Article

An Ensemble Learning based Technique for Bimodal Sentiment Analysis

Shariq Shah 1,*, Hossein Ghomeshi [1], Edlira Vakaj [1], Emmett Cooper [1] and Rasheed Mohammad [1]

[1] School of Computing and Digital Technology, Birmingham City University, 15 Bartholomew Row, Birmingham, B5 5JU, United Kingdom
* Correspondence: shariq.shah@bcu.ac.uk

Abstract: Communication is a key method of expressing one's thoughts and opinions. Amongst many modalities, speech and writing are the most powerful and common forms of human communication. Analysing what and how people think has inherently been an interesting and progressive research domain. This includes bimodal sentiment analysis which is an emerging area in natural language processing (NLP) and has received a great deal of attention in recent years in a variety of areas including social opinion mining, health care, banking, and so on. At present, there are limited studies on bimodal conversational sentiment analysis as it proves to be a challenging area given the complex nature of the way humans express sentiment cues across various modalities. To address this gap, a comparison of the performance of multiple data modality models has been conducted on the MELD dataset, a widely-used dataset for benchmarking sentiment analysis within the research community. Our work then demonstrates the results of combining acoustic and linguistic representations. Lastly, our proposed neural network-based ensemble learning technique is employed over six transformer and deep learning-based models, achieving a State-Of-The-Art (SOTA) accuracy.

Keywords: ensemble learning; bimodal; sentiment analysis; neural network; transformer

1. Introduction

As the digital world advances, there has been an increase in the number of day-to-day interactions over proliferating forms of communication, including newer, nuanced products such as Siri, Alexa, Google Assistant, known as Virtual Personal Assistants (VPA), and the Internet of Things (IoTs) applications [1], [2]. The two key drivers of the digital revolution are Moore's Law – the exponential increase in computing power and solid-state memory, and the substantial progress in enhancing communication bandwidth [3]. This, in fact, has risen the expectations and demands of customers. As customers become more adept and demand higher quality products and services, it is becoming a challenge to keep evolving customer expectations satisfied. Technology has played a major role in the evolution of customer demands since it has ushered in and fostered the 'instant gratification' culture. Customers are better informed with easy access to information and are able to gain access instantly from anywhere.

With the growing advancement and adoption of technology, new opportunities and challenges arise and sentiment analysis becomes an even more important subject. For businesses, it has become vital to understand their customer's thoughts, feelings, and behavior, in order to better model their strategy and align their offerings accurately with customer demands. Advances in Machine Learning (ML) have improved bimodal (employs more than one data modality) and multimodal (employs more than two data modalities) sentiment analysis substantially [4], [5]. Detecting bimodal sentiment can be an important aspect of many applications, mostly in Contact Centres (CCs), which are the first point of contact with organisations for most customers. Although very few studies have addressed the detection and analysis of bimodal sentiment in CC conversations, interest in this area is steadily gaining traction from vertical organisations and the NLP community at large to analyse CC conversations that take place on different communication channels [6], [7]. However, the data produced in CCs are in the form of unstructured data that cannot be fed into an algorithm directly. Compounded with the fact that NLP is nuanced and subjective, the problem becomes much more difficult to solve.

The two primary research areas in bimodal sentiment analysis are (1) how to represent data modalities and (2) how to fuse them together. The recommendation for the former is that a good representation of raw data should capture sentiment features that can be generalised over distinctive semantic content. For the latter purpose, it is recommended to have a fusion mechanism that effectively combines audio and text representations [8]. To represent text and audio data modalities, various low-level features, often referred to as low-level descriptors (LLDs) have been employed previously, such as Word2Vec, GloVe, Mel-frequency Cepstral Coefficients (MFCC), Log-Frequency Power Coefficients (LFPC), energy, pitch, and log-Mel spectrograms [9], [10], [11], [12], [13]. These features have been mostly used as input to models such as Long Short Term Memory (LSTM), Convolutional Neural Networks (CNNs), Hidden Markov Models (HMMs), Recurrent Neural Networks (RNNs), and Deep Neural Networks (DNNs). Most of the prior work used both low-level features and features extracted from Deep Learning (DL) models [14], [15], [16], [17], [18], [19], [20], [21].

In contrast to previous work, this study shows the significance of pre-trained model representations of two modalities (audio and text). We also evaluate the performance of the models following the fusion of text and audio embeddings. The main contribution is the use of the ensemble learning technique, where multiple classifiers are combined using various techniques to establish a well-trained single classifier. Ensemble techniques have demonstrated superiority over techniques in various classification tasks. This is due to their flexibility in training and updating classifiers.

The rest of this paper is organised as follows: Section 2 presents an overview of related research; Section 3 details our methodology; Sections 4, 5, and 6 outlines the experimental setup and results, and draw up a comparison of the state-of-the-art methods used, including the ensemble learning method presented in this paper; Section 7 presents the conclusions and outlines possible future work.

2. Background and Related Work

In this section, we first introduce and briefly discuss the feature extraction approaches used in bimodal sentiment analysis. The theory underlying the models used in this research is then introduced and discussed. Finally, we highlight work closely related to the sentiment classification task associated with data from multiple modalities, specifically sentiment analysis in conversations, particularly the architecture and fusion approaches employed.

2.1. Feature Extraction Approaches

The majority of studies have used a mix of low-level and deep features for bimodal sentiment analysis. The algorithms employed in bimodal sentiment analysis usually involve both feature extraction and fusion methods, since the data vary fundamentally in terms of origin, sequence, and mechanism. This section provides an overview of the various feature extraction approaches used previously.

2.1.1. Low-Level Features

Traditional features such as MFCC, log-Mel spectrograms, and LPCC are popular low-level descriptors (LLDs) used for representing the acoustic content of speech [22], [1]. Acoustic features can be broadly classified into two categories: time-domain features and frequency-domain features. Time-domain features include the short-term energy of the signal, zero crossing rate, maximum amplitude, minimum energy, and entropy of energy [12]. Where there is limited data, frequency domain features reveal deeper patterns in the audio signal, which can potentially help in identifying the emotion underlying the signal. Frequency-domain features include spectrograms, MFCCs, spectral centroid, spectral rolloff, spectral entropy, and chroma coefficients. LFPC can be classified as both time and frequency domain features. Traditional Word2Vec-based models such as Continuous Bag of Words (CBOW), Skipgram, GloVe, and ELMo have been previously used for text feature representation [10], [23], [24], [25]. More recently, BERT Transformer encoder-only models have been used and have been shown to capture meaningful features for text classification tasks [26].

2.1.2. Deep Features

The features extracted using pre-trained DL models are referred to as deep features. Typically, these models are initially trained with one or more than one large, annotated dataset. In previous research works, pre-trained models have been used to extract speech and textual features for the task of sentiment analysis [27], [28], [22], [27], [29]. Such works suggest that deep features can be a better choice in terms of accuracy when compared to low-level features.

2.2. Summary of Models Used in Bimodal Sentiment Analysis

Following our extensive research, we shortlisted three models for this research. We fine-tuned the models, which involved feature extraction, training, and evaluation for each data modality.

2.2.1. RoBERTa

The Robustly optimized BERT approach (RoBERTa) is a pretrained language model, an extension of the original BERT model which has proven to have shown substantially better results according to the General Language Understanding Evaluation (GLUE) benchmark for evaluating natural language understanding systems [30], [26], [31]. The key difference between BERT and RoBERTa is in the training phase of the model. It does not rely on next-sentence prediction and masking is performed during training time, as opposed to during data preparation time for BERT. The 'roberta-base' version of the model is downloaded using the transformers library that loads the model from the open-sourced repository. The model version used has 110M parameters and has been pretrained on a larger English language dataset of 160 GB. With its 24-layer encoder architecture, longer sequences can be used as input [30]. However, the maximum number of tokens remains limited to 512 tokens, similar to the limit for BERT, which are then mapped to an embedding size of 1024 [8].

