison. Recognizing the potential complexity of interactions between text and audio features, we extended our approach to deep learning models, employing techniques such as regularization, batch normalization, dropout, and attention mechanisms to balance complexity and overfitting while capturing intricate patterns in the data.

Simpler Approaches:

- **Logistic Regression, Support Vector Machine, Naive Bayes, Random Forest**

Deep Learning Approaches:

- **Fully Connected Neural Network:** The model has an input layer with 64 neurons, two hidden layers (64 neurons each) with ReLU activation and 50% dropout, and a sigmoid output layer. It is optimized with Adam and trained using binary cross-entropy loss.
- **Enhanced Fully Connected Neural Network:** This model adds L2 regularization, batch normalization, 30% dropout, early stopping, and model checkpoints to prevent overfitting.
- **Hybrid Model:** It combines text and audio input branches, each with a 128-neuron dense layer, ReLU activation, and 20% dropout. The branches merge into a shared 64-neuron layer with dropout and sigmoid output, optimized using Adam and binary cross-entropy loss.
- **Enhanced Hybrid Model:** This model adds L2 regularization, batch normalization, and 30% dropout, with early stopping and checkpoint to prevent overfitting.
- **Hybrid Model with Attention:** Incorporates an Attention Layer to improve feature selection and interpretability.

4.4.2 General Sentiment Classification Model

We then incorporate the sarcasm models into the sentiment model as features (signaling if the data is sarcasm) and aim to enhance the sentiment performance. In addition, it uses the processed features to determine the overall sentiment of the data, classifying it into positive and negative. Below, we describe the conceptual designs of the models:

- **DNN Without Sarcasm Feature:** This model focuses solely on sentiment detection using traditional input features such as word embeddings and sentiment polarity scores. By excluding any additional sarcasm-related features, it serves as a baseline model for understanding the primary sentiment expressed in the text without any adjustments for sarcasm.
- **DNN with Sarcasm Feature Integrated:** In this approach, a sarcasm prediction feature was added to the model's input pipeline. The integration of this feature aimed to account for sarcastic expressions that might alter the true sentiment of the text. By combining this feature with sentiment embeddings, the model was designed to more accurately interpret nuanced and sarcastic language.
- **DNN with Sarcasm-Filtered Sentiment Adjustment:** Building upon the second approach, this model incorporates the sarcasm feature but introduces a sentiment adjustment mechanism. Text identified as highly sarcastic is directly filtered or adjusted to reinterpret its emotional polarity, thereby mitigating false negative sentiment classifications caused by sarcasm. This filtering mechanism enhances the alignment of predictions with the actual sentiment conveyed.

4.5 Intuition and Originality

Our approach aims to outperform existing methods by leveraging multimodal data and incorporating sarcasm detection to better capture complex emotions. By explicitly detecting sarcasm and analyzing sentiment shifts, our pipeline ensures more robust sentiment predictions through the integration of text and audio modalities, which provides a richer context for traning.

Our manual collection of a sarcasm-specific dataset, meticulously designed to capture the nuances of sarcasm and tone, enabling better model training with high quality data. The pipeline's robustness is further enhanced by leveraging state-of-the-art models and integration of features specifically engineered to capture sarcasm such as sentence-level similarity and sentiment shifts. Its originality lies in addressing sarcasm as a confounding factor, combining advanced feature engineering with efficient multimodal fusion techniques.

5 Experiment and Results

In this section, we evaluate the performance of all proposed Sarcasm and Sentimet Models. Performance metrics include Accuracy and F1-score, with a particular emphasis on the F1-score to ensure balanced performance across classes.

