

Multimodal emotion recognition in natural scenarios

Kjartan Thomas Madsen
University of Copenhagen
zhj440
zhj440@alumni.ku.dk

Zuzanna Dubanowska
University of Copenhagen
vpz558
vpz558@alumni.ku.dk

Abstract

Emotion recognition finds many applications in affective computing and Human Computer Interaction. Many systems have been developed to classify emotions based on data coming from different modalities. Lately, multimodal emotion recognition systems, inspired by how humans perceive emotions, have been reporting promising results. These results, however, usually come from evaluating on datasets where noise is limited. In this paper, we propose a system for text, audio and multi-modal emotion classifiers. We explore different approaches to fusing modalities. A dataset consisting of excerpts from naturalistic conversations is used for testing and we evaluate system’s performance in predicting the archetypal emotion (anger, disgust, fear, joy, sadness and surprise) and sentiment (negative, positive, neutral). Results show that the proposed system fails to distinguish emotions or positive sentiment from speech. The text-based model performs the best across our proposed systems, yielding a weighted F1-score of 0.60 and 0.68 for emotion and sentiment respectively. Across the multi-modal models, intermittent fusion shows best recognition of emotion and sentiment, with F1-scores of 0.38 and 0.49 respectively.

Keywords: Multimodal emotion recognition, Speech Emotion Recognition, Multimodal EmotionLines Dataset, Deep Learning, LSTM

Division of labour

Both contributed equally to conducting the experiments and implementing the models. Zuzanna wrote sections 2, 3.1, 4, 6.1.1, 6.1.2, 6.2, & 7.1. Kjartan wrote sections 3.2, 5, 6.1.3, . Sections 1 & 7 were co-written.

1 Introduction

Emotion recognition is a challenging task in Artificial Intelligence, finding numerous applications to areas in human-computer interaction and affective computing. With the evolution of deep learning, new technologies enable researchers to tackle this problem with better and better accuracy, but we are far from reaching the point of machines understanding human emotions. Emotion

recognition could improve human-computer (human-robot) interactions in countless ways: endowing robots with emotional intelligence is essential to make interactions with people more organic in the future; in driver-aid systems, identifying the driver’s emotional state could prevent dangerous accidents from happening.

Emotions are manifested and perceived multimodally. They are expressed through facial expressions, body language, voice, the words we speak or even physiological means like heart rate or blood volume pressure. Humans use contextual information, prior experience, and multiple sensory inputs to infer other peoples’ emotions (Juckel G, 2018). There have been many attempts to textual emotion recognition (Baziotis et al., 2018; Huang et al., 2019), speech emotion recognition (Yeh et al., 2020) as well as multimodal approaches in the past (Spezialetti et al., 2020; Perez-Gaspar et al., 2016a), (Liang et al., 2020).

However, most studies evaluate their models on specially curated datasets where there is no noise, i.e. most variables are controlled: they are spoken by one speaker, or one sentence is uttered, etc. In this paper we want to evaluate one of the most well recognised approaches to Emotion Recognition against an ”in the wild”, noisy dataset, i.e one that was not designed with this or any experiment in mind and with many speakers. We want to explore whether a combination of modalities could result in improved emotion recognition as compared to single modal approaches, and how well state-of-the-art methodologies in ER can deal with noisy, naturalistic data.

In our approach we take two modalities: text and audio obtained from conversational scenes and build deep neural models for single- and multi-modal emotion recognition. We evaluate the performance of the proposed model against existing results on this dataset.

We want to answer the following questions: *Does the use of multiple modalities improve machine recognition of human emotion? Which approach to fusing modalities yields the best performance?* We hypothesise, that the combination of modalities will yield overall better performance than single-modal models.

In Section 2 we provide a survey of related work, in Section 3 we lay out the methodology used in our experiment, in Section 4 we present the dataset, Section 5 introduces the reader to the experimental setting, Section 6 presents the results. Lastly in Section 7 we discuss the

results and suggest where improvements could be made, and conclude our findings.

2 Related Work

2.1 Emotion recognition in text

Emotion recognition in text (ERT) is an important problem in Natural Language Processing (NLP), solutions of which benefit a wide variety of applications in different fields, from psychology through Human-Computer Interaction to data mining and advertising. The approaches vary in emotion analysed (detecting valence, arousal and power, or large sets of derived emotions), corpus type (stories, blog entries, news headlines, conversational) and approaches to modelling (rule-based approaches, machine learning approaches or hybrid approaches), see (Alswaidan and Menai, 2020) for a comprehensive review. We focus on ERT methodologies using Deep Learning (DL) as these are most closely related to our problem formulation.

Deep learning is a branch of machine learning where the model learns patterns by building hierarchies of concepts using internal representations (Goodfellow et al., 2016). DL based ERT is comprised of three main components: text preprocessing (e.g. tokenizing, trimming, treating unknown words), feature extraction (e.g. obtaining word embedding vectors) and classification.