2.2.2. Wav2Vec 2.0

Wav2Vec 2.0 is one of the current SOTA models pre-trained in a self-supervised setting, similar to BERT's masked language modelling [32], [33]. The architecture is made up of two convolutional neural networks; encoder and context networks. The encoder network $f : X \mapsto Z$ takes input raw audio samples $x_i \in X$ and outputs low frequency feature representations $(z_1, z_2, ..., z_T)$ which encode about 30ms of 16kHz audio every 10ms. The context network $g : Z \mapsto C$ converts these low-frequency representations into a higher-level contextual representation $c_i = g(z_i ... z_{i-v})$ for a receptive field v [34], [35]. The overall receptive field after passing through both networks is 210ms for the base version and 810ms for the large version. The main aim of this model, as reported in their first paper, is to improve Automatic Speech Recognition (ASR) performance with fewer labelled training data and enable its use for low-resource languages. The model consists of 35M parameters and was pretrained on 960hrs of unannotated speech data from the LibriSpeech benchmark audiobooks data [36]. The authors set the embedding size to 512 and the maximum audio waveform length to 9.5 seconds.

2.2.3. 2D CNN

Over the years, Convolution Neural Networks (CNNs) have made substantial progress in image recognition tasks as they are good at automatically learning useful features from high-dimensional images [37]. CNNs use shared kernels (weights) to exploit the 2D correlated image data structure. Max pooling is added to CNNs to introduce invariance; thus, only the relevant high-dimensional features are used in classification tasks [38], [39]. This study explores the assumption that CNNs can work well with audio classification tasks due to the fact that when the log Mel filter bank is applied to the Fast Fourier Transform (FFT) representation of raw audio, linearity is produced in the log-frequency axis allowing a convolution operation to be performed along the frequency axis. Otherwise, different filters (or kernels) would have to be used for different frequency ranges. This property along with CNN's good representation power allows the model to learn the underlying patterns effectively within short time frames, resulting in superior performance [1], [21], [40]. To put it simply, the intuition here is to consider the audio segments as input images. The CNN layer will identify local contexts by applying n convolutions over the input audio images along the time axis and generate sequences of vectors. In our paper, we employed two 2D CNNs - one trained on MFCCs and the other on log-Mel spectrograms matrices.

2.3. Sentiment Analysis Architecture and Fusion Approaches

Over the years, sentiment analysis research has shifted from analysing full documents or paragraphs to a finer level of detail - identifying sentiment towards particular phrases or words, audio and visual cues [7], [41]. Correspondingly, sentiment analysis has gained much more research interest, mainly because of its potential application in dialogue systems to produce sentiment-aware and considerate dialogues [41]. However, studies using real-life conversational data are scarce. While this area is attracting a plethora of research work focusing on algorithmic aspects, such studies are typically evaluating a selection of datasets and little effort is dedicated to the expansion of the research scope where bimodal data is explored within the setting of conversational datasets [6]. In this section, we briefly discuss the previous studies on the two different approaches to data modality sentiment analysis.

2.3.1. Bimodal

The bimodal approach utilises two modal input representations to judge the sentiment of each utterance. Current bimodal-based approaches are scarce and few of those who have focused on this have mainly preferred to use textual and acoustic modalities for sentiment analysis tasks.

In [42], a method was used that statistically combined features with N-gram, sentiment words, and domain-specific words to predict user sentiments. Their work applied a combination of acoustic and linguistic rules through a multi-dimensional model on CC data using SVMs, Maxent Entropy, and traditional Bayesian as classifiers. The main contribution of their work is the approach they took to incorporate the results from each of the classifiers while also adding language and acoustic rules to it. The FI score of their proposed method improved to 69.1%, against the baseline F1 score of 65.4%. In [22], a novel transfer learning method was proposed to be used when there is training data, as few as 125 examples per emotion class. Their method is comparable to that in our study, as they combine pre-trained embeddings for both text and audio. In their experiment, sub-words from bert-base and wav2vec2-large-960h representations are aligned through an attention-based recurrent neural network. Their method reported better performance compared to previous SOTA and frequently cited works using feature representations such as LLDs and GloVe embeddings and pre-trained ASR representations. Correspondingly, both audio and textual pretrained representations were fused through an attention mechanism over a bidirectional recurrent neural network - BiLSTM in [21]. The audio features extracted were 34-dimensional feature vectors from each frame, including MFCC, and zero-crossing rate, among others, and GloVe embeddings were used as textual features. Several other works explored fusing bimodal information - linguistic and acoustic. In [29] and [18], an approach was adopted to combine utterance-level audio and textual embeddings before the softmax classification layer. In [27], a different approach was followed which used pre-trained representation from an ASR model with semantic information.

Another study, which was distinct yet relatable to our work, explored the classification of psychiatric illnesses, initially through single data modality (audio and text) and then through hybridization of both modalities [13]. Text features from the RoBERTa model and speech features including MFCC were used. In their hybrid model, a "late fusion" approach was highlighted with a fully connected neural network layer to map the outputs from both models. Their results indicated that their proposed hybrid text model outperformed the single data modality models.

2.3.2. Multimodal

Multimodal methods detect sentiments in conversations through analysis of more than two modal input representations. As the information contained in unimodal data can often be biased and interfered with external noise factors, researchers on multimodal sentiment analysis tasks are receiving widespread attention for their unique approaches to combining different modalities [43]. Although multimodal sentiment analysis is not the scope of this paper, key lessons have been drawn from the studies mentioned in this section.

A wide range of work has been conducted previously, and researchers have proposed models based on DNNs, RNNs, CNNs and LSTMs over multiple data modalities [44], [45], [46] with varying fusion techniques [47], [48], [49], [50], [51], [52]. Recently, the effectiveness of novel DL architectures such as Transformers [53] and Graph Convolution Nets [15] as fusion methods have been explored and highlighted as computationally efficient. The above-mentioned works have used low-level features as input to their models. In [54], BERT-based self-supervised learning (SSL) features were used for text, while other modalities were represented with low-level features, and fusion mechanisms were based on RNN and Self Attention. In the work of [8], pre-trained SSL models were used as feature extractors and a transformer-based fusion mechanism was employed. In contrast to our work, their proposed fusion mechanism represents audio, text, and video modalities. Due to the SSL features being highly dimensional in terms of the size of embedding and large in terms of sequence length, in considering three modalities, they highlight their fusion mechanism as being both efficient and more accurate than other previous SOTA methods. Another proposed fusion method is GraphMFT, where Graph Neural Networks (GNNs) are leveraged to integrate and complement the information from multimodal data [43]. The results indicate that this method achieved better performance than previous SOTA approaches. Other relevant multimodal approaches previously proposed include ConGCN [55], MMGCN [56], MFN [57], ICON [58], BC-LSTM [52], CMN [59], and DialogueRNN [60].

3. Methodology

In this section, we focus on the public benchmark dataset and the features used for comparison.

3.1. Dataset Selection

MELD is a multimodal and multi-party dataset for Emotion Recognition in Conversations [61]. It is an extended version of the EmotionLines [62] dataset and contains dialogue instances similar to those available in EmotionLines, but, unlike EmotionLines, it includes information in the form of text, video, and audio. It comprises over 1400 conversations, a total of 13700 utterances, which are categorized into seven emotions: Neutral, Surprise, Fear, Sadness, Joy, Disgust, and Anger. The utterances are also categorized into three labels (positive, negative, and neutral) which are those in our experiments. There are three or more speakers in each conversation and the extracts are collected from the TV show called 'Friends'. The audio part of the dataset was retrieved from converting MPEG-4 Part 14 files into a WAVE format. Each audio utterance was stored in a 32-bit PCM WAVE format sampled at 16,000 Hz. The training and test set in our experiments consist of 9988 and 1108 audio files and textual utterances, respectively. The textual data is a Comma-Separated Values (CSV) file with two columns: text and sentiment label.

