Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection

Atanu Mandal[1†] and **Gargi Roy**[2‡*] and **Amit Barman**[3†]
Indranil Dutta[4†] and **Sudip Kumar Naskar**[5†]
[†]Jadavpur University, Kolkata, INDIA
[‡]Optum Global Solutions Private Limited, Bengaluru, INDIA
{[1]atanumandal0491, [2]roygargi1997, [3]amitbarman811, [5]sudip.naskar}@gmail.com,
[4]indranildutta.lnl@jadavpuruniversity.in

Abstract

With the recent surge and exponential growth of social media usage, scrutinizing social media content for the presence of any hateful content is of utmost importance. Researchers have been diligently working since the past decade on distinguishing between content that promotes hatred and content that does not. Traditionally, the main focus has been on analyzing textual content. However, recent research attempts have also commenced into the identification of audio-based content. Nevertheless, studies have shown that relying solely on audio or text-based content may be ineffective, as recent upsurge indicates that individuals often employ sarcasm in their speech and writing. To overcome these challenges, we present an approach to identify whether a speech promotes hate or not utilizing both audio and textual representations. Our methodology is based on the Transformer framework that incorporates both audio and text sampling, accompanied by our very own layer called "Attentive Fusion". The results of our study surpassed previous state-of-the-art techniques, achieving an impressive macro F1 score of 0.927 on the Test Set.

1 Introduction

In recent years, the explosive growth of digital communication platforms has facilitated unprecedented levels of information exchange, enabling individuals from diverse backgrounds to interact and share ideas. However, this surge in online interactions has also led to the emergence of a concerning issue: the increase of hate speech (Davidson et al., 2017). Hate speech, characterized by offensive, discriminatory, or derogatory language targeting individuals or groups based on race, ethnicity, religion, gender, or sexual orientation, poses significant challenges to maintaining a safe and inclusive online environment (Schmidt and Wiegand, 2017).

Traditional methods of hate speech detection primarily focused on analyzing text-based content, leveraging natural language processing (NLP) techniques to identify offensive language patterns. While these approaches have yielded some success, they often struggle to capture the nuanced nature of speech, as the exact text might be interpreted differently when considering context, tone, and intent (Fortuna and Nunes, 2018). To address these limitations, researchers are turning to a more holistic approach combining both text and speech modalities to enhance the accuracy and robustness of hate speech detection systems (Rana and Jha, 2022).

Figure 1: Identification of "Hate" or "Not Hate" using multimodality approach

This multidimensional approach referred to as multimodal hate speech detection, leverages not only the textual content of messages but also the acoustic cues and prosodic features present in speech. By simultaneously analyzing both text and speech-based characteristics, this approach aims to capture a more comprehensive representation of communication, considering not only the words

* The work was carried out when the author was in Jadavpur University.

used, but also the emotional nuances conveyed through speech intonation, pitch, and rhythm. Figure 1 illustrates two examples each for “Hate” and “Not Hate” using the Multimodality. In the two cues shown (figure 1), represents the speech cue and represents the text cue.

In this paper, we investigate multimodal hate speech detection exploring the synergies between text and speech for identifying hate speech instances. We examine the challenges posed by hate speech in the digital age, the limitations of traditional text-based detection methods, and the potential advantages of integrating speech data into the detection process. By leveraging insights from various disciplines such as NLP, audio signal processing, and machine learning, multimodal approaches hold promise in achieving higher detection accuracy and reducing false positives, ultimately fostering safer and more inclusive online environments.

Our methodology sets itself apart from other state-of-the-art (SOTA) methodologies in the subsequent manner:

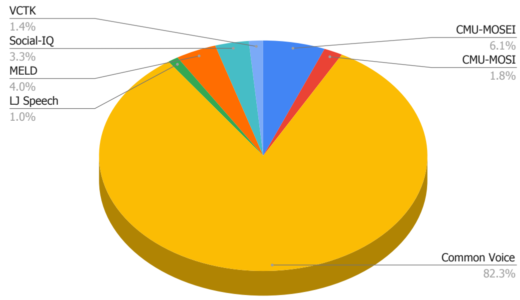
- Our system consists of a sequence of interconnected systems enclosing the Transformer framework[1].
- We have introduced a layer termed “Attentive Fusion” that augments the results.