Wang and colleagues (Wang et al., 2016) introduced a CNN network for a multi-emotion recognition task of 8 basic emotion (anger, anxiety, expect, hate, joy, love, sorrow, surprise) in posts from a Chinese microblogging service. Authors use a skip-gram language model to learn the distributed word representations. The model was evaluated against an SVM classifier baseline and achieved superior performance compared to the traditional machine learning method.

Baziotis et al. (Baziotis et al., 2018) proposed a bidirectional LSTM equipped with a multi-layer attention mechanism for multi-label emotion recognition from English Twitter posts. To represent the features, they obtained word2vec embeddings pretrained on 550 million domain-specific data points. Authors opted for a transfer learning approach and pre-trained the model on a bigger dataset due to limited training data. The solution was a part of the SemEval-2018 shared task "Multi-Label Emotion Classification" and was the winning approach (Mohammad et al., 2018).

For the same shared task, Meisheri et al. (Meisheri and Dey, 2018) proposed a parallel Bidirectional-LSTM architecture where the text input was represented as a matrix concatenated from various standard word embeddings (GloVe, emoji2vec and character-level). The parallel outputs of the LSTMs were concatenated and then fed into two fully connected networks. Their model achieved second place in the aforementioned competition.

Basile et al. (Basile et al., 2019) proposed a deep learning ensemble for emotion recognition in textual conversation. The ensemble was based on four submodels: an input submodel, an output submodel, a sentence-

encoder and a BERT model. Outputs of the four submodels were joined using a normalised SVM and ranked fourth at the SemEval-2019 shared task about emotion classification (Chatterjee et al., 2019).

For the same shared task at SemEval-2019, Ma et al. (Ma et al., 2019) proposed a solution based on bidirectional LSTM networks with emotion-oriented attention. Authors used pre-trained word embeddings to represent the text. The model's performance only outperformed the baseline on one out of three classified emotions, however authors showed that LSTMs are capable of extracting contextual information from text.

Huang et al. (Huang et al., 2019) investigated emotion recognition from conversation on EmotionLines dataset (Chen et al., 2018). Authors implemented a pre-trained BERT model due to limited training data and compared its performance to TF-IDF and BOW baselines. The model surpassed baselines' performance by a significant margin.

Deep et al. (Deep et al., 2019) proposed a Contextual Affect Detection framework based on GRU for emotion classification from conversations on EmotionLines dataset. Authors used contextual word embeddings and other hand-engineered features to represent the text. The model outperformed the state-of-the-art on the EmotionLines (Chen et al., 2018) dataset by 5%.

2.2 Emotion recognition from speech

Speech emotion recognition (SER) is a challenging component of affective computing. Machines endowed with the ability to differentiate emotions from voice would facilitate natural human-computer interaction in e.g. call centers, vehicle driving systems or medical applications. Many techniques have been utilised to classify emotion based on signal, differing in emotions analysed (positive or negative, or specific emotions), corpora (acted, naturalistic) and machine learning methods used, see (Khalil et al., 2019) for a comprehensive review. Recently, DL techniques have been making a mark in the field due to preference towards extraction of low-level features from raw data, ability to deal with un-labeled data and ability to detect complex patterns humans may not differentiate. We focus on deep learning based SER, since it is most closely related to our problem.

DL based SER comprises of feature extraction (e.g. extracting energy-based or prosody features) and classification (Koolagudi and Rao, 2012). In the field of speech processing, researchers have defined some standard features such as spectral features, prosody features or energy features that convey the most information and capture the most variance of the speech signal.

Tian et al. proposed an LSTM for emotion classification, with hierarchical fusion to combine lexical and low-level features. They used a wide range of features, including global prosodic features and acoustic features. The model was evaluated on IEMOCAP (Busso et al., 2008) and AVEC2012 (Schuller et al., 2012) datasets, obtaining better results using this type of fusion than other authors using feature and decision level fusions.

Wöllmer et al. used an LSTM-RNN for emotion recognition in a 3D space of valence, time and activation. They evaluated the model on SEMAINE (McKeown et al., 2012) dataset, observing best results for activation only, while performance of the valence was relatively low. Authors suggested adding additional modalities (such as linguistic features) to improve the performance.

Zhao et al. used 1-D and 2-D CNN-LSTMs for speech emotion recognition. The 1-D CNN was used to extract emotion-related features from speech, while the 2-D CNN was used to learn Mel-spectrogram. The proposed model was evaluated on BerlinDB and IEMOCAP datasets, obtaining better results using the 2-D network with Mel-spectrograms as features with recognition rates above 90% on the former dataset, and 50% and 90% on the latter dataset (speaker independent / speaker dependent cases).