3.2. Feature Selection

Following our extensive analysis, several features were shortlisted. However, we restricted ourselves to using only a few which will be explained in this section.

3.2.1. Text Features

A bidirectional transformer model - RoBERTa-base was used for generating word embeddings as input for our text-based sentiment classification task. Transformer models use word embeddings as input similar to word2vec; however, the models can handle longer input sequences and the relations within these sequences. This ability, combined with the attention mechanism described in the original "attention is all you need" BERT paper [26] enables it to find long-range dependencies in the text, leading to more robust language models. As explained above, RoBERTa is an enhanced version of BERT, as it is much better trained and designed which reduces the overall training time required [30].

3.2.2. Audio Features

MFCC and Mel-Spectrograms were extracted for audio-based sentiment classification tasks. MFCC represents the short-term power spectrum of a sound by transforming the audio signal to mimic the human cochlea. The Mel scale as opposed to linear scales approximates human-based perception of the sound [63]. The filter-source theory defines the source to be the vocal cords and the filter represents that vocal tract. The length and shape of the vocal tract determine how sound is produced by a human and the cepstrum can describe the filter, i.e., represent sound in a structured manner [64]. Mel-Frequency Cepstral Coefficients (MFCC) are coefficients that capture the envelope of the short-time power spectrum. On the other hand, a spectrogram represents the frequency and time of an audio signal. The Mel-spectrogram is a representation of the audio signal on a Mel scale. The logarithmic form of the Mel-spectrogram helps in understanding emotions better because humans perceive sound on a logarithmic scale. As different emotions exhibit different MFCC and Mel-spectrogram patterns, their represented images were used individually as input to deep learning 2D CNN network, thus, helping us to evaluate which of the two feature representations is better suited to classify the emotion with audio data.

In relation to Wav2Vec 2.0, the features are extracted using a latent feature encoder that converts the raw waveform into a sequence of feature vectors every 20 milliseconds. This is then fed to the context network (transformer encoder) and processed through 12 and 24 transformer blocks for the base and large versions, respectively. The dimension size increases from 512 (output of the CNN) to 768 for the base and 1024 for the large as the input goes through a feature projection layer.

4. Experiments

In this section, we describe several experiments conducted in this research, including all details related to implementation. We classify them into; textual, acoustic, hybrid, and ensemble learning.

4.1. Implementation Details

This section presents the details of the model's implementation and the experimental setup. We implemented our model by following the official guidance on Hugging Face's transformer library [65] - Python library providing detailed code explanations for building BERT-based models [66]. A sequential data processing and modelling framework was built using mainly the PyTorch and Keras libraries. Other libraries that were used for extracting acoustic features and model training included TensorFlow, scikit-learn, and Librosa. The training was conducted using a single NVIDIA Quadro T1000 GPU, and the operating system was Windows 10. For evaluation purposes, we calculated accuracy, loss, precision, recall, and f1-score and generated a confusion matrix graph.

4.2. Textual

In this experiment, we extracted features from the MELD text data modality and used them as input into the model as described in section 2.2.1. The features extracted are vectors represented in tensors: 'input ids' and 'attention_mask'. The model also adds $[SEP]$ which is a marker for the ending of a sentence and $[CLS]$ to the start of each input, so the transformer model knows it is a classification problem. In addition to that, a special token called $[PAD]$ is also added, which defines the maximum length of a sequence that the transformer can accept. All the sequences that are greater in length than max_length are truncated while shorter sequences are padded with zeros. Another token is also added which encodes unknown tokens as $[UNK]$.

In the first experiment, we used the roberta-base model with a basic configuration. The size of an extracted embedding was 768. In our second experiment, we used the concatenated outputs of the last four hidden layers. The size of an extracted embedding increased to 3072. In both experiments, the maximum training sequence length was set to 60, batch size to 16, the number of epochs to 10, and early stopping with the patience of 3 epochs. The training set contains utterances of less than 60 tokens. In addition, the loss was computed using a softmax layer with cross-entropy and 1/5 training steps were set as warm-up steps. While training, we used Adam optimization with a learning rate of 0.00002.

4.3. Acoustic

The MFCC and Mel-spectrogram features for 2D CNN implementation on the MELD audio data modality were extracted using the librosa library - a python package commonly used for music and audio analysis. For computational reasons, the audio duration to be loaded was set to 41 seconds. The maximum audio length for MELD data is also 41 seconds. The number of MFCCs returned was set to 30 and the Mel-spectrogram was set to 60. The batch size was set to 16. The 20 epochs training of 2D CNNs started with four convolutional layers, max pooling with a dropout rate of 0.2 and activation function as "relu". The final layer used activation softmax, loss was computed with cross-entropy, and the optimizer was set as Adam, with a learning rate of 0.001.

In the case of Wav2Vec 2.0, the feature encoder has a total receptive field of 400 samples or 25 ms of audio at a sample rate of 16,000 Hz. We used a maximum sequence length of MELD audio data as input to the network. The variable audio lengths used as input were passed through a temporal CNN network as explained in section 2.2.2. The batch size was set to 16, the maximum number of epochs to 100, and early stopping with the patience of 30 epochs. The size of each dimensional vector was 768, the same as in the RoBERTa, since both are transformer-based models. The extracted embeddings were then passed to 6 linear layers before passing to a softmax layer. The loss criterion was set as cross-entropy, 1/5 training steps were calculated as warm-up steps and the optimizer was Adam, with a learning rate of 0.00002.

4.4. Bimodal

In our bimodal experiment, we applied two settings immediately before modelling. In our first setting, RoBERTa's pooler output embeddings were merged with Wav2Vec 2.0 embeddings. In our second setting, RoBERTa's last four hidden layers' outputs were concatenated, and the embeddings were merged with Wav2Vec 2.0 embeddings. This was done by first casting layers to a tuple and concatenating over the last dimension. Following that, the mean of the concatenated vector over the token dimension was taken as the final output. For modelling, we introduced a simple concatenation function at the start of the modelling architecture, where embeddings from text and audio modalities are joined and fed as a single input for model training. In both settings, the batch size was set to 16, the maximum number of epochs to 100, and early stopping with the patience of 30 epochs. The size of each dimensional vector was 768, but in the case of the RoBERTa's last four hidden layers, the size was 3072. For the first setting, the concatenated dimension size was 1536, while for the second, it was 3840. The concatenated embeddings were then passed to 5 linear layers before passing to a softmax layer. The loss criterion was set as cross-entropy, 1/5 training steps were calculated as warm-up steps and the optimizer was Adam, with a learning rate of 0.00002.

4.5. Ensemble Learning

Following our unimodal and bimodal experimentation, ensemble learning methods were applied to further improve the model's performance. Max voting, averaging, weighted averaging and our proposed neural-network ensemble learning were the chosen ensemble learning methods. The 'max voting' method is generally used for classification problems. In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a 'vote'. The predictions which are nominated by the majority of the models are classified as the final prediction. Similarly, in the 'max voting' technique, multiple predictions are made for each data point in averaging. In this method, an average of predictions from all the models is calculated and used to create the final prediction. 'Weighted averaging' is an extension of the averaging method. All models are assigned different weights defining the importance and accuracy of each model. In our proposed ensemble learning technique, a neural network model was created and trained on the predictions of all six models with 250 epochs and a batch size of 10.

All steps were completed by first saving the predictions from each of the six models i.e. RoBERTa pooler output, RoBERTa last four hidden layers, Wav2Vec base, Wav2Vec large, 2D CNN (MFCC), 2D CNN (spectrogram). The saved predictions were then loaded as a single dataset to perform all the ensemble methods mentioned above.

5. Results and Discussion

In this section, we present the results of the experiments conducted on the MELD dataset. This work aims to present comparable results of text and audio modalities while also presenting the results of a simple bimodal fusion technique as well as ensemble learning. We aimed to highlight the effectiveness of ensemble learning in the task of bimodal sentiment analysis. At the time of writing, we did not find any work similar to our focus on using ensemble learning on diverse models of two modalities for enhancing the task of sentiment analysis classification.

