big data and cognitive computing

MDPI

Article

An Ensemble-Learning-Based Technique for Bimodal Sentiment Analysis

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

School of Computing and Digital Technology, Birmingham City University, Birmingham B5 5JU, UK; edlira.vakaj@bcu.ac.uk (E.V.)
* Correspondence: shariq.shah@bcu.ac.uk

Abstract: Human communication is predominantly expressed through speech and writing, which are powerful mediums for conveying thoughts and opinions. Researchers have been studying the analysis of human sentiments for a long time, including the emerging area of bimodal sentiment analysis in natural language processing (NLP). Bimodal sentiment analysis has gained attention in various areas such as social opinion mining, healthcare, banking, and more. However, there is a limited amount of research on bimodal conversational sentiment analysis, which is challenging due to the complex nature of how humans express sentiment cues across different modalities. To address this gap in research, a comparison of multiple data modality models has been conducted on the widely used MELD dataset, which serves as a benchmark for sentiment analysis in the research community. The results show the effectiveness of combining acoustic and linguistic representations using a proposed neural-network-based ensemble learning technique over six transformer and deep-learning-based models, achieving state-of-the-art accuracy.

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

Citation: Shah, S.; Ghomeshi, H.; Vakaj, E.; Cooper, E.; Mohammad, R. An Ensemble-Learning-Based Technique for Bimodal Sentiment Analysis. *Big Data Cogn. Comput.* **2023**, *7*, 85. https://doi.org/10.3390/bdcc7020085

Academic Editor: Domenico Ursino

Received: 21 March 2023
Revised: 25 April 2023
Accepted: 26 April 2023
Published: 30 April 2023

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, and 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 raised 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 customers' thoughts, feelings, and behaviors, 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 essential aspect of many applications, mostly in contact centers (CCs), which are the first point of contact with organizations 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 organizations and the NLP community at large,

to analyze 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 generalized 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 [9–13]. 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 [14–21].

In contrast to previous work, this study shows the significance of pretrained 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 better performance in diverse classification tasks [22]. This is attributed to their ability to combine the predictions of multiple models, thereby reducing the impact of individual model errors and biases [23]. Ensemble techniques are also flexible in terms of training and updating classifiers, which allows them to adapt to changing data patterns and improve over time [24]. Numerous studies have demonstrated the superiority of ensemble techniques over individual classifiers in various domains. For example, in a study on the classification of gene expression data, an ensemble of support vector machines (SVMs) outperformed individual SVMs and other popular classification methods, achieving higher accuracy and stability across different datasets [25]. Another study, on land use/land cover classification, showed that a hybrid ensemble of decision trees and SVMs achieved higher accuracy than individual classifiers and other ensemble methods [26].

Furthermore, ensemble techniques have been applied successfully in various real-world applications, such as intrusion detection in network security [27], diagnosis of breast cancer [28], and prediction of stock prices [24]. In real-world applications, models often encounter multiple orientations and other language nuances that can be difficult for individual classifiers to capture accurately. One of the key advantages of ensemble techniques is their ability to handle noisy and ambiguous data, which helps in identifying complex patterns in large volumes of data. By combining the predictions of multiple classifiers, ensemble methods can reduce the impact of such noise and improve overall accuracy. Ensemble techniques also provide a flexible framework for incorporating new data and updating the model over time. As new data becomes available, ensemble methods can adapt and re-weight the base classifiers to better capture the changing sentiment trends and dynamics in the data [29,30]. This adaptability is particularly useful in sentiment analysis applications that involve analyzing social media streams, where the sentiment can change rapidly in response to events and trends. The aforementioned advantages demonstrate the flexibility and effectiveness of ensemble techniques in solving today's complex classification problems of bimodal sentiment analysis.

The study seeks to achieve the following:

- Compare multiple pretrained text model representations—BERT, ALBERT, and RoBERTa.
- Compare multiple pretrained audio model representations—2D CNN and Wav2Vec 2.0.

- Among the models that are tested, conduct an analysis of potential models to be incorporated in the bimodal approach and compare the performance of individual models, with a bimodal strategy.
- To further improve the accuracy of their analysis, our proposed ensemble learning technique will be employed to combine multiple models and minimize the impact of individual model errors and biases.

The study aims to enhance our understanding of how emotions are expressed in conversations and improve the reliability of sentiment analysis methods in real-world applications. The study argues that the proposed approach can better handle noisy and ambiguous data, adapt to changing sentiment trends and dynamics, and improve overall accuracy, making it a valuable contribution to the field of sentiment analysis. The MELD dataset was chosen as the primary data source for this study, as it was considered to be a suitable alternative to real-world conversations due to its similarity in structure and content. The rest of this paper is organized as follows: Section 2 presents an overview of related research; Section 3 details our methodology; Sections 4 and 5 outline 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 6 presents the conclusions and outlines possible future work.

2. 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

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 [31]. These sets of features are computationally efficient and provide useful information for detecting certain types of acoustic events [31,32]. However, they have limited discriminative power and do not capture higher-level features [33,34]. 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 [1,35]. Frequency-domain features include spectrograms, MFCCs, spectral centroid, spectral roll-off, spectral entropy, and chroma coefficients. LFPC can be classified as both time- and frequency-domain features [31]. Frequency-domain features capture fine-grained information about the frequency content of the speech signal, provide useful information for detecting certain types of acoustic events, and can capture phonetic and phonological information [31,32]. However, they are computationally expensive, sensitive to noise reverberation, and may require careful normalization and parameter tuning [33,34]. The selection of low-level features for speech analysis depends on the specific task and characteristics of the speech signal. Time-domain features may be useful in detecting speech in noisy environments, while frequency-domain features may be more informative for analyzing prosodic features of speech. Works that have used low-level features include [1,12,35–38].