Yeh et al. (Yeh et al., 2020) proposed a novel Dialogical Emotion Decoding (DED) inference algorithm for speech emotion recognition in conversations. Treating dialogues as sequences, the model consecutively decoded the emotion states of each utterance over time using a recognition engine. This decoder is trained by incorporating intra- and inter-speakers emotion influences within a conversation. The model was tested on IEMOCAP and MELD (Poria et al., 2018) datasets, achieving outstanding performance on the former and above baseline results on the latter.

2.3 Multimodal emotion recognition

The studies on speech emotion recognition and emotion recognition from text are done largely independently of each other. Human interaction, though, is inherently multimodal: we communicate with words but use intonation, gestures or facial expressions to emphasise the message we are trying to convey. As a consequence, more recent studies have been focusing on multimodal emotion recognition (MER) in hope of creating better performing ER systems.

Until recently, multi-modal ER largely encompassed combining textual or speech modalities with physiological data (Perez-Gaspar et al., 2016b; Val-Calvo et al., 2020; Barros et al., 2015; Perez-Gaspar et al., 2016a). Lately, the focus has been shifting towards modalities that human perception uses to process emotions: vision, audio, text, etc. Here, we discuss approaches to multimodal emotion recognition focusing on these 'perception' modalities and employing DL methodologies as these are most closely related to our problem formulation.

Yoon et al. proposed a dual encoder model that utilises textual and auditory modalities concurrently for multimodal speech emotion recognition. The architecture extracts information from speech data from signal to language level thus utilising the information in the data comprehensively. The proposed model outperformed previous state-of-the-art methods in assigning data to one of four emotion categories (i.e., angry, happy, sad and neutral) when evaluated against IEMOCAP dataset, with accuracies ranging from 68.8% to 71.8%.

Sun et al. (Sun et al., 2021) proposed a Multimodal Cross- and Self-Attention network for Multimodal Emotion Recognition using audio and text. The model employed cross and self-attention modules that combine the two modalities. The model was composed of an audio and text encoders, the attention modules and fully connected layers to classify the outputs. The audio features used were 40 MFCCs and textual features consisted of 300 dimensional GloVe embeddings. Authors evaluated model's performance on IEMOCAP and MELD and achieved state-of-the-art F1 results on both. Additionally, authors performed analysis of individual modalities' contribution to the predictive abilities of the model and found that removing either results in significant performance drops.

Ho et al. (Ho et al., 2020) proposed a multimodal approach to emotion recognition based on Multi-Level Multi-Head Fusion Attention mechanism and RNN from two modalities: audio and text. Authors used MFCCs to represent the audio signal and BERT embeddings to encode textual information. Two single-modal RNNs were trained in parallel and then the outputs were fused using the multi-head attention technique to predict emotional states. Authors evaluated their model on IEMOCAP, MELD and CMU-MOSEI datasets and concluded that combination of modalities yields better performance than using either of the modalities on its own. The proposed network performs favourably against state-of-the-art methods.

Padi et al. (Padi et al., 2022) leverage transfer learning for Multimodal Emotion Recognition. Authors proposed a neural network-based ER model that uses a late fusion of transfer-learned and fine-tuned models from speech (ResNet) and text modalities (BERT). Authors show, that a late fusion strategy leads to improvement in emotion recognition performance. Moreover, transfer learning proves beneficial as it helps mitigate the data scarcity, thereby improving the performance of emotion recognition models further. Padi et al. evaluated the proposed multimodal framework on IEMOCAP dataset. Experimental results indicated that both audio and text-based models improve the emotion recognition performance. The proposed multimodal solution achieved state-of-the-art results on the benchmark dataset.

3 Methodology

3.1 Long Short Term Memory Network with Attention

In our approach, we use an Long Short Term Memory network with attention for single- and multi-modal emotion recognition. LSTM (Hochreiter and Schmidhuber, 1997) is a special form of an RNN, which is especially suited to deal with sequential data. It contains recurrent units, or feedback connections, that allow the network to process entire sequences of data by retaining past information. LSTMs are recognised models for Emotion Recognition both from textual and audio data (see Related Work). Moreover, their ability to work with both textual and auditory modality enable us to re-use the

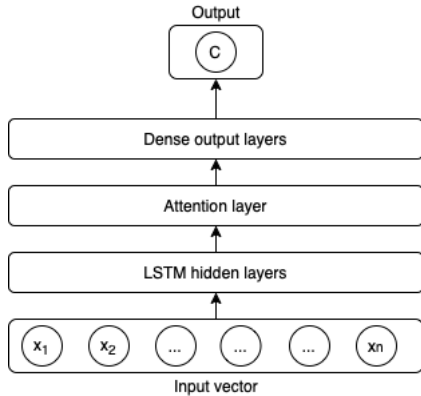


Figure 1: Model architecture.

same model for all experiments. Thanks to that, we can clearly evaluate how well emotion and sentiment can be predicted from each modality and from the combination of them, keeping the model itself a dependent variable.