5.1. Text Data Experiment

As mentioned in section 4.2, the RoBERTa model was used for text-only input data experiments. BERT and ALBERT were chosen as baseline models for this experiment. This eventually helped in comparing the model results and shortlisting the best-performing model for the fusion experiment. Table 1 shows the results of all models on the MELD dataset. The RoBERTa model produced state-of-the-art results for the MELD dataset when compared with the benchmark model results uploaded on the 'paperswithcode' website. The RoBERTa model with basic configuration achieved almost the same performance with a different configuration where outputs of its last four hidden layers were concatenated, as opposed to using the pooler output layer. The results shown in Figure 1 confirms that RoBERTa performs best in its basic configuration.

Table 1. Experimental results of MELD text dataset.

Model	Accuracy %	Loss	Precision	Recall	F1-score
BERT	70.0	0.84	0.70	0.70	0.70
ALBERT	69.1	0.85	0.69	0.69	0.69
RoBERTa	**71.0**	0.84	0.71	0.71	0.71
RoBERTa-last four hidden layers	70.5	0.84	0.70	0.71	0.70

5.2. Audio Data Experiment

As defined in section 4.3 section, 2D CNN was trained with two different features, i.e. MFCC and Mel-spectrogram. Further, pretrained Wav2Vec 2.0 base and large models were also experimented with. In this experiment, only audio data was used for the task of model training and predictions. Both models were compared against each other as opposed to using baseline models. An evaluation was conducted using metrics such as accuracy, loss, precision, recall, and f1-score. As illustrated in Table 2, the outperforming model was the one trained on the 2D CNN - spectrogram. 2D CNN's performance was close to that of the best-performing model. Figure 2 provides insights into predicted vs actual classes. When comparing the accuracy of audio data modality models with text data modality models, it is clear that the latter shows superior performance, as presumed.

Table 2. Experimental results of MELD audio dataset.

Model	Accuracy %	Loss	Precision	Recall	F1-score
2D CNN-MFCC	50.3	1.04	0.46	0.50	0.46
2D CNN-Mel-spectrogram	**51.3**	1.02	0.46	0.51	0.42
Wav2Vec 2.0-base	49.0	1.01	0.48	0.49	0.46
Wav2Vec 2.0-large	50.0	1.02	0.47	0.50	0.47

5.3. Bimodal Experiment

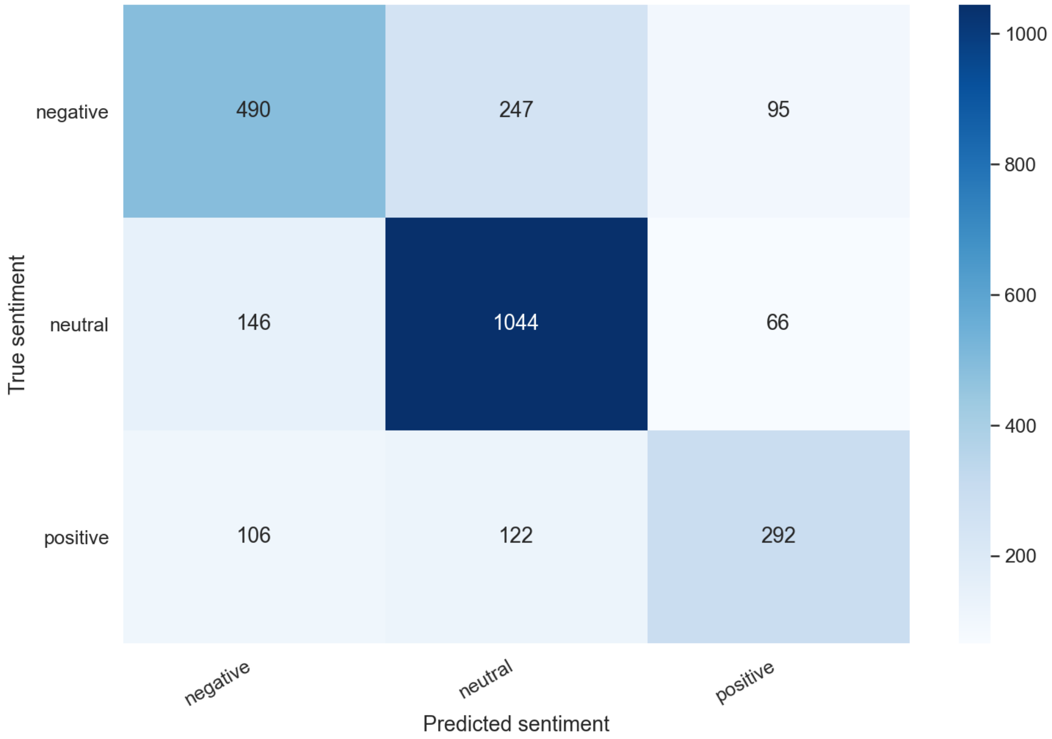
As defined in Section 4.4, RoBERTa and Wav2Vec 2.0 were the chosen models for this experiment in which both text and audio data were used for the task of model training and predictions. The outperforming model was the one trained on RoBERTa's concatenated last four hidden layers and Wav2Vec 2.0 outputs, as illustrated in Table 3. Figure 3 provides insights into predicted vs actual classes. When comparing the accuracy of this experiment with unimodal experiments, it is seen that although bimodal models outperform audio data modality-only models, they still do not show better performance than text data modality-only models.

Table 3. Experimentatal results of bimodal (text + audio) MELD dataset.

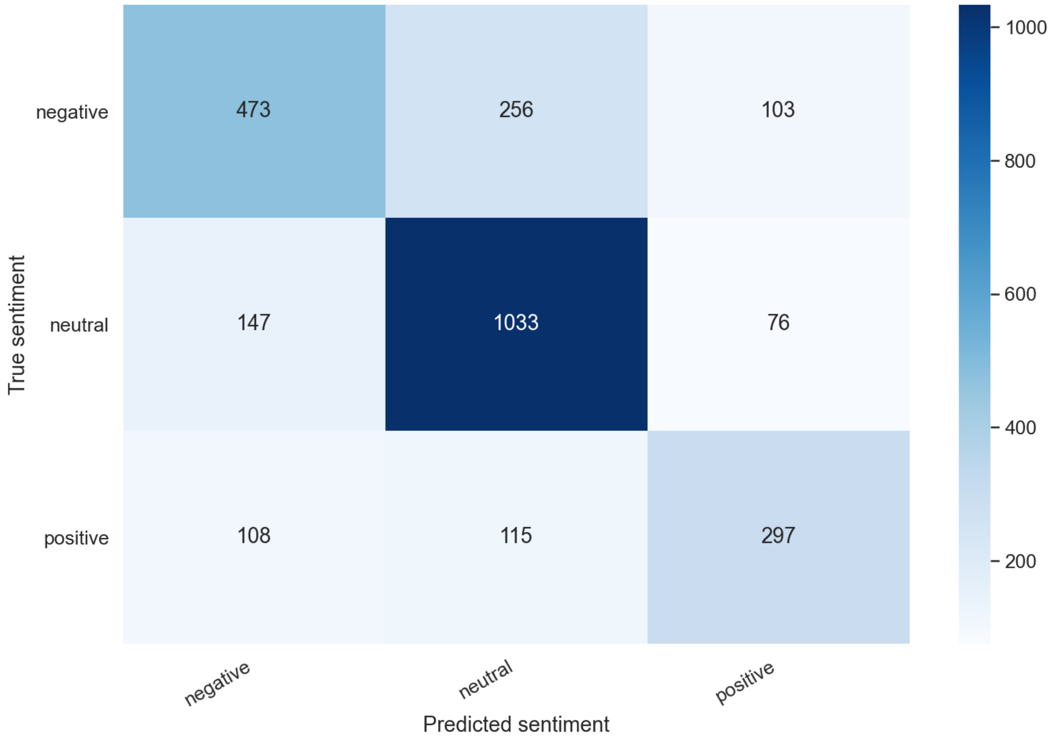
Model	Accuracy %	Loss	Precision	Recall	F1-score
RoBERTa + Wav2Vec 2.0	67.9	0.86	0.68	0.68	0.68
RoBERTa last four hidden layers + Wav2Vec 2.0	**68.4**	0.85	0.68	0.68	0.68

Figure 1. Test results of the four transformer-based models used for textual data modality sentiment classification highlighting predicted vs actual classes.