5.1 Sarcasm Detection Model

Table 2: Model Results on Text Audio Features

Model Name	Accuracy (%)	F1 Score (%)
Logistic Regression with Text Feature Only	69	71
Logistic Regression with Text and Audio Features	77	78

We first implemented Logistic Regression as the baseline model for this binary classification task, running it separately with text-only data and with combined text and audio data. The results demonstrated that the inclusion of audio data significantly improved performance, yielding higher accuracy and better precision-recall metrics as shown in the figures. Besides, based on the analysis of significant features in logistic regression with text and audio features, such as PCA_MFCC_1 and PCA_MFCC_8 (audio) and PCA_W2V_6 and sentence_level_similarity_word (text), confirmed that audio cues provide valuable contextual and tonal information, referred to the figures. This finding validated the multimodal approaches in capturing nuanced patterns of sarcasm that may be missed when relying solely on textual features.

Table 3: Sarcasm Model

Model Name	Accuracy (%)	F1 Score (%)
Logistic Regression	77	78
SVM	79	79
Naive Bayes	77	78
Random Forest	81	80
FCNN	79	79
Enhanced FCNN	81	81
Hybrid Model	79	79
Enhanced Hybrid Model	83	82
Enhanced Hybrid Model with Attention	67	67

Simpler Model Among the traditional models, Random Forest exhibited the highest accuracy (81%) and F1 score (80%) comapring to other method as shown in the figures. This can be attributed to its ability to capture non-linear relationships between features, which is particularly beneficial when working with our PCA-reduced data.

Deep learning Model Among the advanced models, the Enhanced Hybrid Model achieved the best performance, with the highest accuracy (83%) and F1 score (82%), effectively balancing precision and recall as shown in the figures. In contrast, attention-based models exhibited the lowest performance, with both accuracy and F1 score at 67%, reflecting the challenges of training complex architectures on small datasets and PCA-reduced features. The improved performance of the enhanced variants highlights the effectiveness of regularization in our scenario.

Comparison and Insights The Enhanced Hybrid Model slightly outperformed Random Forest. However, we selected Random Forest as the final Sarcasm Model to be included in our General Sentiment model due to its simpler interpretability, lower computational complexity, and comparable performance to the advanced models. While the Enhanced Hybrid Model process text and audio features independently, Random Forest offers a practical balance of simplicity and reliability, making it the preferred choice despite the marginal performance difference.

5.2 General Sentiment Classification Model

Table 4: Sentiment Model

Model Name	Accuracy (%)	F1 Score (%)
DNN without Sarcasm Feature (Pure Sentiment Model)	84	84
DNN with Sarcasm Feature Integrated	88	89
DNN with Sarcasm-Filtered Sentiment Adjustment	73	65

DNN Without Sarcasm Feature (Pure Sentiment Model) This baseline model achieved an accuracy of 0.84 and a weighted F1-score of 0.84. The F1-score, however, was 0.78, indicating that performance across classes was not entirely uniform. While this model captures sentiment effectively, the lack of a sarcasm-related feature limits its ability to accurately classify texts that include sarcastic expressions. This is evident from the slightly lower macro average F1-score, as the model struggles to balance performance between positive and negative sentiment classes when sarcasm is present.

DNN with Sarcasm Feature Integrated The integration of a sarcasm detection feature improved both accuracy and F1-scores. This model achieved an accuracy of 0.88 and a F1-score of 0.89, reflecting better balance in classification performance across classes. The inclusion of sarcasm features allowed the model to better interpret texts with nuanced expressions, reducing misclassifications that were prevalent in the baseline model. This improvement highlights the importance of incorporating context-sensitive features when handling sarcastic content.

DNN with Sarcasm-Filtered Sentiment Adjustment This model introduced an adjustment mechanism for sarcastic texts, achieving an accuracy of 0.73 and a F1-score of 0.65, revealing a stark imbalance in the model's performance. This indicates that the sarcasm-filtering mechanism introduced significant bias, effectively ignoring the minority class and leading to poor overall performance despite reasonable accuracy.