2 Dataset Description

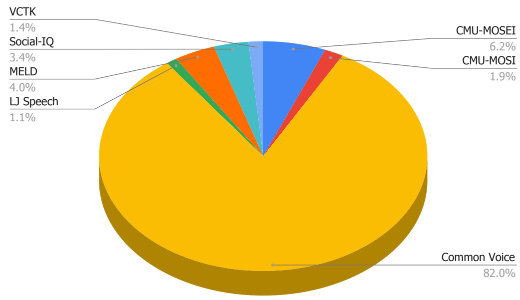
Dataset	Number of Samples		
	Train	Dev	Test
CMU-MOSEI (Bagher Zadeh et al., 2018)	597	133	130
CMU-MOSI (Zadeh et al., 2016)	181	40	39
Common Voice (Ardila et al., 2020)	8,050	1,768	1,733
LJ Speech (Ito and Johnson, 2017)	102	23	23
MELD (Poria et al., 2019)	393	87	85
Social-IQ (Zadeh et al., 2019)	325	74	69
VCTK (Yamagishi et al., 2019)	138	31	30
	9,786	2,156	2,109

Table 1: Statistics of the dataset used for Identification of Hatred.

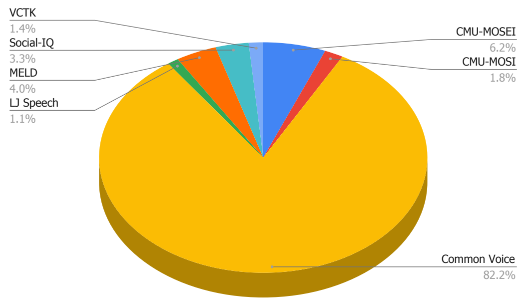
For our experiments, we used fragments of the DeToxy dataset (Ghosh et al., 2022)[2], a dataset for detecting Hatred within spoken English speech. This dataset is derived from diverse open-source datasets. The specifics regarding the number of samples utilized from various datasets are precisely outlined in Table 1.



(a) Training Data



(b) Development Data



(c) Test Data

Figure 2: Pictorial representation of the contribution of datasets

Our experiments were carried out on a comprehensive dataset that encompassed all seven datasets combined. Each dataset contained entries that fell into either the "Hate" or "Not-Hate" category, along with a transcription for each audio. To facilitate understanding, we have depicted the distribution

[1] Code is publicly available in GitHub.

[2] Ghosh et al. (2022) used 20,271 data consisting of CMU-MOSEI, CMU-MOSI, Common Voice, IEMOCAP, LJ Speech, MELD, MSP-Improv, MSP-Podcast, Social-IQ, Switchboard, and VCTK of which IEMOCAP, MSP-Improv, MSP-Podcast, Switchboard are not open-sourced therefore we were unable to use the dataset.

Dataset	Hate			Not Hate		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
CMU-MOSEI	149	33	35	448	100	95
CMU-MOSI	47	10	10	134	30	29
Common Voice	2,013	442	433	6,037	1,326	1,300
LJ Speech	28	6	6	74	17	17
MELD	99	22	21	294	65	64
Social-IQ	83	18	19	242	56	50
VCTK	34	8	8	104	23	22
	2,453	539	532	7,333	1,617	1,577

Table 2: Data Statistics of "Hate" and "Not Hate"