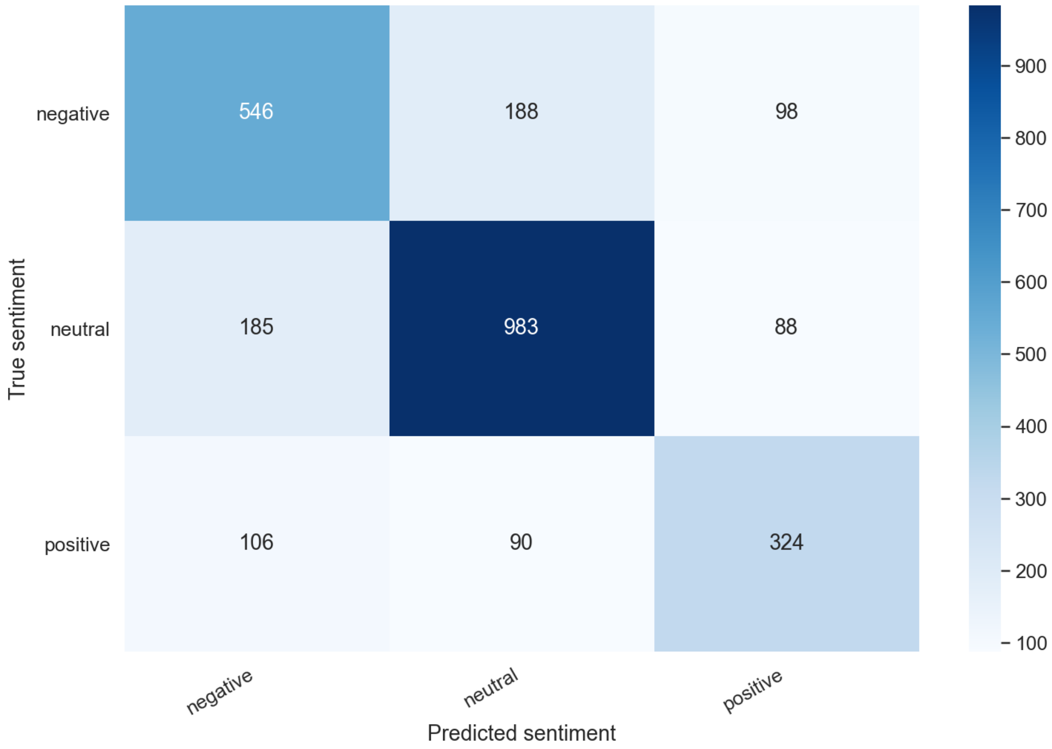
(a) BERT



(b) ALBERT



(c) RoBERTa



(d) RoBERTa-lfhl

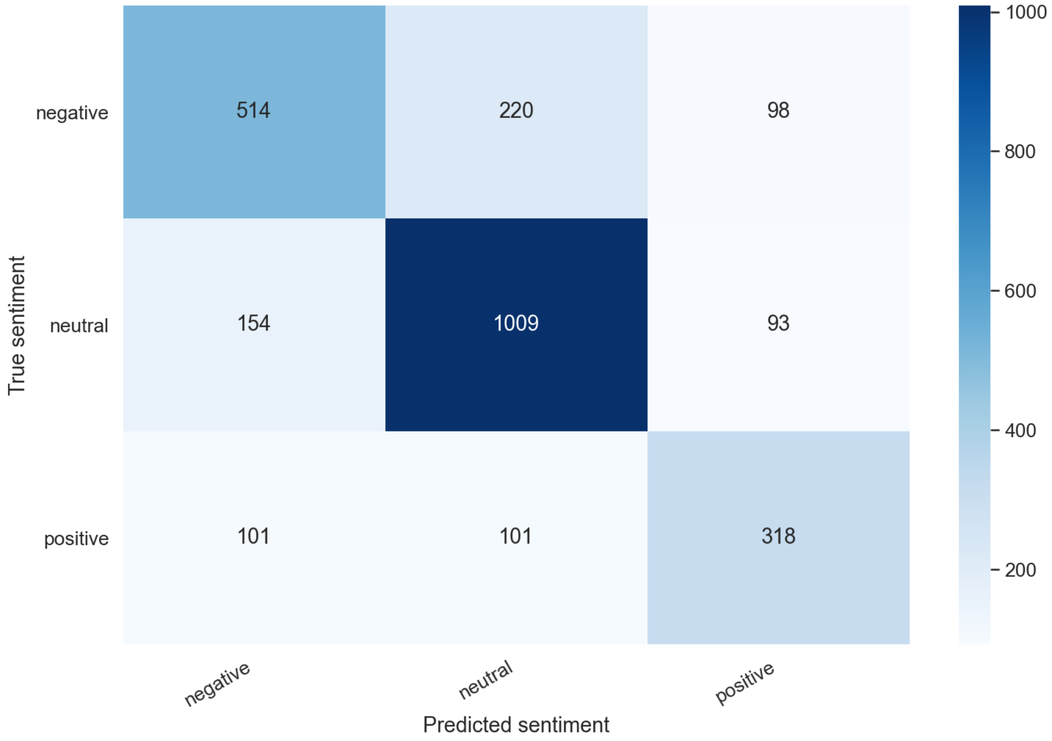
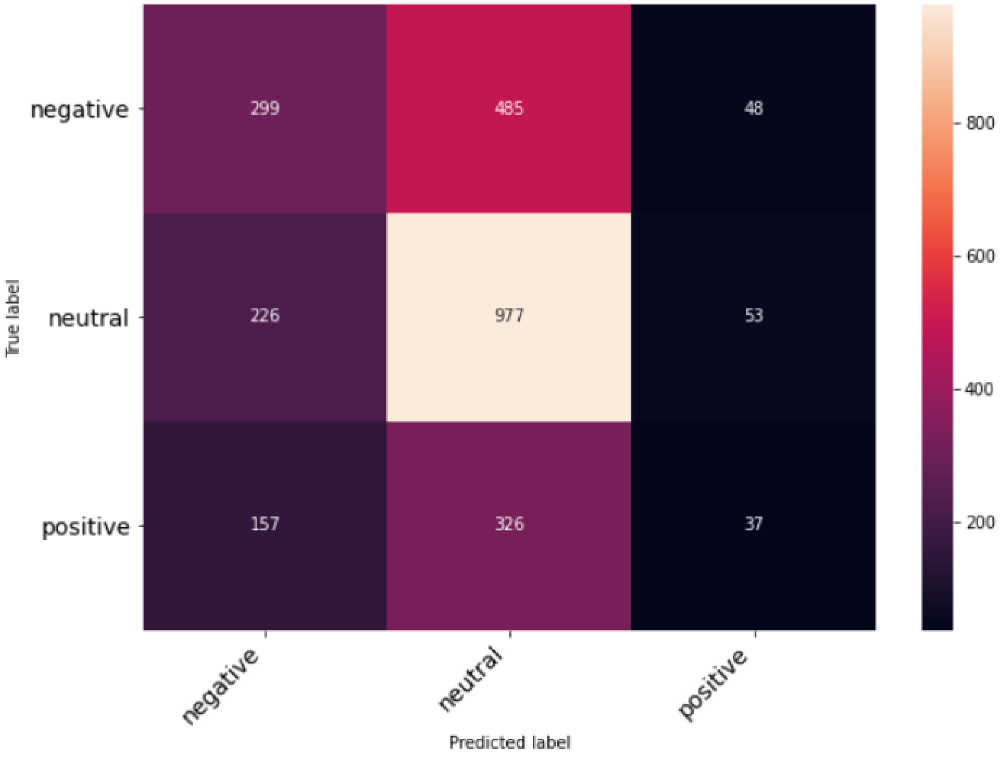
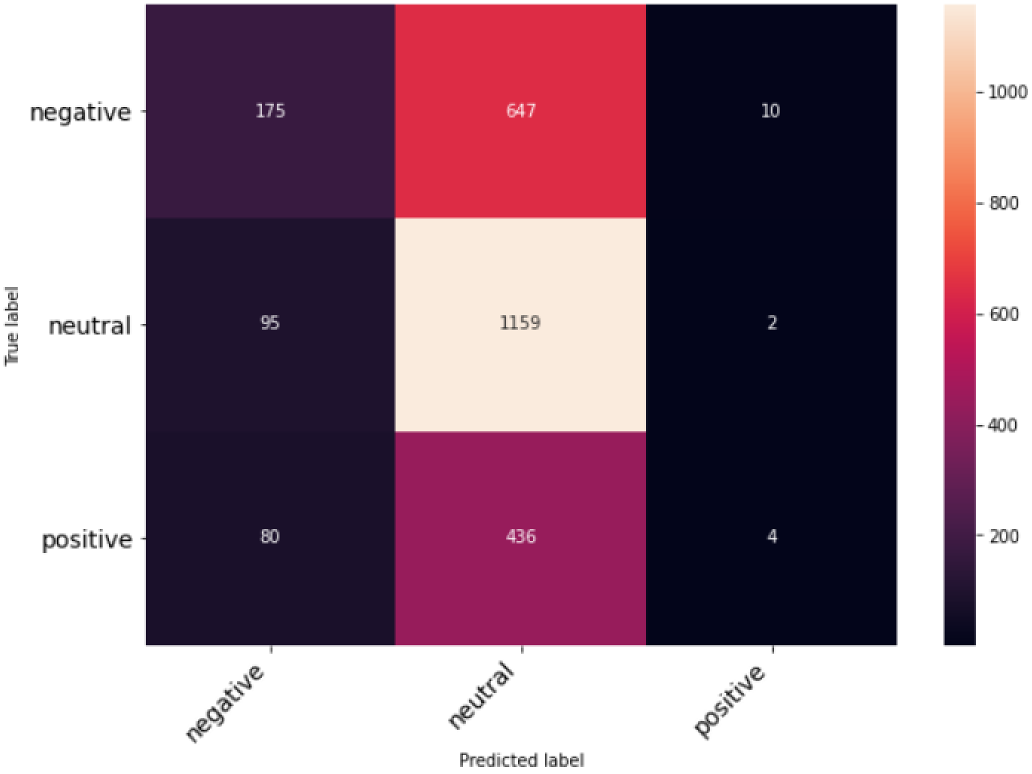