For text-based features, traditional Word2Vec-based models such as continuous bag of words (CBOW), Skipgram, GloVe, and ELMo have been previously used [10,39–41]. Contrary to other studies , ref. [42] proposed a network-based pattern analysis approach to extract more salient insights from Reddit textual data. More recently, transformer encoder-only models have become a popular choice, due to their ability to capture meaningful features for text classification tasks [43].

2.1.2. Deep Features

The features extracted using pretrained 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, pretrained models have been used to extract speech and textual features for the task of sentiment analysis [35,43–46]. Such works suggest that deep features can be a better choice in terms of accuracy when compared to low-level features. Similarly, BERT, GPT, and ELMo models have been used to extract deep features for various text classification tasks, including sentiment analysis, topic modeling, and named entity recognition. For instance, in [47], pretrained BERT features were used for text classification in the medical domain, and in [48], pretrained GPT-2 features were used for sentiment analysis. In [49], pretrained WaveNet features were used for speech recognition in noisy environments, and in [50], pretrained transformer features were used for sentiment analysis of spoken language. Deep features have demonstrated superior performance in numerous domains in comparison to other types of features. Their popularity for feature extraction is on the rise due to their capacity to learn significant representations more effectively [51–55].

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 shown substantially better results according to the general language understanding evaluation (GLUE) benchmark for evaluating natural language understanding systems [43,56,57]. 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 the training time, as opposed to during the 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-source repository. The model version used has 110 M 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 [56]. However, the maximum number of tokens remains limited to 512 tokens, similar to the limit for BERT, which is 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 pretrained in a self-supervised setting, similar to BERT's masked language modeling [58,59]. 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 $(z_1, z_2, \ldots, z_T)$, which encode about 30 ms of 16 kHz audio every 10 ms. The context network $g : Z \mapsto C$ converts these low-frequency representations into a higher-level contextual representation $c_i = g(z_i \ldots z_{i-v})$ for a receptive field v [60,61]. The overall receptive field, after passing through both networks, is 210 ms for the base version and 810 ms 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 labeled training

data and enable its use for low-resource languages. The model consists of 35 M parameters and was pretrained on 960 h of unannotated speech data from the LibriSpeech benchmark audiobooks data [62]. The authors set the embedding size to 512 and the maximum audio waveform length to 9.5 s.

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 [63]. 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 [64,65]. 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 [1,21,66]. 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-spectrogram matrices.

2.3. Sentiment Analysis Architecture and Fusion Approaches

Over the years, sentiment analysis research has shifted from analyzing full documents or paragraphs to a finer level of detail—identifying sentiment towards particular phrases or words, audio and visual cues [7,67,68]. 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 [67]. 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 utilizes two modal input representations to judge the sentiment of each utterance. Current bimodal-based approaches are scarce and the few of those who have focused on this have mainly preferred to use textual and acoustic modalities for sentiment analysis tasks.

In [69], a method was used that statistically combined features with N-grams, 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 the model. The F1 score of their proposed method improved to 69.1%, against the baseline F1 score of 65.4%. In [35], a novel transfer learning method is proposed to be used when there is training data with as few as 125 examples per emotion class. Their method is comparable to that in our study, as they combine pretrained 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 pretrained 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 [18,46], an approach was adopted to combine utterance-level audio and textual embeddings before the softmax classification layer. In [44], a different approach was followed, which used pretrained 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. Table 1 provides a summary of bimodal sentiment analysis studies.

Table 1. Summary of works on bimodal sentiment analysis.

Work	Model	Dataset	Result	Drawback
[69]	SVMs, MaxEnt entropy, Bayesian classifiers	China Mobile CC corpus	It combines multiple classifiers to improve F1 score to 69.1%.	Large amount of training data is required.
[35]	BERT, Wav2Vec 2.0, RNNs	IEMOCAP	It achieved high unweighted accuracy of 73.9% with relatively few training examples.	It may not perform as well on larger-size data.
[21]	GloVe, CNN, TDNN, BiLSTM	IEMOCAP	It combines both audio and textual unimodels and significantly improves weighted accuracy to 70.4%.	High computational demands.
[46]	BERT, BiLSTM, XLNet	IEMOCAP	Their best bimodal achieves 73.5% and 71.4% accuracy by using pretrained models.	It may not generalize well to different datasets or languages.
[18]	ARE, TRE, MDRE, MDREA	IEMOCAP	It effectively achieves better accuracy than other SOTA methods, with accuracy ranging from 68.8% to 71.8%.	The small dataset used in the study restricts its wide applicability.
[44]	CNN, LSTM, RNN and logistic model tree (LMT)	IEMOCAP, SWBD-sentiment	The use of pretrained models improved weighted accuracy to 71.7% and 70.10% for IEMOCAP and SWBD-sentiment, respectively.	The model's complexity poses a challenge in applications where quick processing is important.
[13]	belabBERT, neural network	Manual data collection	Hybridization of belabBERT with a basic audio classification network pushed its accuracy of 75.68% to 77.0%.	The results are only demonstrated through a very small Dutch language dataset.

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 by external noise factors, researchers on multimodal sentiment analysis tasks are receiving widespread attention for their unique approaches to combining different modalities [70]. 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 [71–73] with varying fusion techniques [74–79]. Recently, the effectiveness of novel DL architectures, such as transformers [80] and graph convolution nets [15], as fusion methods have been explored and highlighted as computationally efficient. In [81], 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], pretrained 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. The authors emphasize that their proposed fusion mechanism is both efficient and more accurate than previous SOTA methods when dealing with high-dimensional SSL features, in terms of the

size of embedding and large sequence length when working with three modalities. Another proposed fusion method is GraphMFT, where graph neural networks (GNNs) are leveraged to integrate and complement the information from multimodal data [70]. The results indicate that this method achieved better performance than previous SOTA approaches. Other relevant multimodal approaches previously proposed include ConGCN [82], MMGCN [83], MFN [84], ICON [85], BC-LSTM [79], CMN [86], and DialogueRNN [87]. Table 2 provides a summary of bimodal sentiment analysis studies.