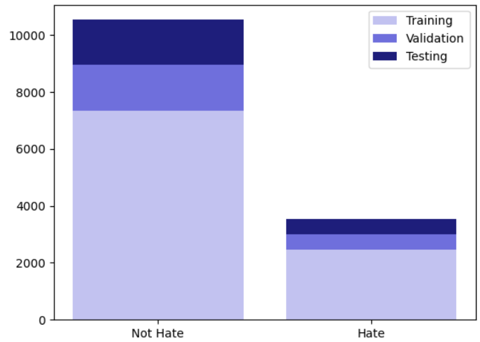


Figure 3: Sample count for "Hate" and "Not Hate"

of each dataset's contribution to our framework through a pie chart, as showcased in Figure 2. Figures 2a, 2b, and 2c accordingly illustrate the respective contributions of the training data, development data, and testing data. There exists a significant disparity in the number of samples across the various datasets but, the proportional representation of the training, development, and test datasets remains consistent. Notably, Common Voice comprises the majority of the data, while LJ Speech is the least represented. The statistical analysis of the "Hate" and "Not Hate" classes is presented in Table 2. Meanwhile, the bar plot showcasing the sample count for both classes can be seen in Figure 3. A comprehensive description of datasets is described in Section 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, and 2.7.

2.1 CMU-MOSEI

Carnegie Mellon University - Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) (Bagher Zadeh et al., 2018) is considered the largest and most extensive dataset for emotion recognition tasks and multimodal sentiment analysis. Figure 4a provides the information on the number of samples for "Hate" and "Not Hate" and Figure 5a provides the pictorial information of the number of samples to audio duration.

2.2 CMU-MOSI

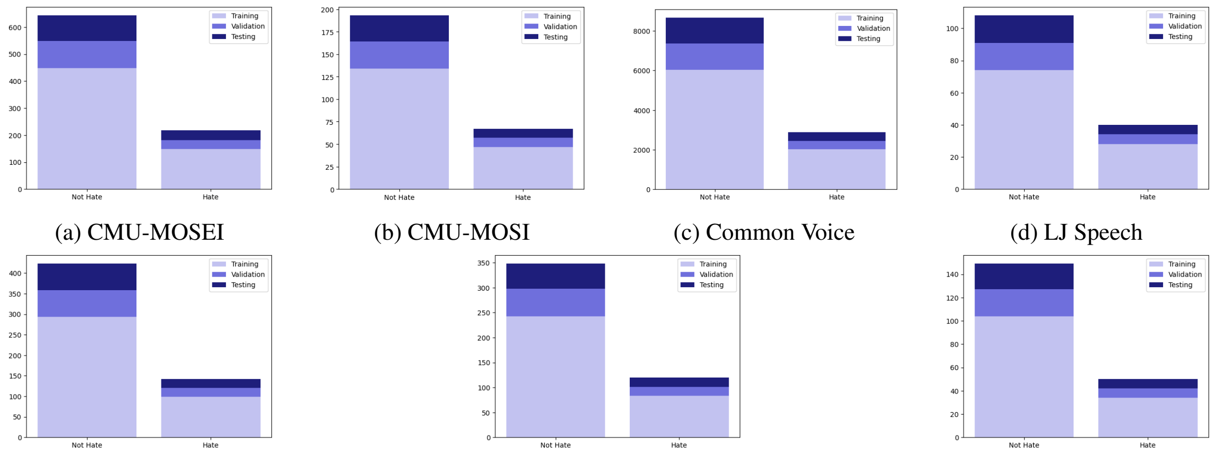
Carnegie Mellon University - Multimodal Corpus of Sentiment Intensity (CMU-MOSI) (Zadeh et al., 2016) is another dataset by Carnegie Mellon University, which consists of 2199 video clips of different opinions, annotated with sentiment. It is annotated in the range $[-3, 3]$, using various parameters for sentiment intensity, subjectivity, and per-millisecond annotations of audio features. It contains 97% non-toxic and nearly 3% toxic utterances. Figure 4b provides the information on the number of samples for "Hate" and "Not Hate" and Figure 5b provides the pictorial information of the number of samples to audio duration.

2.3 Common Voice

This dataset (Ardila et al., 2020) by Mozilla Developer Network is an open-source, dataset of voices of multiple languages for the use of training speech-enabled systems, with 20,217 hours of recorded audio and 14,973 hours of validated speech audios. Figure 4c provides the information on the number of samples for "Hate" and "Not Hate" and Figure 5c provides the pictorial information of the number of samples to audio duration.