Figure 2. Test results of the four models used for audio data modality sentiment classification highlighting predicted vs actual classes.

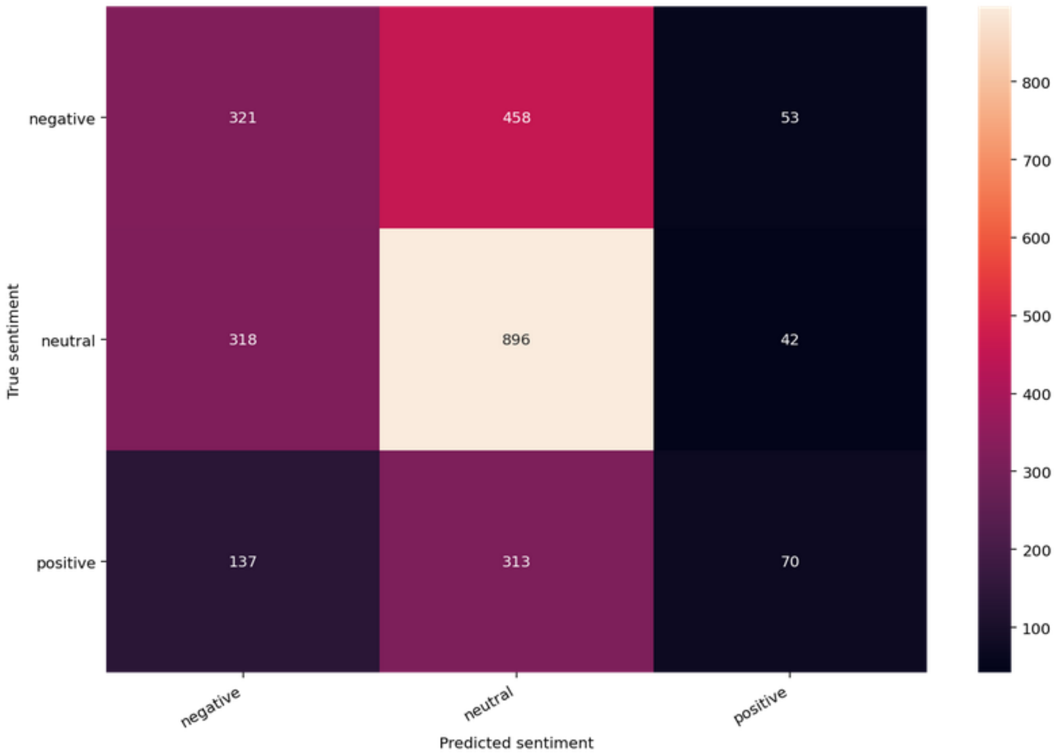
(a) 2D CNN mfcc



(b) 2D CNN spectrogram



(c) Wav2Vec 2.0 base



(d) Wav2Vec 2.0 large

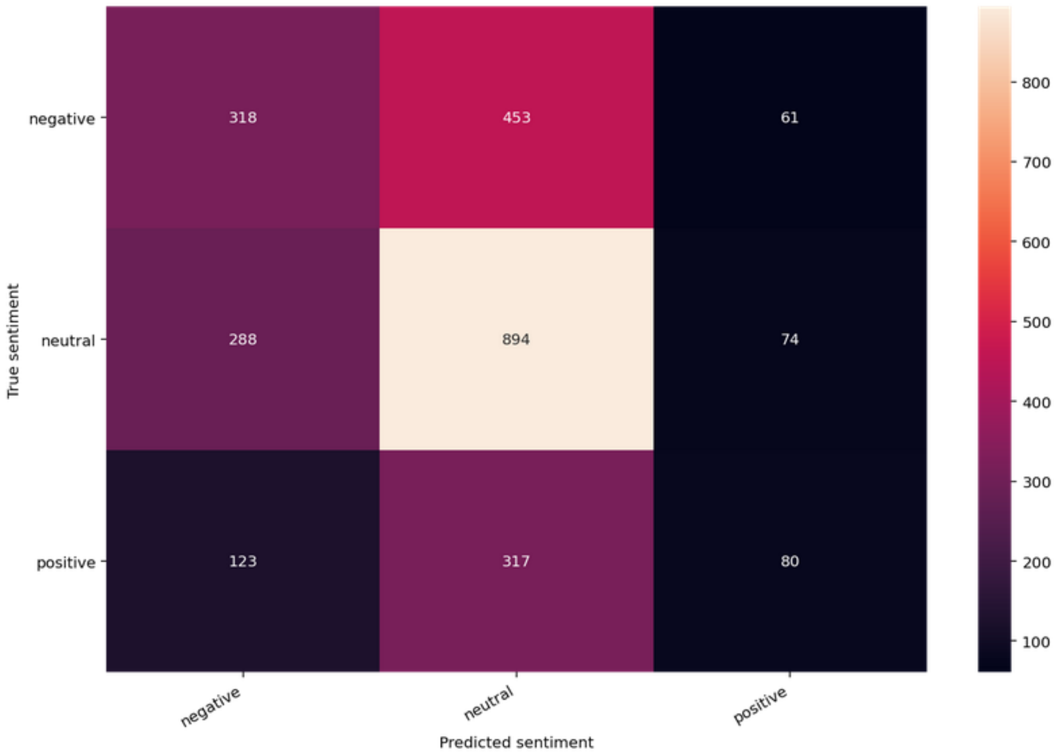
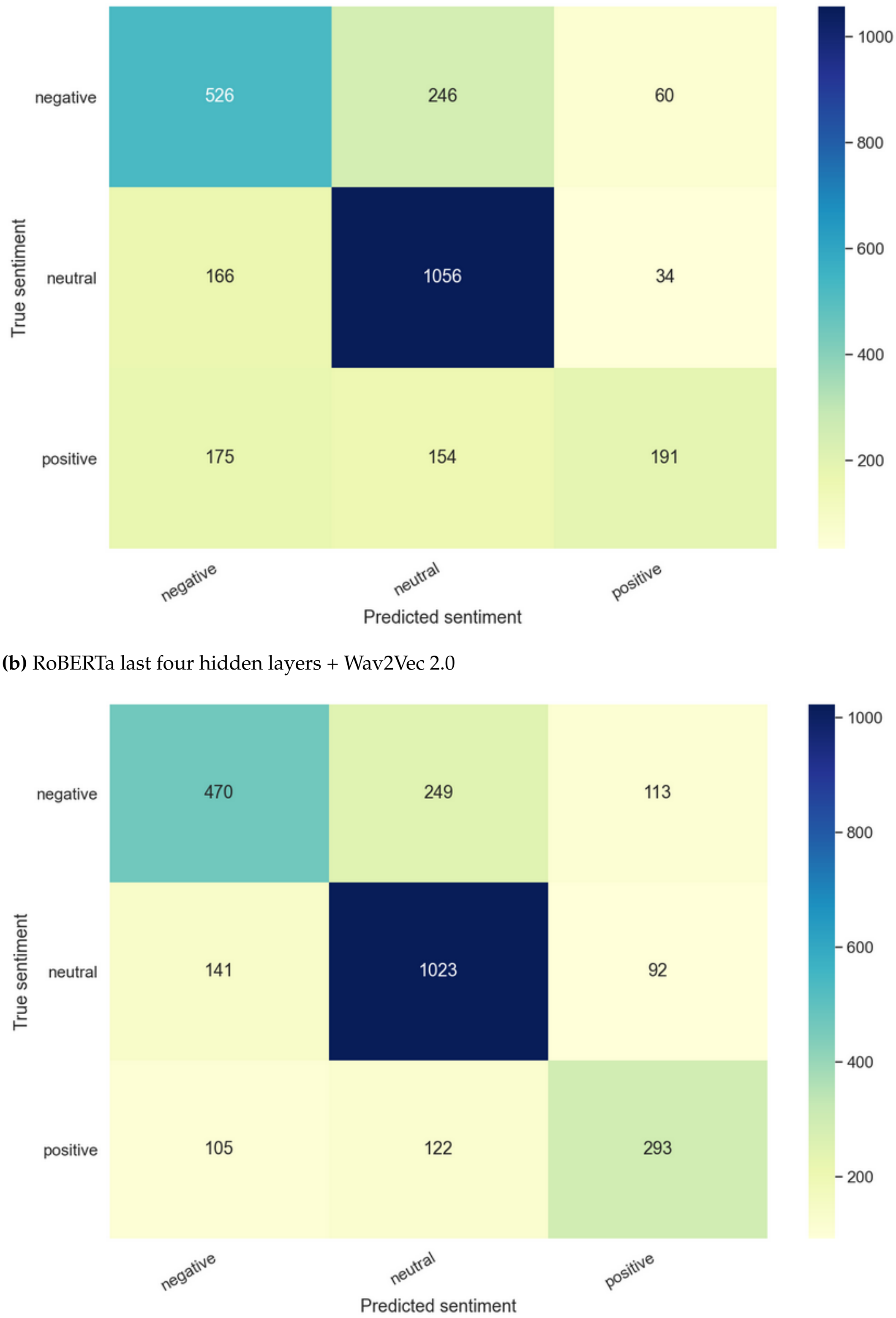


Figure 3. Test results of bimodal sentiment classification highlighting predicted vs actual classes.



(a) RoBERTa + Wav2Vec

(b) RoBERTa last four hidden layers + Wav2Vec 2.0

5.4. Ensemble Experiment

As explained in section 4.5, four different experiments were conducted to assess the impact of ensemble learning on the prediction results of the models previously experimented on in this study. As shown in Table 4, we can see there is no improvement with "Max Voting" and "Averaging" techniques. However, with "Weighted Averaging", there is an improvement of 0.3%, while with our proposed "Neural Network" technique, there is an improvement of 3.42% when compared with the text-only highest outperforming model.

Table 4. Experimental results of ensemble learning techniques.

Ensemble Learning Technique	**Accuracy** %
Max Voting	59.9
Averaging	68.6
Weighted Averaging	71.3
Neural Network based ensemble learning	**74.4**

Table 5. All experimental results put together.

Model	**Accuracy** %
BERT	70.0
ALBERT	69.1
RoBERTa	71.0
RoBERTa-last four hidden layers	70.5
2D CNN-MFCC	50.3
2D CNN-Mel-spectrogram	51.3
Wav2Vec 2.0-base	49.0
Wav2Vec 2.0-large	50.0
RoBERTa + Wav2Vec 2.0	67.9
RoBERTa last four hidden layers + Wav2Vec 2.0	68.4
Max Voting	59.9
Averaging	68.6
Weighted Averaging	71.3
Neural Network based ensemble learning	**74.4**

6. Conclusions and Future Work

In this paper, we presented multiple techniques for classifying sentiment using different types of data. As illustrated in Table 5, a set of experiments was conducted to compare the performance of different models including state-of-the-art models for sentiment classification using uni and bimodal datasets and variations of those models. We calculated accuracy, loss, precision, recall, and f1-score for evaluating the best-performing models in our experiments. According