Table 2. Summary of works on multimodal sentiment analysis.

Work	Model	Dataset	Result	Drawback
[70]	BiLSTM, GNNs	IEMOCAP, MELD	An improvement of 2.73% on IEMOCAP and 1.99% on MELD recorded.	Computationally intensive and untested generalizability to other datasets.
[71]	DNNs, ELM	IEMOCAP	DNN-based approach significantly enhanced weighted accuracy, to 54.3%, against other approaches.	Absence of comparison with the latest methods, small dataset size for evaluation.
[72]	CNN	IEMOCAP	Results demonstrate an improved performance of 71.8% weighted accuracy.	Lack of comparison with other advanced methods, and the evaluation is conducted on a relatively small dataset.
[73]	CNN, LSTM	RECOLA	Achieves significantly better performance in comparison to traditional designed features.	Focuses on only two emotions, limiting its generalizability to other emotion categories.
[78]	GloVe, LSTM, neural network	CMU-MOSI	An improvement of 4.0% on CMU-MOSI highlighted.	Lack of interpretability, and the absence of a detailed comparative evaluation.
[79]	LSTM	MOSI, MOUD, IEMOCAP	A significant improvement of 5–10% over the SOTA, and it exhibits high robustness in terms of generalizability.	Complex model, challenging training and optimization, and uses a single evaluation metric only.
[80]	Transformers	CMU-MOSI, CMU-MOSEI, IEMOCAP	The model outperforms prior SOTA approaches in both word-aligned and unaligned scenarios.	Modest experiment gains over existing approaches, and more recent models may have surpassed its performance.
[15]	Graph neural networks	IEMOCAP, AVEC, MELD	SOTA performance reported against previous methods.	Computationally expensive, highly dependent on input quality, and lacks interpretability.
[81]	RNN, BERT	IEMOCAP, MELD, CMU-MOSEI	Achieves better performance than previous approaches.	It does not include ablation studies to assess the performance thoroughly.
[8]	RoBERTa, Wav2Vec, FABNET	CMU-MOSEI, CMU-MOSI, IEMOCAP, MELD	The models, through the effective fusion mechanism, outperformed prior multimodal approaches.	Computationally expensive and requires a significant amount of data for training.

3. Methodology

In this section, we focus on the public benchmark dataset, and the features used for comparison. Figure 1 highlights the pipeline used in the study, consisting of multiple steps followed to obtain the results of each experimental approach.

3.1. Dataset Selection

MELD is a multimodal and multi-party dataset for emotion recognition in conversations [88]. It is an extended version of the EmotionLines [89] 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 13,700 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. As shown in Table 3, the training, development, and test sets in our experiments consist of 9988 and 1108, and 2608 audio files and textual utterances, respectively. The textual data is a comma-separated values (CSV) file with two columns: text and sentiment label.

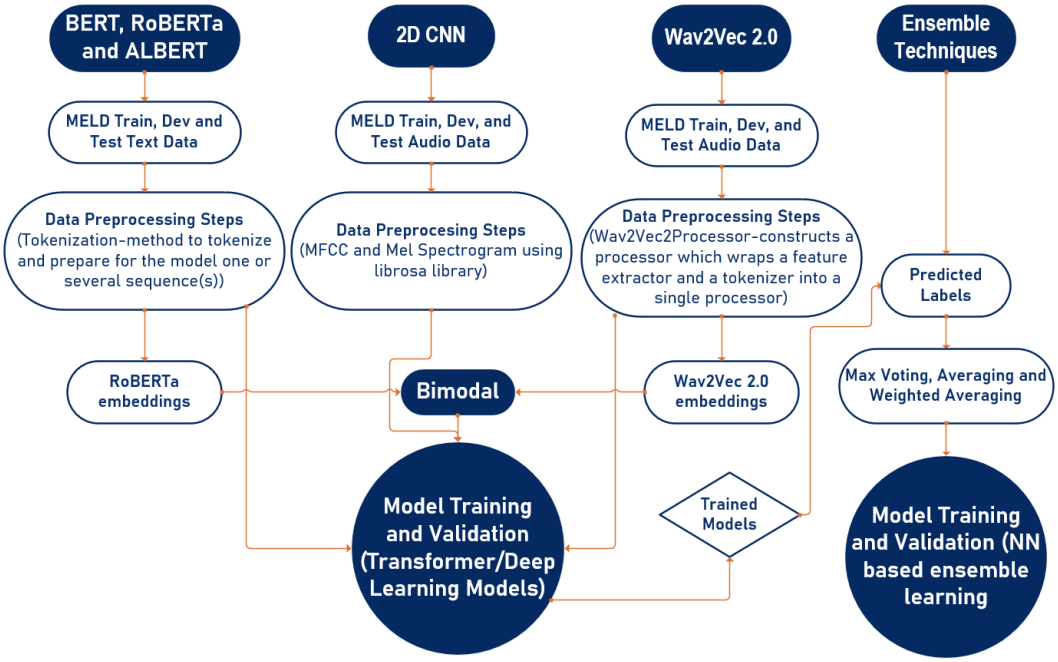


Figure 1. The architecture of the pipeline for an ensemble-learning-based technique for bimodal sentiment analysis.

Table 3. Description of the MELD dataset.

Data Type	Description
Modalities	Audio, text
Sentiment classes	Positive, neutral, negative
# training set	9988/8.72 h
# development set	1108/0.96 h
# testing set	2608/2.31 h
# positive sentiment	3087
# neutral sentiment	6434
# negative sentiment	4183
Elicitation	Play-acted

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 [43], 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 [56].