2.4 LJ Speech

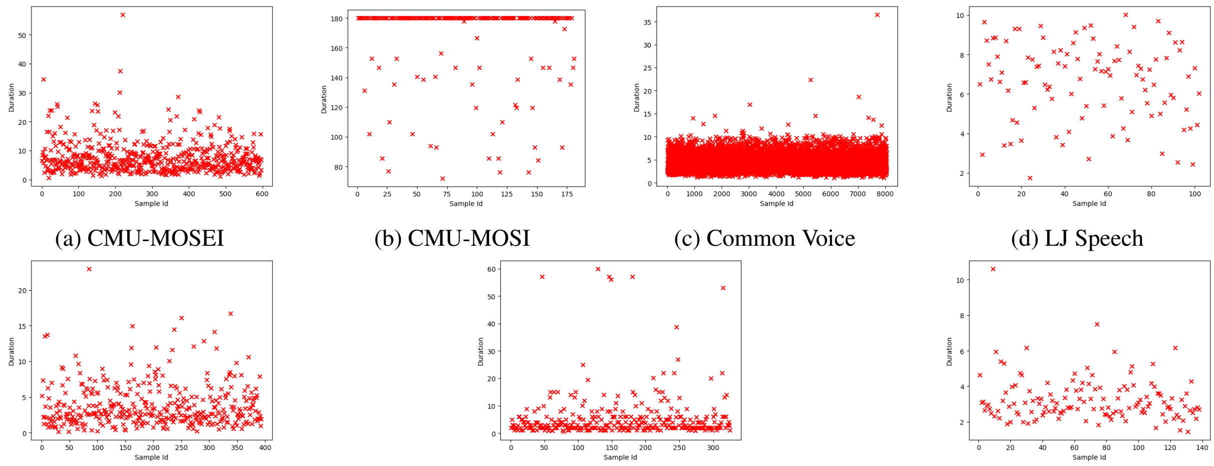
This is another open-source dataset (Ito and Johnson, 2017) which has 13,100 clips of short audio segments, with one speaker reading texts from a collection of seven books of non-fiction. Every clip is transcribed and has a varying length of 1 to 10 seconds. Figure 4d provides the information on the number of samples for "Hate" and "Not Hate" and



(a) CMU-MOSEI (b) CMU-MOSI (c) Common Voice (d) LJ Speech

(e) MELD (f) Social-IQ (g) VCTK

Figure 4: Sample count for “Hate” and “Not Hate”



(a) CMU-MOSEI (b) CMU-MOSI (c) Common Voice (d) LJ Speech

(e) MELD (f) Social-IQ (g) VCTK

Figure 5: Scatter representation of Datasets according to audio length

Figure 5d provides the pictorial information of the number of samples to audio duration.

2.5 MELD

Multimodal Emotion Lines Dataset (MELD) (Poria et al., 2019) has over 1,400 dialogues and 13,000 dialogues from the television show “Friends”. Each utterance in dialogue has been labelled by one of the emotions – Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. MELD also has annotations for sentiments – positive, negative, and neutral. Figure 4e provides the information on the number of samples for “Hate” and “Not Hate” and Figure 5e provides the pictorial information of the number of samples to audio duration.

2.6 Social-IQ

Another dataset (Zadeh et al., 2019) by Carnegie Mellon University has videos that are thoroughly validated and annotated, along with questions, answers, and annotations for the level of complexity of the said questions and answers. Figure 4f provides the information on the number of samples for “Hate” and “Not Hate” and Figure 5f provides the pictorial information of the number of samples to audio duration.

2.7 VCTK

The VCTK corpus (Yamagishi et al., 2019) contains 110 speakers’ speech data spoken in English, having various accents. Every single speaker reads a passage, selected from newspapers, archives, and so on. Figure 4g provides the information on the number of samples for “Hate” and “Not Hate” and Figure 5g provides the pictorial information of the number of samples to audio duration.

3 Experiments

This section demonstrates our innovative techniques for detecting Hatred within a speech. The section is divided into numerous subsections for understanding our approach with ease. Section 3.1 presents the methods we used to prepare the dataset for our suggested framework. Section 3.2 describes our suggested framework. Section 3.3 discusses the parameters used for our proposed framework and Section 4 discusses the results of our approach with other benchmark frameworks.