to the experimental results, RoBERTa achieved the highest average accuracy compared to other models in the text-only setting, whereas 2D CNN trained on Mel-spectrogram features achieved the highest in the audio-only setting. In our bimodal approach - based on the output of the model trained on RoBERTa's concatenated last four hidden layers along with Wav2vec 2.0 features, this version showed higher performance than the other approach adopted. Lastly, to overcome the shortcomings of models with lower accuracy, we proposed simple neural network-based ensemble learning. To evaluate our ensemble learning approach, we considered other ensemble learning techniques in our final set of experiments. Our proposed ensemble learning approach outperformed the second-best technique by 3.12%.

The main limitation of this work currently lies in the fact that it relies extensively on the quality as well as the number of distinctive models for generating a good-performing neural network-based ensemble learning model. An interesting avenue to explore in the future could involve relying on a few highly-performing models for this experiment.

In the future, we would like to use our method to reproduce the results of this study on other datasets and ensure our proposed method is consistent. Further, as progress is made in the field of deep learning techniques, our work can provide a basis to enhance bimodal sentiment analysis, particularly audio sentiment analysis, and explore the benefit of implementing ensemble learning on such bimodal datasets.

Author Contributions: Shariq Shah made substantial contributions to the acquisition, analysis, and interpretation of results. As the first author, he was also responsible for draft creation and revisions. Hossein Ghomeshi contributed to the conception or design of the work while other authors critically proofread, provided feedback, and approved the version to be published.

Funding: This paper has been produced as part of the broader UK Research and Innovation (UKRI) sponsored Knowledge Transfer Partnership project between FourNet and Birmingham City University. The research has received funding from UKRI under project reference number 512001.