Comparison and Insights The results demonstrate that integrating a sarcasm detection feature significantly enhances sentiment classification performance, as seen in the DNN with Sarcasm Feature Integrated model. However, the additional adjustment mechanism in the DNN with Sarcasm-Filtered Sentiment Adjustment model introduced substantial challenges, particularly in handling minority classes. While this mechanism aimed to refine sentiment predictions, its implementation needs further refinement to prevent the observed class imbalance.

Overall, the DNN with Sarcasm Feature Integrated model offers the best balance of accuracy and F1-score, making it the most effective approach for general sentiment classification tasks involving sarcastic content.

6 Conclusion

This study distinguishes itself from conventional sentiment analysis by emphasizing sarcasm detection through the integration of audio features alongside textual data. This methodology has exhibited promising results in sentiment analysis, evidenced by its contribution to the enhanced accuracy of baseline sarcasm detection models through the inclusion of audio data. Moreover, the observed improvements in both accuracy and F1 scores underscore the effectiveness of integrating sarcasm features as supplementary elements, thereby refining the sentiment classification process. However, the achieved accuracy of 88.9% and the constraints posed by a limited dataset indicate substantial scope for further improvements and advancements in this field.

Limitation and Future Work

Dataset Expansion: The relatively small dataset curtails the model's capability to fully capture intricate sentiment patterns. Future efforts would involve collecting more datapoints, which would enhance the model's sensitivity to subtle nuances and also enable the use of more complex models without the risk of overfitting.

Contextual Analysis Enhancement: Enhancing sentence prediction accuracy through contextual analysis could significantly aid in identifying sarcasm, often marked by the juxtaposition of positive words in negative contexts. This modification would likely improve the overall efficacy of the sentiment classification model.

Model Expansion: Integrating LSTMs for text processing in the hybrid model could be advantageous, as they are ideal for sequential data and NLP tasks. Their initial exclusion due to feature engineering constraints suggests future research could explore alternative methods for text sequence extraction to support LSTM integration.

Multi-Class Sentiment Classification: Future iterations of the model could extend beyond binary to multi-class sentiment classification. Plans include the integration of specialized models that are adept at recognizing complex emotions, thereby enhancing classification accuracy and expanding the model's applicability to sectors that require nuanced emotional analysis.

References

Alyzehkazmi (2024). Sentiment analysis from audio: An exploration of different machine learning algorithms. Accessed: 2024-01-27.

Bouazizi, M. and Ohtsuki, T. (2016). Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in twitter. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE.

Luitel, S. and Anwar, M. (2022). Audio sentiment analysis using spectrogram and bag-of- visual-words. In *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 200–205.

Nandwani, P. and Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81–.

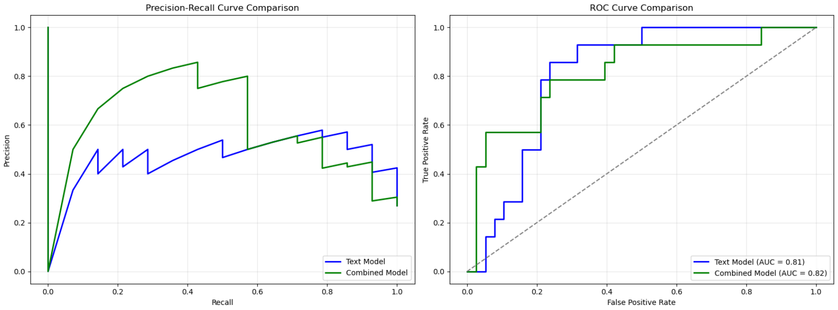
Soumya, S. and Pramod, K. (2020). Sentiment analysis of malayalam tweets using machine learning techniques. *ICT Express*, 6(4):300–305.

Wu, C.-H. and Liang, W.-B. (2011). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21.