3.1 Dataset Pre-processing

In the task of pre-processing, we carefully selected the data that possessed comparable lengths of audio. We disregarded instances with excessively long or short duration. Our inclination to overlook excessively long audio duration stemmed from the understanding that it would necessitate extensive computational resources. Conversely, audio with extremely short duration lacked the richness of audio features.

	Values
Sample Rate	16,000 Hz
Number of FFT	400
Number of MELs Channel	80
Hop Length	160
Chunk Length	30
Number of Samples	4,80,000
Number of Frames	3,000
Number of Samples per Token	320
Frames per Second	10 ms
Tokens per Second	25 ms

Table 3: Audio feature extraction parameters

We conducted a series of experiments with our framework, exploring various methods of extracting features such as Mel-frequency cepstral coefficients (MFCCs) and filter banks. However, we discovered that the "log mel spectrogram" yields superior accuracy in comparison to other alternatives, as it captures auditory information in a manner akin to human perception. To extract these features, we established the optimal parameters empirically, which are detailed in Table 3.

For feature extraction from text, we used the pre-trained Albert Tokenizer (Lan et al., 2020) from IndicBART (Dabre et al., 2022) developed by AI4Bharat. To tokenize each sentence, we introduced the symbols "$< s >$" and "$< /s >$" at the start and end, respectively, signifying the commencement and conclusion of the sentences.

3.2 Framework

We have used the Transformer (Vaswani et al., 2017) framework which has gained widespread recognition and is considered SOTA in the domains of Speech Recognition and Machine Translation (MT) due to its exceptional ability to handle the complexities of these complex tasks. To provide a clear overview of our methodology, Figure 6 presents an overview of our framework.

The Speech Feature is extracted by the "log mel spectrogram" technique, which has been discussed in section 3.1. This technique involves the computation of a spectrogram that represents the frequency content of an audio signal over time, using a logarithmic scale for the frequency axis. The resulting spectrogram has a dimension of "$(80 \times time\_step)$" and is then passed to the Speech Sampling Block. The Speech Sampling Block is responsible for selecting a subset of the input spectrogram, based on certain criteria (described in Section 3.2.1). On the other hand, the tokenized Text, which is obtained through a process described in section 3.1, has a dimension of "$(max\_length \times 1)$" and is passed to the Text Sampling Block (discussed in Section 3.2.2). The Text Sampling Block performs a similar function as the Speech Sampling Block but on the tokenized Text instead of the spectrogram.

The resulting subset of Speech Sampling is fed to the Encoder of the first Transformer module and the Decoder of the second Transformer module. Similarly, the resulting subset of Text Sampling is fed to the Decoder of the first Transformer module and the Encoder of the second Transformer module. The motivation behind this approach is to investigate whether the text in the Decoder can learn from the audio in the Encoder, and vice versa. This is inspired by the idea of MT, where the target text in the Decoder learns from the source text in the Encoder. By applying this concept to the audio and text domains, we aim to explore the potential for cross-modal learning and the transfer of knowledge between different modalities.

To further process the outputs of the two Transformer modules, we introduce a Long short-term memory (LSTM) block that consists of a single LSTM layer. This LSTM block is responsible for sequentially learning the knowledge from each step