The attention mechanism is inspired by human behaviour and imitates how our brain assigns more attention to some things than others, e.g. when reading an article one pays more attention to certain words. In machine learning, the attention mechanism works by assigning different weights to features. The greater the weight, the more the feature contributes to the predictions. Deep Learning architectures with attention are recognised in many NLP tasks and are widely implemented for emotion recognition (Alswaidan and Menai, 2020).


Our model obtained the attention weights of the features (words, audio features or both) through the attention layer, highlighting words or audio features that convey the most information about emotion or sentiment.

The model takes in 1D feature vectors. It consists of shared LSTM-layers (number of layers is a hyperparameter) with dropout to avoid over-fitting. The LSTM layers are followed by a 1D batch normalisation layer in order to increase stability of the network. The output of the normalisation layer is fed to the attention layer which is followed by two dense output layers with a ReLU activation in between them. The output of the second dense output layer is a binary vector of class labels.

3.2 Feature Selection

We focus on textual and auditory modalities for multi-modal emotion recognition, considering different features for each of the modalities. We consider features previously applied for and widely recognised in ER.


For the textual modality we use embeddings produced by a pre-trained BERT transformer encoder from the dialogue transcripts. BERT and variations of BERT (RoBERTa, ALBERT, etc.) has been shown to outperform other embeddings when paired with a LSTM-type model in classification tasks evaluated by F1-score (Wang et al., 2020). Furthermore, BERT is extremely versatile and a vast number of pre-trained models are readily available through the huggingface interface. Lee



Utterance: "Become a drama critic!"

Emotion: Joy **Sentiment:** Positive

Text	Audio	Visual
Ambiguous	Joyous tone	Smiling Face



Utterance: "Great, now he is waving back"

Emotion: Disgust **Sentiment:** Negative

Text	Audio	Visual
Positive/Joy	Flat tone	Frown

Figure 2: Example dialogue from the dataset, where each image with a label corresponds to one utterance. Source: (Poria et al., 2018)

& Lee (Lee and Lee, 2021) achieve state-of-the-art performance on the MELD dataset using RoBERTa embeddings in their model.

For the audio modality we use MFCC and MEL spectrogram features extracted from the audio recordings. These are the gold standard features in speech recognition, giving higher recognition accuracy in speech recognition systems as compared to other techniques (Gamit and Dhameliya, 2015).

4 Dataset

The dataset used in the experiments is the Multi-modal EmotionLines Dataset (MELD) (Poria et al., 2018). MELD contains approximately 13,000 clips from more than 1,400 dialogues from the American TV-series Friends (see Figure 2 for an example dialogue). Each utterance was labeled with an emotion label (anger, disgust, sadness, joy, neutral, surprise and fear) and sentiment label (negative, positive, neutral) and consists of textual, audio and visual modalities in the form of video clips and textual transcriptions. The exact statistics of the dataset can be found in Table 1.

5 Experimental Setting

We use the same LSTM architecture for textual, auditory and multi-modal emotion recognition. The model was implemented using PyTorch. We performed two multi-class classification experiments on the data: emotion recognition and sentiment recognition (see labels in Table 1). Because the dataset is not balanced (see Dataset section), we applied weights to each class through the loss function in all experiments. The weights were set to 1 - the probability of the class, so for the neutral class in the emotion recognition experiments the weight is $1 - 4710/9989 \approx 0.53$. By applying label smoothing (see (Szegedy et al., 2015)) through the loss function we got better result for some of the experiments, and we mention where this was applied. In all experiments the models were trained for 100-120 epochs, and we use the model weights of the epoch that achieved the highest

Emotion Data			
<i>Emotion / Data split</i>	<i>Train</i>	<i>Val</i>	<i>Test</i>
Overall	9989	1109	2610
Neutral	4710	470	1256
Anger	1109	153	345
Surprise	1205	150	281
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Sadness	683	111	208

Sentiment Data			
Overall	9989	1109	2610
Neutral	4710	470	1256
Positive	2334	233	521
Negative	2945	406	833

Table 1: Statistics on the number of utterances in the respective datasets.

F1-score on the validation set. The primary evaluation metric is weighted F1.

5.1 Audio-LSTM

First, the audio was extracted from the video files using ffmpeg (Tomar, 2006) and clipped (or padded) to 5s length, the signal was normalised by subtracting its mean and dividing by the standard deviation. Then, the signal was augmented using Additive White Gaussian Noise, as adding noise can reduce overfitting in the networks and improve their generalisation capabilities (Reed and Marks, 1999). MFCC and MEL spectrogram representations were created for every signal using Librosa (McFee et al., 2015). Because of the model architecture, the feature vectors had to be flattened to 1D. The feature vector representing each signal had length 104 (40 MFCCs and 64 MELs). Our model was trained with Adam optimiser with initial learning rate = 0.01 and weight decay of 0.001. The model had 3 recurrent layers and the hidden state size was 128, the dropout rate $p = 0.5$ to prevent overfitting. We trained the model for 100 epochs, with a batch size of 64. The same procedure was repeated for both emotion and sentiment recognition. Label smoothing was set to 0.01 for both experiments.