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 [90]. 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 [91]. 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 a 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 ms. 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 their implementation. We classify them into; textual, acoustic, bimodal, and ensemble learning, as shown in Figure 2.

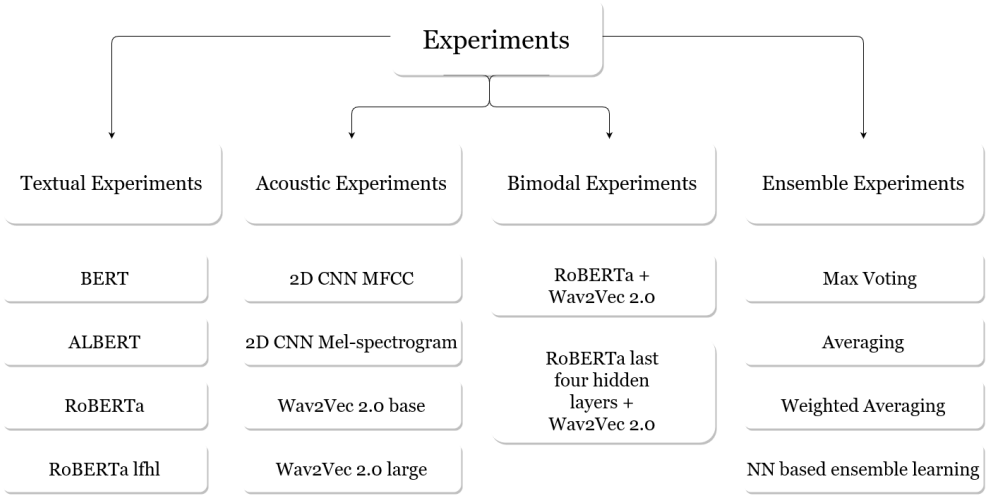


Figure 2. Flow chart portraying all experiments carried out.

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 [92]—a Python library providing detailed code explanations for building BERT-based models [93]. A sequential data processing and modeling 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 weighted 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 one-fifth of the 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 s. The maximum audio length for MELD data is also 41 s. 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-epoch 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 RoBERTa, since both are transformer-based models. The extracted embeddings were then passed to six linear layers before passing to a softmax layer. The loss criterion was set as cross-entropy, one-fifth of the 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 modeling. 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 achieved 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 modeling, we introduced a simple concatenation function at the start of the modeling 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 five linear layers before passing to a softmax layer. The loss criterion was set as cross-entropy, one-fifth of the 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 experiments for sentiment analysis, ensemble learning methods were employed to further enhance the model's performance. Four different ensemble learning methods were used, including max voting, averaging, weighted averaging, and our proposed neural-network-based ensemble learning method.

Max voting is a classification method that utilizes multiple models to make predictions for each data point. Each model's prediction is considered a 'vote', and the final prediction is determined by the majority of the models. A similar approach is taken in the averaging method, where predictions from all models are averaged to generate the final prediction.

Weighted averaging extends the averaging method by assigning different weights to each model, based on their importance and accuracy. In our proposed neural-network-based ensemble learning method, a neural network model was created and trained using the predictions of all six models for 250 epochs and a batch size of 10.

To perform these ensemble methods, the predictions from each of the six models were saved and then loaded as a single dataset. The ensemble methods were then applied to this dataset to generate the final prediction.

5. Results and Discussion

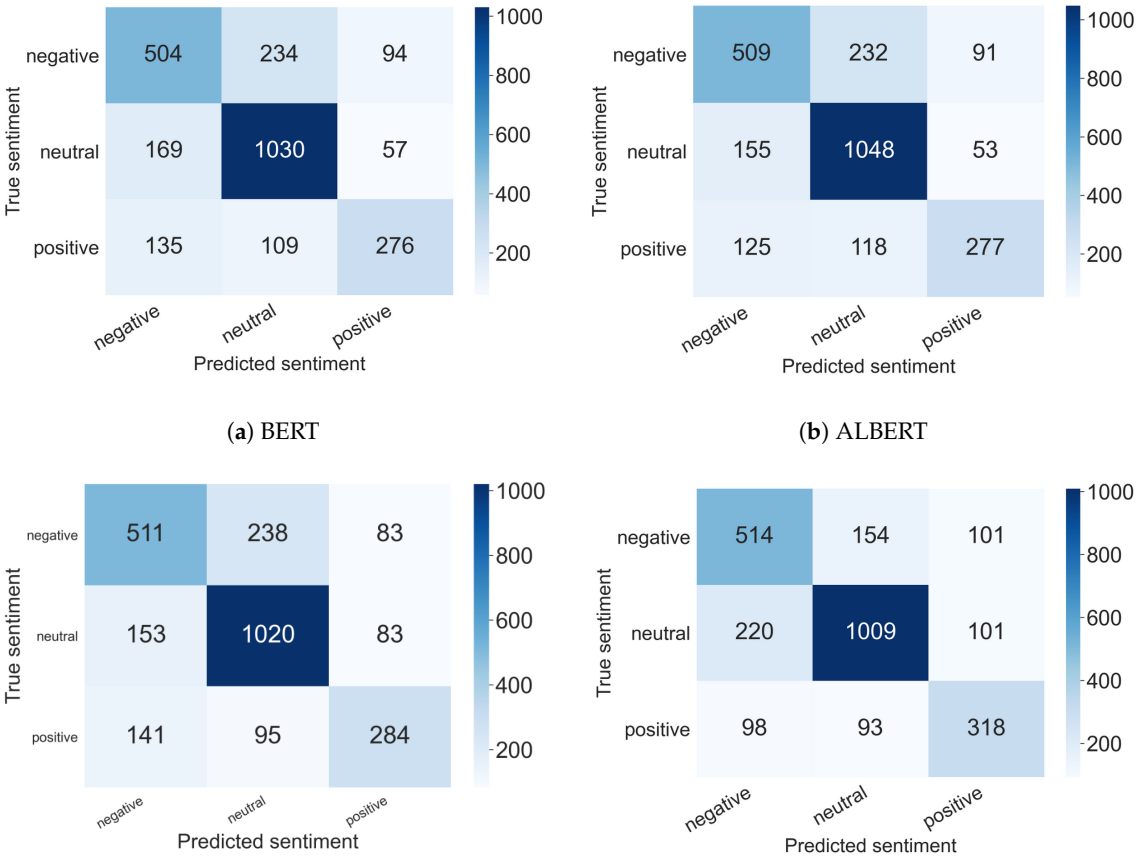
In this section, we present the results of the experiments conducted on the MELD dataset. This work aims to present comparable results of text and audio modalities while also presenting the results of a simple bimodal fusion technique as well as ensemble learning. We aimed to highlight the effectiveness 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 4 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 as 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 3 confirm that RoBERTa performs best in its basic configuration. RoBERTa outperforms other models mainly due to three reasons. Firstly, it was trained on a significantly larger corpus of text data, for an extended period, which enabled it to learn more intricate patterns and improve its ability to generalize. Secondly, it employs dynamic masking instead of static masking, which increases the maximum sequence length, resulting in more efficient learning. Furthermore, it does not utilize the next-sentence-prediction task, which again contributes to more efficient learning. Lastly, the model weights are randomly initialized instead of using pretrained weights for all tasks, making it more adaptable to new tasks and improving its overall performance [56].