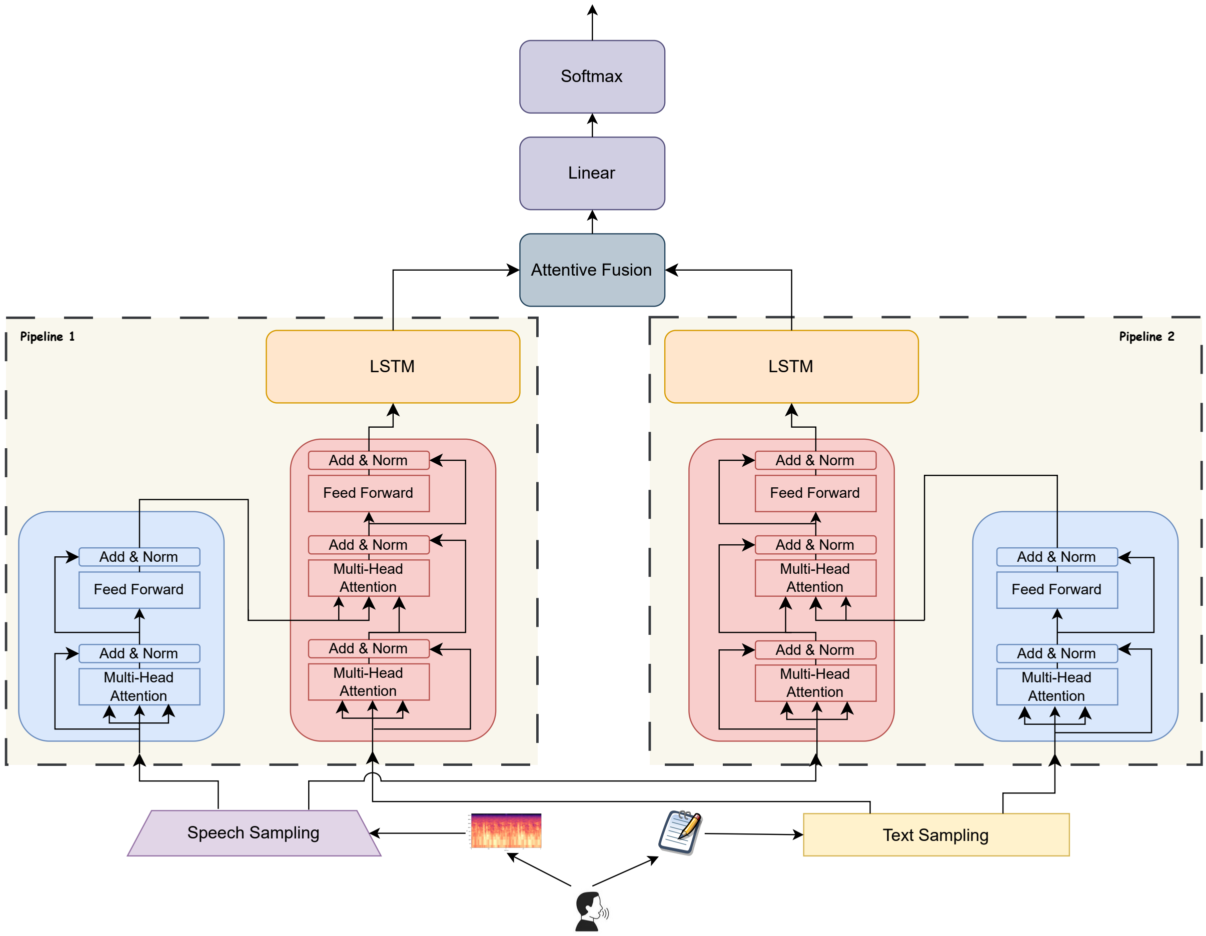


Figure 6: Overview of our approach

of the output. After going through this process, we obtain two outputs: one from the first LSTM and another from the second LSTM. The combination of the first Transformer with the first LSTM is represented as "Pipeline 1" and the combination of the second Transformer with the second LSTM is represented as "Pipeline 2". These two outputs are then passed to the proposed "Attentive Fusion" Layer (described in Section 3.2.3). The Attentive Fusion Layer is designed to learn the knowledge from both outputs in a joint manner, combining the information from the two pipelines. The output of the Attentive Fusion Layer is then fed to a Linear Layer with Softmax activation, where it undergoes further processing and classification according to the specific classes.

A comprehensive exploration of the Audio Sampling Module, Text Sampling Module, and our proposed Attentive Fusion layer are presented in Section 3.2.1, 3.2.2, and 3.2.3, respectively. All hyperparameter configurations can be found in Section 3.3.