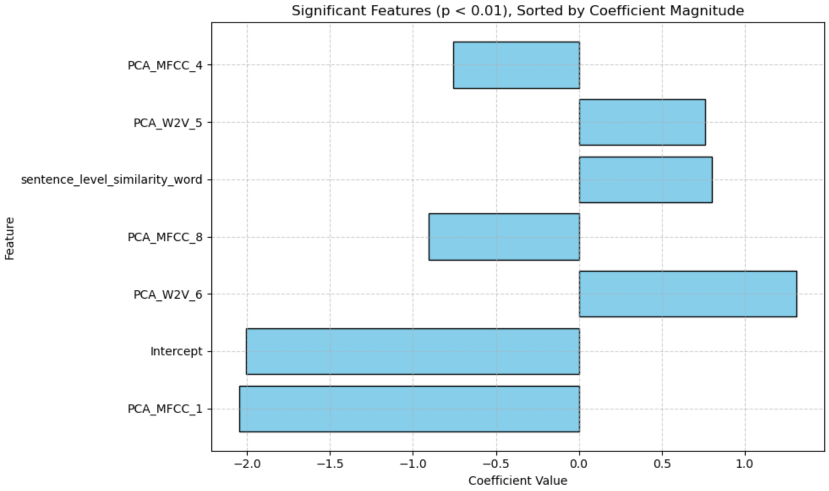
A Appendix

A.1 Features Results

A.1.1 Comparative Precision-Recall and ROC Curve: Text-Only vs. Text + Audio Prediction Models

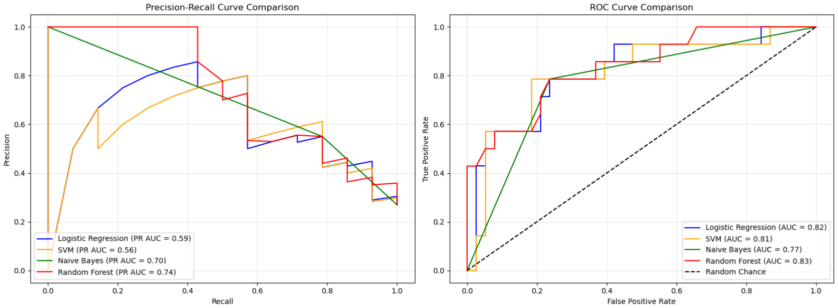


A.1.2 Feature Coefficient Value: Text + Audio Prediction Models

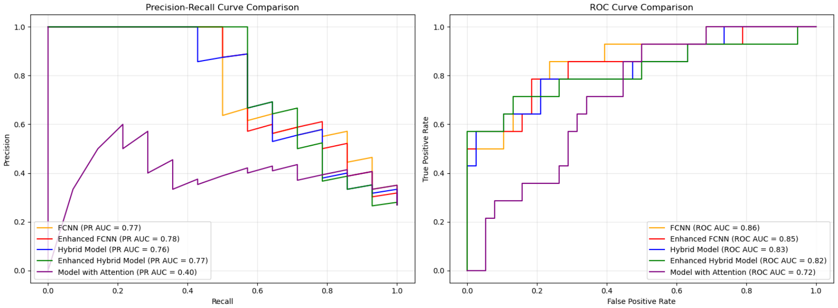


A.2 Model Result

A.2.1 Simpler Models Precision-Recall and ROC Curve



A.2.2 Deep Learning Models Precision-Recall and ROC Curve


Precision-Recall Curve Comparison
Precision
Recall
FCNN (PR AUC = 0.77)
Enhanced FCNN (PR AUC = 0.78)
Hybrid Model (PR AUC = 0.76)
Enhanced Hybrid Model (PR AUC = 0.77)
Model with Attention (PR AUC = 0.40)
ROC Curve Comparison
True Positive Rate
False Positive Rate
FCNN (ROC AUC = 0.86)
Enhanced FCNN (ROC AUC = 0.85)
Hybrid Model (ROC AUC = 0.83)
Enhanced Hybrid Model (ROC AUC = 0.82)
Model with Attention (ROC AUC = 0.72)