Table 4. Experimental results of MELD text dataset.

Model	Weighted 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



(**a**) BERT (**b**) ALBERT

(**c**) RoBERTa (**d**) RoBERTa-lfhl

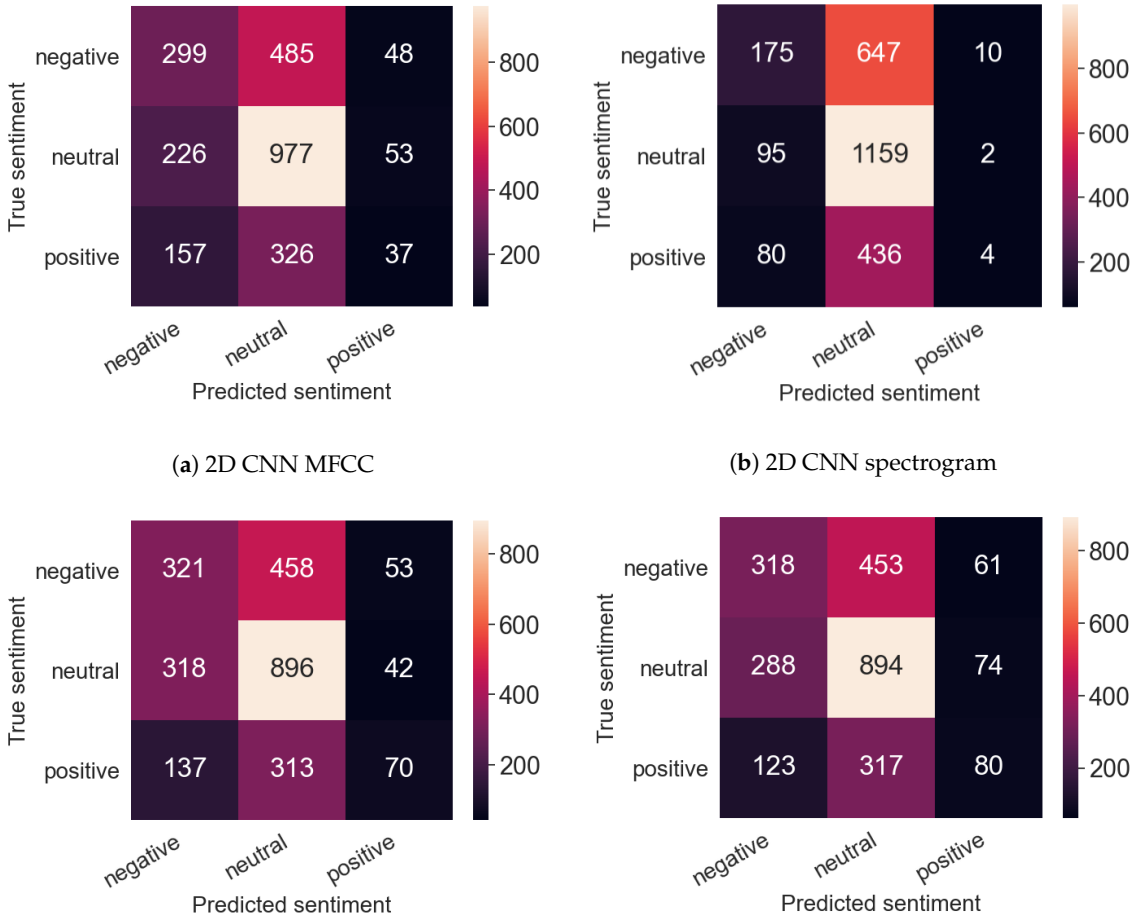
Figure 3. Test results of the four transformer-based models used for textual data modality sentiment classification, highlighting predicted vs. actual classes.

5.2. Audio Data Experiment

As defined in Section 4.3 , 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 weighted accuracy, loss, precision, recall, and F1 score. As illustrated in Table 5, the best-performing model was the one trained on the Mel-spectrogram. The performance of 2D CNN was close to that of the best-performing model. Figure 4 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 show superior performance, as presumed.

Table 5. Experimental results of MELD audio dataset.