3.2.1 Speech Sampling

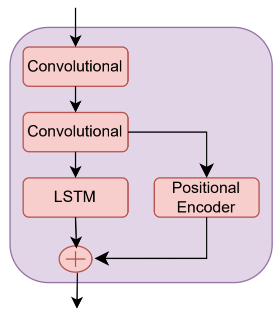


Figure 7: Overview of Speech Sampling

Our module for "Speech Sampling" was influenced by the work of Radford et al. (2022) with minor modifications. This module comprises a pair of Convolutional layers, with each layer being accompanied by a Gaussian Error Linear Unit (GELU) activation function. The outcome of the Convolutional layer was passed through a Positional Encoder and an LSTM layer separately. The results from the Positional Encoder and LSTM were combined. The Speech Sampling framework is shown in Figure 7.

3.2.2 Text Sampling

Our "Text Sampling" module comprises a simplistic approach containing Word Embedding and a Positional Encoder. The raw text was tokenized by appending with "$< s >$" and "$< /s >$" at the commencement and conclusion of the sentences (refer to Section 3.1) and passed on to Word Embedding. The subsequent output is then directed to the Positional Encoder. Subsequently, the output of the Word Embedding and the Positional Encoder are combined and conveyed to the subsequent hierarchical module. The representation of the "Text Sampling" framework can be seen in Figure 8.

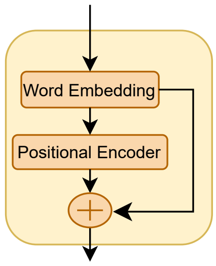


Figure 8: Overview of Text Sampling

3.2.3 Attentive Fusion Layer

The layer we have named the "Attentive Fusion" layer is a layer that we have devised for the purpose of detecting hatred within a speech. In our methodology (as illustrated in equation 1), we have seamlessly integrated the outcomes from Pipeline 1 and Pipeline 2, allowing them to flow into their respective Linear layers individually, thereby ensuring the preservation of their unique characteristics.

$$\begin{aligned} L_1 &= Linear(x_1) \\ L_2 &= Linear(x_2) \end{aligned} \tag{1}$$

The result, L_1 and L_2 underwent a process of cross multiplication, after which a hyperbolic tangent function ($tanh$) was used. To enhance the disparity of each tensor value (w_i), an exponential (e) function was applied, as demonstrated in equation 2.

$$w_i = e^{(tanh(L_1 \times L_2))} \tag{2}$$

The outcome of equation 2 w_i underwent division by the summation of every element of w_i. To prevent division by zero, we introduced an epsilon (ϵ). The entire outcome was then subjected to multiplication with w_i itself. Equation 3 illustrates our approach.

$$w_i^{'} = \frac{w_i}{\sum_i w_i + \epsilon} \times w_i \tag{3}$$

The value, $w_i^{'}$ obtained from equation 3 was introduced into the subsequent module that incorporates a Linear Layer to differentiate between different classes.

3.3 Hyperparameters

3.3.1 Speech Sampling

For the two Convolutional layers, we used filter sizes of "4096" and "1024", respectively and kernel size of "3" for both. Strides of "1" for 1^{st} and "2" for 2^{nd} Convolutional layer was used. For LSTM layer units of "512" with activation function "tanh" and recurrent activation of "sigmoid" was used. For the Positional Encoder vocab size of "64,014", the hidden dimension of "512" was passed.

3.3.2 Text Sampling

For the Word Embedding, we used a vocab size of "64,014", and a hidden dimension of "512" with True mask zero. For the Positional Encoder vocab size of "64,014", the hidden dimension of "512" was passed.

3.3.3 Transformer

For the transformer framework, the number of heads and the number of layers for the Encoder and Decoder were kept "4". The hidden dimensions were kept "512", and the dropout (Srivastava et al., 2014) rate of "0.3".

3.3.4 LSTM

For the LSTM layer, units of "512" with activation function "tanh" and recurrent activation of "sigmoid" were used, with a dropout rate of "0.3". Use of bias and Forget bias with return sequence kept True.