Institutional Review Board Statement: Not Applicable

Informed Consent Statement: Not Applicable

Data Availability Statement: No new data were created. Also, the dataset used in the experiments can be downloaded from https://affective-meld.github.io/. The code for all the above experiments can be accessed from https://github.com/SShah30-hue/sentiment-analysis-ensemble.git.

Acknowledgments: The authors wish to thank David Bealing and Micheal Babalola for their assistance and advice. We also thank our anonymous reviewers for their comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Venkataramanan, K.; Rajamohan, H.R. Emotion recognition from speech. *arXiv preprint arXiv:1912.10458* **2019**.
2. Hendler, J.; Mulvehill, A.M. Social machines: the coming collision of artificial intelligence, social networking, and humanity, 2016.
3. Hey, T.; Trefethen, A. The data deluge: An e-science perspective. *Grid computing: Making the global infrastructure a reality* **2003**, *72*, 809–824.
4. Picard, R.W. Affective computing, 2000.
5. Goodfellow, I.; Bengio, Y.; Courville, A. Deep learning, 2016.
6. Deschamps-Berger, T.; Lamel, L.; Devillers, L. End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings. In Proceedings of the 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2021, pp. 1–8.
7. Poria, S.; Majumder, N.; Mihalcea, R.; Hovy, E. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access* **2019**, *7*, 100943–100953.
8. Siriwardhana, S.; Kaluarachchi, T.; Billinghurst, M.; Nanayakkara, S. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access* **2020**, *8*, 176274–176285.
9. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the 2014 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2014, pp. 960–964.
10. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
11. Sundarprasad, N. Speech emotion detection using machine learning techniques **2018**.
12. Nwe, T.L.; Foo, S.W.; De Silva, L.C. Detection of stress and emotion in speech using traditional and FFT based log energy features. In Proceedings of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint. IEEE, 2003, Vol. 3, pp. 1619–1623.
13. Wouts, J.V. Text-based classification of interviews for mental health–juxtaposing the state of the art. *arXiv preprint arXiv:2008.01543* **2020**.
14. Nagarajan, B.; Oruganti, V. Deep net features for complex emotion recognition. *arXiv preprint arXiv:1811.00003* **2018**.
15. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540* **2019**.
16. Sarkar, P.; Etemad, A. Self-supervised learning for ecg-based emotion recognition. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 3217–3221.
17. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017, pp. 2227–2231.

18. Yoon, S.; Byun, S.; Jung, K. Multimodal speech emotion recognition using audio and text. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 112–118.
19. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the Interspeech, 2017, pp. 1089–1093.
20. Sarma, M.; Ghahremani, P.; Povey, D.; Goel, N.K.; Sarma, K.K.; Dehak, N. Emotion Identification from Raw Speech Signals Using DNNs. In Proceedings of the Interspeech, 2018, pp. 3097–3101.
21. Xu, H.; Zhang, H.; Han, K.; Wang, Y.; Peng, Y.; Li, X. Learning alignment for multimodal emotion recognition from speech. *arXiv preprint arXiv:1909.05645* **2019**.
22. Boigne, J.; Liyanage, B.; Östrem, T. Recognizing more emotions with less data using self-supervised transfer learning. *arXiv preprint arXiv:2011.05585* **2020**.
23. Mikolov, T.; Yih, W.t.; Zweig, G. Linguistic regularities in continuous space word representations. In Proceedings of the Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2013, pp. 746–751.
24. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* **2013**.
25. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. NAACL-HLT, 2018.
26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
27. Lu, Z.; Cao, L.; Zhang, Y.; Chiu, C.C.; Fan, J. Speech sentiment analysis via pre-trained features from end-to-end asr models. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7149–7153.
28. Tits, N.; Haddad, K.E.; Dutoit, T. Asr-based features for emotion recognition: A transfer learning approach. *arXiv preprint arXiv:1805.09197* **2018**.
29. Heusser, V.; Freymuth, N.; Constantin, S.; Waibel, A. Bimodal speech emotion recognition using pre-trained language models. *arXiv preprint arXiv:1912.02610* **2019**.
30. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**.
31. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* **2018**.
32. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* **2019**.
33. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* **2020**, *33*, 12449–12460.
34. Baevski, A.; Schneider, S.; Auli, M. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453* **2019**.
35. Baevski, A.; Auli, M.; Mohamed, A. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912* **2019**.
36. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 5206–5210.
37. Heravi, E.J.; Aghdam, H.H.; Puig, D. Classification of Foods Using Spatial Pyramid Convolutional Neural Network. In Proceedings of the CCIA, 2016, pp. 163–168.
38. Huang, Z.; Dong, M.; Mao, Q.; Zhan, Y. Speech emotion recognition using CNN. In Proceedings of the Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 801–804.
39. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control* **2019**, *47*, 312–323.
40. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE transactions on multimedia* **2014**, *16*, 2203–2213.
41. Pang, B.; Lee, L.; et al. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval* **2008**, *2*, 1–135.
42. Sun, J.; Xu, W.; Yan, Y.; Wang, C.; Ren, Z.; Cong, P.; Wang, H.; Feng, J. Information fusion in automatic user satisfaction analysis in call center. In Proceedings of the 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). IEEE, 2016, Vol. 1, pp. 425–428.
43. Li, J.; Wang, X.; Lv, G.; Zeng, Z. GraphMFT: A Graph Attention based Multimodal Fusion Technique for Emotion Recognition in Conversation. *arXiv preprint arXiv:2208.00339* **2022**.
44. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the Interspeech 2014, 2014.
45. Li, P.; Song, Y.; McLoughlin, I.V.; Guo, W.; Dai, L.R. An attention pooling based representation learning method for speech emotion recognition **2018**.

doi:10.20944/preprints202303.0460.v1

46. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016, pp. 5200–5204.
47. Seng, J.K.P.; Ang, K.L.M. Multimodal emotion and sentiment modeling from unstructured Big data: Challenges, architecture, & techniques. *IEEE Access* **2019**, *7*, 90982–90998.
48. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.M.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the Proceedings of the 6th international conference on Multimodal interfaces, 2004, pp. 205–211.
49. Wöllmer, M.; Metallinou, A.; Eyben, F.; Schuller, B.; Narayanan, S. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In Proceedings of the Proc. INTERSPEECH 2010, Makuhari, Japan, 2010, pp. 2362–2365.
50. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* **2017**, *37*, 98–125.
51. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* **2017**.
52. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.P. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), 2017, pp. 873–883.
53. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Proceedings of the conference. Association for Computational Linguistics. Meeting. NIH Public Access, 2019, Vol. 2019, p. 6558.
54. Ho, N.H.; Yang, H.J.; Kim, S.H.; Lee, G. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access* **2020**, *8*, 61672–61686.
55. Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; Zhou, G. Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In Proceedings of the IJCAI, 2019, pp. 5415–5421.
56. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779* **2021**.
57. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2018, Vol. 32.
58. Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. Icon: Interactive conversational memory network for multimodal emotion detection. In Proceedings of the Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 2594–2604.
59. Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.P.; Zimmermann, R. Conversational memory network for emotion recognition in dyadic dialogue videos. In Proceedings of the Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting. NIH Public Access, 2018, Vol. 2018, p. 2122.
60. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. Dialoguernn: An attentive rnn for emotion detection in conversations. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 6818–6825.
61. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* **2018**.
62. Chen, S.Y.; Hsu, C.C.; Kuo, C.C.; Ku, L.W.; et al. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379* **2018**.
63. Shaw, A.; Vardhan, R.K.; Saxena, S. Emotion recognition and classification in speech using artificial neural networks. *International Journal of Computer Applications* **2016**, *145*, 5–9.
64. Tomas, G.S. Speech Emotion Recognition Using Convolutional Neural Networks. PhD thesis, PhD thesis, Institute of Language and Communication, Technical University of Berlin, 2019.
65. Hugging face: The AI community building the future.
66. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* **2019**.