Model	Weighted 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



(**a**) 2D CNN MFCC

(**b**) 2D CNN spectrogram

(**c**) Wav2Vec 2.0 base

(**d**) Wav2Vec 2.0 large

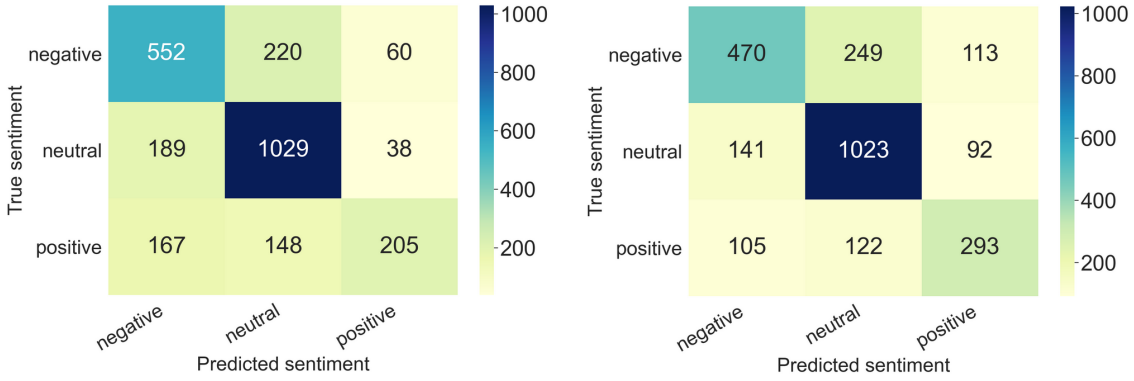
Figure 4. Test results of the four models used for audio data modality sentiment classification, highlighting predicted vs. actual classes.

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 best-performing model was the one trained on RoBERTa's concatenated last four hidden layers and Wav2Vec 2.0 outputs, as illustrated in Table 6. Figure 5 provides insights into predicted vs. actual classes. When comparing the weighted accuracy of this experiment with the unimodal experiments, it is seen that although the bimodal models outperform the audio data modality-only models, they still do not show better performance than the text data modality-only models.

Table 6. Experimental results of bimodal (text + audio) MELD dataset.

Model	Weighted 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



(**a**) RoBERTa + Wav2Vec 2.0 (**b**) RoBERTa last four hidden layers + Wav2Vec 2.0

Figure 5. Test results of bimodal sentiment classification, highlighting predicted vs. actual classes.

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 7, 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-performing model. This is because neural-network-based ensemble learning can learn complex relationships and patterns in the data, automatically select informative features, and adjust parameters to improve performance, while traditional ensemble techniques can only model linear relationships and require manual feature engineering and hyperparameter tuning [23].

Table 7. Experimental results of ensemble learning techniques.

Ensemble Learning Technique	Weighted Accuracy %
Max voting	59.9
Averaging	68.6
Weighted averaging	71.3
Neural-network-based ensemble learning	**74.4**

5.5. Implications

From the experimental results, it can be determined that ensemble methods can improve the accuracy and robustness of sentiment analysis models by reducing the impact of individual model biases and errors, and by leveraging the strengths of different models to capture diverse aspects of sentiment.

Our study shows that by combining the predictions of multiple models, ensemble methods can achieve more accurate and reliable sentiment analysis results in various applications, such as product review analysis, social media monitoring, and customer feedback analysis. Numerous studies in the domain have demonstrated that by combining the predictions of multiple classifiers, ensemble methods can reduce the risk of overfitting, improve accuracy and robustness, and provide a more balanced view of outputs [22,29].

By leveraging the strengths of multiple models, ensemble techniques can provide more authentic sentiment analysis results that can be applied to a wide range of applications, from CC conversation analysis to brand reputation management. It can also help companies to monitor and analyze CC customer agent performance, identify emerging trends and issues, and respond quickly and effectively to customer needs and preferences, eventually helping businesses to make more informed decisions and improve their overall performance and competitiveness [94,95]. It can also help in tracking the effectiveness of their communication

and engagement strategies with customers and help in personalizing them, which can result in building stronger and more positive relationships with customers [96].

Overall, the implications are broad and far-reaching. As sentiment analysis continues to grow in importance and popularity, ensemble techniques are likely to become an increasingly important tool for achieving more accurate and effective results.

6. Conclusions and Future Work

In this paper, we presented multiple techniques for classifying sentiment using different types of data. As illustrated in Table 8, 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 in those models. We calculated weighted 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 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%.

Table 8. Performance comparison of unimodal and bimodal models with proposed neural-network-based ensemble learning technique on the MELD dataset.

Model	Weighted 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**

Our proposed approach is well-suited for analyzing bimodal data, which encompasses sentiment expressed through both spoken language and nonverbal cues, a common feature of CC interactions. By combining the results of both text-based and audio-based sentiment analysis, ensemble learning bimodal sentiment analysis can achieve a more comprehensive and nuanced understanding of the sentiment expressed, leading to more reliable insights for decision-making purposes. It offers various potential applications in different domains, particularly in identifying the sentiment of customer conversations within the CC domain. Additionally, it can be used in social media analysis to monitor the sentiment of online conversations related to a particular topic or brand, helping companies to identify and respond to customer complaints or concerns.

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 be 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: S.S. made substantial contributions to the acquisition, analysis, and interpretation of results. As the first author, he was also responsible for draft creation and revisions. H.G. contributed to the conception or design of the work while E.V., E.C. and R.M. critically proofread, provided feedback, and approved the version to be published. All authors have read and agreed to the published version of the manuscript.