3.3.5 Learning Rate

We use "AdamW" optimizer (Loshchilov and Hutter, 2019) with β_1 of "0.9", β_2 of "0.98", ϵ of "1×10^{-6}" and decay of "0.1" with adaptive learning rate.

$$\begin{aligned} arg_1 &= \sqrt{cs} \\ arg_2 &= cs \times ws^{-1.5} \\ lr &= \sqrt{d_{model}} \times \min(arg_1, arg_2) \\ lrate &= \min(lr, 0.0004) \end{aligned} \tag{4}$$

In equation 4, cs is current step, warmup steps (ws) were set to "2048" and d_{model} of "512".

4 Results

Table 4 provides the benchmark macro F1 scores of the frameworks proposed by Ghosh et al. (2022). The proposed frameworks were trained only with the audio sequences. Our study suggests that only the audio sequences cannot provide a better understanding of hatred in speech. Current trends show that persons use hateful words in spoken sentences but the tones, frequency and amplitudes are kept normal, which can also be remarked as "Sarcasmic Behaviour". To overcome the situation we used multimodality where an audio specimen along with its transcripts are used. Using multimodality has an increase in macro F1 score compared with the F-Bank framework proposed by Ghosh et al. (2022).

In cases of unfrozen wav2vec-2.0. the differences are very nominal as wav2vec-2.0 provides an embedding knowledge of each token of the audio specimen. In contrast, we didn't use any embedding knowledge of speech tokens, which will be experimented with in our upcoming works. The researcher has shown two wav2vec-2.0 among which is one wav2vec-2.0 (9 layer). In this system, the researcher took the representation token from the 9^{th} layer.

System	Category	Dev	Test
F-Bank	-	0.610	0.620
wav2vec-2.0	Freezed	0.448	0.457
	Unfreezed	0.877	0.869
wav2vec-2.0 (9 layer)	Unfreezed	0.897	0.877
Proposed Framework	-	**0.931**	**0.927**

Table 4: Evaluation and Result of Different Systems Proposed by Ghosh et al. (2022). *Unfrozen wav2vec-2.0 setup with representations taken from the 9^{th} layer as reported by Ghosh et al. (2022)*

5 Ablation Studies

5.1 "Attentive Fusion layer" vs "Concatenate layer"

Instead of using our proposed Attentive Fusion Layer, we used the Concatenate Layer to study the effectiveness of the Attentive Fusion layer and found it even outperformed the Concatenate layer. The differences in results are shown in Table 5.

	Dev	Test
Concatenate Layer	0.908	0.909
Attentive Fusion Layer	**0.931**	**0.927**

Table 5: Macro F1 Score Result on replacing "Attentive Fusion layer" with "Concatenate layer"

5.2 Using Pipeline 1 and Pipeline 2 separately

We also checked whether alone text learning from audio representation or audio learning from textual representation outperformed our baseline result or not. For the evaluation, we used Pipeline 1 and Pipeline 2 separately followed by Linear Layer, and found that it is unable to score at par result to our baseline. The differences in results are shown in Table 6.

	Dev	Test
Pipeline 1	0.910	0.909
Pipeline 2	0.910	0.899
Our Baseline	**0.931**	**0.927**

Table 6: Macro F1 Score Result on Pipeline 1 and Pipeline 2 separately and compared with our baseline.

6 Conclusion

In this work, we proposed a framework that can classify whether a speech promotes Hatred or not. For the speech feature extraction, we used a log mel spectrogram feature extraction technique. Our framework consists of Speech Sampling and Text sampling followed by two separate transformer frameworks that serve different efforts. Each Transformer framework is followed by an LSTM layer, the output of which is fed to our proposed layer, and further sent to Linear Layer for Classification. The whole framework was able to outperform the existing benchmark macro F1 score. The only limitation of our approach is it is limited to the English language. In future work, we would like to test its robustness for other languages.

Acknowledgements

This research was supported by the TPU Research Cloud (TRC) program, a Google Research initiative.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Sreyan Ghosh, Samden Lepcha, S Sakshi, Rajiv Ratn Shah, and Srinivasan Umesh. 2022. DeToxy: A Large-Scale Multimodal Dataset for Toxicity Classification in Spoken Utterances. In *Proc. Interspeech 2022*, pages 5185–5189.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *ArXiv*, abs/2202.06218.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8799–8809.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.