5.2 Text-LSTM

First, each utterance in the dataset was transformed into BERT embeddings using the pre-trained BERT model (Devlin et al., 2018). The embeddings were retrieved using the Huggingface library. The model for emotion detection consists of an input layer of the size of the embeddings (768), 2 recurrent layers with dimension 128, and has a dropout rate of 0.5. The model was then trained with Adam optimizer with initial learning rate = 0.0001 and a weight decay of 0.001. We apply label smoothing of 0.01 for both experiments. We trained the

model for 120 epochs and batch size = 64.

5.3 Bi-Modal-LSTM

We explored three approaches to multi-modal data fusion: early fusion, intermittent fusion, and late fusion. For early fusion, the word embeddings from the textual modality and the feature vectors from the auditory modality were concatenated to form one large feature vector of size $768 + 104 = 872$. The model is a LSTM with 3 recurrent layers of size 256. We trained the LSTM model end-to-end with label smoothing = 0.01 in both experiments. The learning rate was set to 0.0001 with weight decay of 0.001. The model was trained for 100 epochs with a batch size of 64.

The intermittent fusion approach consists of two LSTMs, one for text and one for audio, that are not connected to each other. The output of the two LSTMs are connected to the same neural layer, where their outputs are concatenated and fed through the same neural network that is attached to the text-LSTM (see section 3.1). The architecture allows for different sizes of the LSTMs, but the model performed best when they both had 3 recurrent layers with size 128. Label smoothing was not applied in this model. The learning rate was 0.0001 with weight decay 0.001.

For the late fusion approach, we trained two single-modal LSTM models concurrently and then concatenated the last layers and then used these features as input into a single neural layer for prediction (only the weights in this neural layer was trained, and the LSTMs remains the same). Label smoothing was set to 0.01. The learning rate was 10^{-5} with weight decay 0.001.

6 Results

The chosen metric for our experiments is the weighted F1-score and F1-score per class. All results are summarized in Table 2. A comparison of our results with other authors is included in Table 3.

6.1 Emotion Recognition

6.1.1 Speech Emotion Recognition

Our audio model reached an F1-score of 0.32 across the 7 classes in emotion recognition. The model overfits heavily to the 'neutral' class, failing to correctly categorise all except this label and 'anger' label to a very limited extent (0.03 F1-score). (Poria et al., 2018) achieve better performance using an LSTM model, with weighted F1 of 0.39, and emotion F1-scores ranging from 0.03 for fear, to 0.61 for neutral label. Similarly to us, their weighted F1 score mostly owes to the overpowering score on neutral label. Ho et al. (Ho et al., 2020) use a RNN-type model with attention, using MFCC features and narrowly outperforms Poria et al.'s approach.

6.1.2 Text Emotion Recognition

The text model performed the best overall across all of our proposed models, achieving a weighted F1-score of 0.60 across all classes. The text model also has the best

per class score of all our models. The model still overfits to the 'neutral' label, achieving an F1-score of 0.77 on this label, however it is capable of distinguishing between other emotions as well. The model achieves 0.57 on 'joy' label, closely followed by 'surprise' with an F1-score of 0.52, and 'anger' with a score of 0.41. It performs substantially worse in classifying the remainder of the labels: it achieves a 0.29 F1-score on classifying sadness, 0.22 on fear, and 0.13 on disgust. These results fall in line with (Poria et al., 2018). Authors achieve a weighted F1-score of 0.56 using an LSTM model. When it comes to specific emotions, the model achieves the lowest F1-scores on fear (0.07), disgust (0.22) and sadness (0.27). Similarly to our results, Poria et al.'s model's F1-score for anger, surprise and joy are around 0.5 (0.42, 0.48, 0.54 respectively) and 0.72 for the 'neutral' label. Lee and Lee (Lee and Lee, 2021) achieve the highest weighted F1-score, 0.67, on textual modality.

6.1.3 Multimodal Emotion Recognition

We have trained three different multimodal models. One employing early fusion denoted by *Multimodal (E)* in the table, intermittent fusion denoted by *Multimodal (I)* and late fusion denoted by *Multimodal (L)*. The multimodal approaches have similar performance to the speech only approach. Intermittent model performs the best achieving a score of 0.38, followed by 0.37 and 0.32 for the early, and late fusion. The models failed to recognise sadness, fear and disgust. All models achieve the best F1-score on neutral label. The early fusion model achieves the next best scores on anger, surprise followed by joy. In the case of the intermittent fusion model, it scores best on joy, followed by anger and then by surprise. Late fusion model has almost negligible recognition of surprise, anger and joy (0.08, 0.03 and 0.02 respectively).