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. The dataset used in the experiments can be downloaded from https://affective-meld.github.io/ (accessed on 2 November 2021). The code for all of 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 Michael 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* **2019**, arXiv:1912.10458.
2. Hendler, J.; Mulvehill, A.M. *Social Machines: The Coming Collision of Artificial Intelligence, Social Networking, and Humanity*; Apress: New York, NY, USA, 2016.
3. Hey, T.; Trefethen, A. The data deluge: An e-science perspective. *Grid Comput. Mak. Glob. Infrastruct. Real.* **2003**, *72*, 809–824.
4. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000.
5. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 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), Nara, Japan, 28 September–1 October 2021; IEEE: Piscataway, NJ, USA, 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. [CrossRef]
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. [CrossRef]
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), Florence, Italy, 4–9 May 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 960–964.
10. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
11. Sundarprasad, N. Speech Emotion Detection Using Machine Learning Techniques. Master's Thesis, San Jose State University, San Jose, CA, USA, 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, Singapore, 15–18 December 2003; IEEE: Piscataway, NJ, USA, 2003; Volume 3, pp. 1619–1623.
13. Wouts, J.V. Text-based classification of interviews for mental health—Juxtaposing the state of the art. *arXiv* **2020**, arXiv:2008.01543.
14. Nagarajan, B.; Oruganti, V. Deep net features for complex emotion recognition. *arXiv* **2018**, arXiv:1811.00003.
15. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv* **2019**, arXiv:1908.11540.

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), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 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), New Orleans, LA, USA, 5–9 March 2017; IEEE: Piscataway, NJ, USA, 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), Athens, Greece, 18–21 December 2018; IEEE: Piscataway, NJ, USA, 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, Stockholm, Sweden, 20–24 August 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, Hyderabad, India, 2–6 September 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* **2019**, arXiv:1909.05645.
22. Wang, G.; Sun, J.; Ma, J.; Xu, K.; Gu, J. Sentiment classification: The contribution of ensemble learning. *Decis. Support Syst.* **2014**, *57*, 77–93. [CrossRef]
23. Kazmaier, J.; Van Vuuren, J.H. The power of ensemble learning in sentiment analysis. *Expert Syst. Appl.* **2022**, *187*, 115819. [CrossRef]
24. Chen, W.; Zhang, H.; Mehlawat, M.K.; Jia, L. Mean–variance portfolio optimization using machine learning-based stock price prediction. *Appl. Soft Comput.* **2021**, *100*, 106943. [CrossRef]
25. Nair, A.J.; Rasheed, R.; Maheeshma, K.; Aiswarya, L.; Kavitha, K. An ensemble-based feature selection and classification of gene expression using support vector machine, K-nearest neighbor, decision tree. In Proceedings of the 2019 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 17–19 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1618–1623.
26. Jozdani, S.E.; Johnson, B.A.; Chen, D. Comparing deep neural networks, ensemble classifiers, and support vector machine algorithms for object-based urban land use/land cover classification. *Remote Sens.* **2019**, *11*, 1713. [CrossRef]
27. Bhati, B.S.; Chugh, G.; Al-Turjman, F.; Bhati, N.S. An improved ensemble based intrusion detection technique using XGBoost. *Trans. Emerg. Telecommun. Technol.* **2021**, *32*, e4076. [CrossRef]
28. Hosni, M.; Abnane, I.; Idri, A.; de Gea, J.M.C.; Alemán, J.L.F. Reviewing ensemble classification methods in breast cancer. *Comput. Methods Programs Biomed.* **2019**, *177*, 89–112. [CrossRef] [PubMed]
29. Huang, J.; Xue, Y.; Hu, X.; Jin, H.; Lu, X.; Liu, Z. Sentiment analysis of Chinese online reviews using ensemble learning framework. *Clust. Comput.* **2019**, *22*, 3043–3058. [CrossRef]
30. Kandasamy, V.; Trojovskỳ, P.; Machot, F.A.; Kyamakya, K.; Bacanin, N.; Askar, S.; Abouhawwash, M. Sentimental analysis of covid-19 related messages in social networks by involving an n-gram stacked autoencoder integrated in an ensemble learning scheme. *Sensors* **2021**, *21*, 7582. [CrossRef]
31. Rabiner, L.; Juang, B.H. *Fundamentals of Speech Recognition*; Prentice-Hall Inc.: Hoboken, NJ, USA, 1993.
32. Purwins, H.; Li, B.; Virtanen, T.; Schlüter, J.; Chang, S.Y.; Sainath, T. Deep learning for audio signal processing. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 206–219. [CrossRef]
33. Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M.P.; Shyu, M.L.; Chen, S.C.; Iyengar, S.S. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
34. Torfi, A.; Shirvani, R.A.; Keneshloo, Y.; Tavaf, N.; Fox, E.A. Natural language processing advancements by deep learning: A survey. *arXiv* **2020**, arXiv:2003.01200.
35. Boigne, J.; Liyanage, B.; Östrem, T. Recognizing more emotions with less data using self-supervised transfer learning. *arXiv* **2020**, arXiv:2011.05585.
36. Sayedelahl, A.; Fewzee, P.; Kamel, M.S.; Karray, F. Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features. In Proceedings of the Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, 9–12 October 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 407–414.
37. Davis, N.; Suresh, K. Environmental sound classification using deep convolutional neural networks and data augmentation. In Proceedings of the 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), Thiruvananthapuram, India, 6–8 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 41–45.
38. Jain, M.; Narayan, S.; Balaji, P.; Bhowmick, A.; Muthu, R.K. Speech emotion recognition using support vector machine. *arXiv* **2020**, arXiv:2002.07590.
39. Mikolov, T.; Yih, W.T.; Zweig, G. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 746–751.
40. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
41. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. NAACL-HLT. *arXiv* **2018**, arXiv:1802.05365.