Poria et al. (Poria et al., 2018) achieve a weighted F1-score 0.6 using an intermittent fusion approach. Their model outperforms ours on all emotions, however follows a similar pattern: fear, disgust, and sadness have the lowest F1-scores (our models fail to recognise any of the three) and surprise, anger and joy have higher F1-scores, yet still worse than the neutral label. Ho et al. (Ho et al., 2020) take on a late fusion approach similar to ours, but with significantly more complex architecture and they also train it end-to-end, in contrast to our late-fusion approach. Their approach slightly outperforms that of Poria et al. with a weighted F1-score of 0.01 higher overall.

6.2 Sentiment Recognition

6.2.1 Speech Sentiment Recognition

The model achieves a weighted F1-score of 0.42. It recognises the neutral label best, achieving a score of 0.42, followed by negative with 0.16. The model fails to classify positive label altogether.

Poria et al.'s (Poria et al., 2018) model outperforms our model by a significant margin. Their weighted F1-score 0.5, with 0.62 for the neutral label, 0.45 for negative

label and 0.25 for positive label. These results fall in line with ours.

6.2.2 Text Sentiment Recognition

The sentiment textual model performs best across our Sentiment Recognition models. The model achieves an overall F1-score of 0.68, recognising the neutral label best (0.72 score), followed by negative (0.62) and neutral (0.57).

Our model outperforms Poria et al.'s (Poria et al., 2018) model, their proposed method achieves a weighted F1 of 0.67. The model recognises the neutral label with F1 of 0.74, negative with 0.6 and positive with 0.54 F1-scores. These results indicate their model behaved similarly to ours in the Text Sentiment Recognition task. (Lee and Lee, 2021) achieve a weighted F1-score of 0.73, which to out knowledge is the highest recorded on this dataset.

6.2.3 Multimodal Sentiment Recognition

Amongst the multimodal models, the intermittent model performs best again, achieving an F-1 score of 0.49. The early fusion model achieves a score of 0.46 and the late fusion a score of 0.32. All classifiers recognise the neutral label best, the intermittent fusion model achieves a score of 0.47 on negative label and 0.26 on positive label. The early fusion achieves a score of 0.46 and 0.13 on negative and positive label respectively. The late fusion classifier fails to recognise the positive label and achieves a negligible F1-score on negative (0.04).

This study has not confirmed Poria et al.'s results (Poria et al., 2018). Their proposed LSTM achieves an overall score of 0.68 in text+audio sentiment recognition. Authors achieve the best F1-score on the positive label, followed by neutral and followed by negative. The positive labels recognition is better by a significant margin (F1-score difference of 0.14 to the next best predicted sentiment).

Emotion Recognition					
F1 / Modality	Speech	Text	Multimodal (E)	Multimodal (I)	Multimodal (L)
Overall	0.32	0.60	0.37	0.38	0.32
Neutral	0.65	0.77	0.64	0.63	0.57
Anger	0.03	0.41	0.23	0.20	0.03
Surprise	0.0	0.52	0.14	0.13	0.08
Disgust	0.0	0.13	0.0	0.0	0.0
Fear	0.0	0.22	0.0	0.0	0.0
Joy	0.0	0.57	0.13	0.25	0.02
Sadness	0.0	0.29	0.0	0.0	0.0
Sentiment Recognition					
Overall	0.42	0.68	0.46	0.49	0.32
Neutral	0.61	0.76	0.56	0.60	0.64
Positive	0.0	0.57	0.18	0.26	0.0
Negative	0.16	0.62	0.46	0.47	0.04

Table 2: Results of the emotion recognition and sentiment recognition experiments. The metric presented is F1-score. The (E), (I) and (L) in Multimodal Recognition stand for early, intermittent and late fusion respectively. We highlighted for each row the best performing multimodal approach.

Emotion Recognition			
<i>Author / Modality</i>	<i>Speech</i>	<i>Text</i>	<i>Multimodal</i>
Us	0.32	0.60	0.36
Lee&Lee	-	0.67	-
Ho et al.	0.45	0.59	0.61
Poria et al.	0.42	0.57	0.6
Sentiment Recognition			
Us	0.42	0.68	0.49
Lee&Lee	-	0.73	-
Poria et al.	0.5	0.67	0.68

Table 3: A comparison of our results with other emotion and sentiment recognition obtained on this dataset. The metric presented is F1 score. In the mulitmodal column we list our best result out of the three approaches

7 Discussion and Conclusions

Contrary to what we hypothesised, our study did not show that multi-modal emotion recognition systems perform better than single-modal ones. We employed methodology well recognised as state of the art in audio- and text- emotion recognition. We tested three approaches to modality fusion and found that intermittent fusion yields the best results among the three, but still below expectations. We are aware that our research might have had its limitations. The first one is the simple design of our model, which might have failed to capture the complexity of the data. The second one is the choice of audio features to represent the signal. The third one is the approach to fusion we have chosen to merge modalities during training. Additionally, the imbalance of the dataset is another possible source of error. We proceed with a discussion of the limitations that might have negatively influenced our results.

It is plausible, that a number of limitations arose because our methodology was too simple and failed to capture the complexity of the data. Other authors (Poria et al., 2018; Ho et al., 2020), evaluating models on the same dataset achieve better performance employing two modalities as compared to one. Poria et al.’s proposal however is a significantly more complex model, employing two-step hierarchical fusion to unite the modality data during training. Ho et al.’s proposed system is a multi-level multi-head fusion RNN, also more elaborate than our system. The poor result we obtained was not anticipated, nevertheless we still believe that multi-modal approaches to ER hold promise. Our findings, contrasted with other results suggest that emotion recognition ‘in the wild’ is too complex a phenomenon to model for simple architectures like one proposed in this paper.

Across all models, our results show that the poorest SER performance. We argue that the SER model’s performance was heavily affected by the imbalance of the dataset (see Table 1) as it only successfully retrieved

the neutral label constituting about 50% of the available data, which shows that our methods for counteracting this (label-smoothing, weighting) were not good enough. While these results fall somewhat in line with existing studies (see Table 3), Poria et al.’s (Poria et al., 2018) audio model, evaluated on the same dataset was capable of recognising the emotions to a greater extent than ours. A possible reason is that MFCCs and MEL spectral features do not accentuate differences between emotions uttered well enough for the model to recognise them. Poria et al. use raw discretized signals to produce feature vectors. Another possible reason is our model not being able to extract meaningful information from the auditory data, due to its simple architecture.

Moreover, as Poria et al. noted in their discussion, the audio might be negatively impacting the multimodal models’ performance. We hypothesise further, this could be the factor influencing intermittent and late fusion models’ performance negatively. From an architectural standpoint, these approaches are very dependent on how well the audio features reflect the emotion. In the early fusion, we argue that the audio features counteracted the performance of the classifier, by simply adding noise to the textual features instead of meaningful information. We suggest that future work expands the feature set used to represent the audio in order to capture more variance of the data.

Our results indicate that some emotions or sentiments are harder to recognise from text and audio than others. Similar findings hold for multimodal models: different fusions achieve different results across specific labels. The textual model recognised positive emotions (joy, surprise) with greater ease than negative emotions (anger, disgust, fear, sadness). While the neutral label still has the highest F1-score, attributed to the ‘neutral’ label greatly outnumbering the other labels, the model recognises other emotion and sentiment. The size of ‘surprise’ and ‘joy’ classes could explain this behaviour. The case is different for sentiment recognition: here, the textual model performs better on negative sentiment than on positive sentiment. A possible explanation is that, summed up, the negative emotions constitute a larger part of the dataset than positive emotions do so the model was exposed to more examples of the class. A possible solution to mitigate these issues in the future could be to use a different, balanced dataset, or remove the ‘neutral’ label and see the behaviour of the models then. In the case of multimodal models, the type of fusion seems to influence the emotion or sentiment being recognised best. We hypothesise, the choice of fusion of features accentuates different emotions/sentiment. For example, the late fusion classifier failed almost entirely to capture any but the neutral label. The early fusion classifier performed the best on anger, surprise and substantially better on negative than positive sentiment. The intermittent fusion classifier performed best overall, with the biggest contributions to weighted F1-scores being joy in the case of emotion and negative sentiment

(excl. neutral). Future work could explore why the type of fusion influences which emotion is recognised best. Future studies could also include more elaborate fusion techniques, such as these employed by Poria et al. and Ho et al. (Poria et al., 2018; Ho et al., 2020).

7.1 Conclusion

In this paper, we introduced an LSTM-based system for text-, audio- and multimodal emotion recognition. We evaluated our system against MELD, a naturalistic conversational dataset. The dataset has numerous sources of noise, sometimes eliminated in other data used to evaluate ER systems, like multiple speakers, different sentences uttered and many emotions showcased. The results of the study indicate that emotion recognition from data where a lot of noise is present is a complex task, and simple methodologies, like the one proposed in this paper might not be capable of modelling it. We failed to show that multi-modal emotion recognition systems outperform single-modal systems, however, these findings might not be representative of the real state of things. We argue that the discrepancy of our results and existing studies stems from our methodology being too simple, the auditory features not capturing enough variance of the data to represent the different emotions and the naïve approach to modality fusion. Our observations have several implications for future research. It indicated that for successful speech emotion recognition one needs to employ more sophisticated modelling techniques. It also suggests that naïve approaches to fusing modalities fail to represent the data in this complex case. Finally, it points to dataset imbalance as a large source of additional noise resulting in overfitting to certain labels. Future studies should target emotion recognition with these issues in mind.