42. Cauteruccio, F.; Corradini, E.; Terracina, G.; Ursino, D.; Virgili, L. Extraction and analysis of text patterns from NSFW adult content in Reddit. *Data Knowl. Eng.* **2022**, *138*, 101979. [CrossRef]
43. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
44. 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), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 7149–7153.
45. Tits, N.; Haddad, K.E.; Dutoit, T. Asr-based features for emotion recognition: A transfer learning approach. *arXiv* **2018**, arXiv:1805.09197.
46. Heusser, V.; Freymuth, N.; Constantin, S.; Waibel, A. Bimodal speech emotion recognition using pre-trained language models. *arXiv* **2019**, arXiv:1912.02610.
47. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef]
48. Qian, T.; Xie, A.; Bruckmann, C. Sensitivity Analysis on Transferred Neural Architectures of BERT and GPT-2 for Financial Sentiment Analysis. *arXiv* **2022**, arXiv:2207.03037.
49. Rethage, D.; Pons, J.; Serra, X. A wavenet for speech denoising. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 5069–5073.
50. Yang, B.; Wu, L.; Zhu, J.; Shao, B.; Lin, X.; Liu, T.Y. Multimodal sentiment analysis with two-phase multi-task learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2015–2024. [CrossRef]
51. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
52. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
53. Wang, H.; Tian, K.; Wu, Z.; Wang, L. A short text classification method based on convolutional neural network and semantic extension. *Int. J. Comput. Intell. Syst.* **2021**, *14*, 367–375. [CrossRef]
54. Tang, D.; Qin, B.; Liu, T. Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1422–1432.
55. Maas, A.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 142–150.
56. 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* **2019**, arXiv:1907.11692.
57. 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* **2018**, arXiv:1804.07461.
58. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *arXiv* **2019**, arXiv:1904.05862.
59. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
60. Baevski, A.; Schneider, S.; Auli, M. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv* **2019**, arXiv:1910.05453.
61. Baevski, A.; Auli, M.; Mohamed, A. Effectiveness of self-supervised pre-training for speech recognition. *arXiv* **2019**, arXiv:1911.03912.
62. 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), South Brisbane, QD, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 5206–5210.
63. Heravi, E.J.; Aghdam, H.H.; Puig, D. Classification of Foods Using Spatial Pyramid Convolutional Neural Network. In Proceedings of the CCIA, Barcelona, Spain, 19–21 October 2016; pp. 163–168.
64. Huang, Z.; Dong, M.; Mao, Q.; Zhan, Y. Speech emotion recognition using CNN. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 801–804.
65. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control.* **2019**, *47*, 312–323.
66. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213. [CrossRef]
67. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
68. Ashimi, O.; Dridi, A.; Vakaj, E. Financial Sentiment Analysis on Twitter During Covid-19 Pandemic in the UK. In Proceedings of the International Conference of Advanced Computing and Informatics, Glasgow, UK, 22–26 August 2022; Springer: Berlin/Heidelberg, Germany, 2022.

69. 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), Hangzhou, China, 27–28 August 2016; IEEE: Piscataway, NJ, USA, 2016; Volume 1, pp. 425–428.
70. Li, J.; Wang, X.; Lv, G.; Zeng, Z. GraphMFT: A Graph Attention based Multimodal Fusion Technique for Emotion Recognition in Conversation. *arXiv* **2022**, arXiv:2208.00339.
71. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the Interspeech, 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
72. Li, P.; Song, Y.; McLoughlin, I.V.; Guo, W.; Dai, L.R. An attention pooling based representation learning method for speech emotion recognition. In Proceedings of the ISCA Conference, International Speech Communication Association, Hyderabad, India, 2–6 September 2018.
73. 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), Shanghai, China, 20–25 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 5200–5204.
74. 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.
75. 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 6th International Conference on Multimodal Interfaces, State College, PA, USA, 13–15 October 2004; pp. 205–211.
76. 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 INTERSPEECH 2010, Makuhari, Japan, 26–30 September 2010; pp. 2362–2365.
77. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
78. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
79. 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 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 873–883.
80. 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 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; NIH Public Access: Bethesda, MD, USA, 2019; Volume 2019; p. 6558.
81. 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. [CrossRef]
82. 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, Macao, China, 10–16 August 2019; pp 5415–5421.
83. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv* **2021**, arXiv:2107.06779.
84. 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 AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–3 February 2018; Volume 32.
85. Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. Icon: Interactive conversational memory network for multimodal emotion detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp 2594–2604.
86. 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 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 1, pp. 2122–2132.
87. 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 AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6818–6825.
88. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv* **2018**, arXiv:1810.02508.
89. Chen, S.Y.; Hsu, C.C.; Kuo, C.C.; Ku, L.W. Emotionlines: An emotion corpus of multi-party conversations. *arXiv* **2018**, arXiv:1802.08379.
90. Shaw, A.; Vardhan, R.K.; Saxena, S. Emotion recognition and classification in speech using artificial neural networks. *Int. J. Comput. Appl.* **2016**, *145*, 5–9. [CrossRef]
91. Tomas, G.S. Speech Emotion Recognition Using Convolutional Neural Networks. Ph.D. Thesis, Institute of Language and Communication, Technical University of Berlin, Berlin, Germany, 2019.

92. Hugging Face: The AI Community Building the Future. 2021. Available online: https://huggingface.co/ (accessed on 3 December 2021).
93. 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* **2019**, arXiv:1910.03771.
94. Alrehili, A.; Albalawi, K. Sentiment analysis of customer reviews using ensemble method. In Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS), Aljouf, Saudi Arabia, 3–4 April 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
95. Araque, O.; Corcuera-Platas, I.; Sánchez-Rada, J.F.; Iglesias, C.A. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Syst. Appl.* **2017**, *77*, 236–246. [CrossRef]
96. Ullah, M.A.; Munmun, K.; Tamanna, F.Z.; Chowdhury, M.S.A. Sentiment Analysis using Ensemble Technique on Textual and Emoticon Data. In Proceedings of the 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), Istanbul, Turkey, 23–25 November 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 255–259.