References

- [Alswaidan and Menai2020] Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.
- [Barros et al.2015] Pablo Barros, Cornelius Weber, and Stefan Wermter. 2015. Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 582–587.
- [Basile et al.2019] Angelo Basile, Marc Franco-Salvador, Neha Pawar, Sanja Štajner, Mara Chineza-Ríos, and Yassine Benajiba. 2019. Symantoresearch at semeval-2019 task 3: combined neural models for emotion classification in human-chatbot conversations. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 330–334.
- [Baziotis et al.2018] Christos Baziotis, Athanasios Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 245–255, New Orleans, Louisiana, June. Association for Computational Linguistics.
- [Busso et al.2008] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, December.
- [Chatterjee et al.2019] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.
- [Chen et al.2018] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao, Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations.
- [Deep et al.2019] Kumar Shikhar Deep, Asif Ekbal, and Pushpak Bhattacharyya. 2019. A deep neural framework for contextual affect detection. In *International Conference on Neural Information Processing*, pages 398–409. Springer.
- [Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- [Gamit and Dhameliya2015] Mayur R Gamit and Kinnal Dhameliya. 2015. Isolated words recognition using mfcc, lpc and neural network. *International journal of Research in Engineering and technology*, 4(6):146–149.
- [Goodfellow et al.2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.
- [Ho et al.2020] Ngoc-Huynh Ho, Hyung-Jeong Yang, Soo-Hyung Kim, and Guesang Lee. 2020. Multi-modal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access*, 8:61672–61686.
- [Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- [Huang et al.2019] Yen-Hao Huang, Ssu-Rui Lee, Mau-Yun Ma, Yi-Hsin Chen, Ya-Wen Yu, and Yi-Shin Chen. 2019. Emotionx-idea: Emotion bert—an affective model for conversation. *arXiv preprint arXiv:1908.06264*.
- [Juckel G2018] Welpinghus A Brüne M. Juckel G, Heinisch C. 2018. Understanding another person’s emotions-an interdisciplinary research approach. *Front Psychiatry*, 9.
- [Khalil et al.2019] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345.

- [Koolagudi and Rao2012] Shashidhar G. Koolagudi and K. Sreenivasa Rao. 2012. Emotion recognition from speech: A review. 15(2).
- [Lee and Lee2021] Joosung Lee and Woojin Lee. 2021. Compn: Context modeling with speaker’s pre-trained memory tracking for emotion recognition in conversation.
- [Liang et al.2020] Jingjun Liang, Ruichen Li, and Qin Jin, 2020. *Semi-Supervised Multi-Modal Emotion Recognition with Cross-Modal Distribution Matching*. Association for Computing Machinery, New York, NY, USA.
- [Ma et al.2019] Luyao Ma, Long Zhang, Wei Ye, and Wenhui Hu. 2019. PKUSE at SemEval-2019 task 3: Emotion detection with emotion-oriented neural attention network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 287–291, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.
- [McFee et al.2015] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.
- [McKeown et al.2012] Gary McKeown, Michel F. Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3:5–17.
- [Meisheri and Dey2018] Hardik Meisheri and Lipika Dey. 2018. TCS research at SemEval-2018 task 1: Learning robust representations using multi-attention architecture. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 291–299, New Orleans, Louisiana, June. Association for Computational Linguistics.
- [Mohammad et al.2018] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June. Association for Computational Linguistics.
- [Padi et al.2022] Sarala Padi, Seyed Omid Sadjadi, Dinesh Manocha, and Ram D. Sriram. 2022. Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models.
- [Perez-Gaspar et al.2016a] Luis-Alberto Perez-Gaspar, Santiago-Omar Caballero-Morales, and Felipe Trujillo-Romero. 2016a. Multimodal emotion recognition with evolutionary computation for human-robot interaction. *Expert Systems with Applications*, 66:42–61.
- [Perez-Gaspar et al.2016b] Luis-Alberto Perez-Gaspar, Santiago-Omar Caballero-Morales, and Felipe Trujillo-Romero. 2016b. Multimodal emotion recognition with evolutionary computation for human-robot interaction. *Expert Systems with Applications*, 66:42–61.
- [Poria et al.2018] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations.
- [Reed and Marks1999] Russell D. Reed and Robert J. Marks. 1999. *Neural Smthing: Supervised Learning in Feedforward Artificial Neural Networks*. MIT Press.
- [Schuller et al.2012] Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2012. Avec 2012: The continuous audio/visual emotion challenge - an introduction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI ’12*, page 361–362, New York, NY, USA. Association for Computing Machinery.
- [Spezialetti et al.2020] Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. 2020. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7.
- [Sun et al.2021] Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. 2021. Multimodal cross- and self-attention network for speech emotion recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4275–4279.
- [Szegedy et al.2015] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision.
- [Tomar2006] Suramya Tomar. 2006. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10.
- [Val-Calvo et al.2020] Mikel Val-Calvo, José Ramón Álvarez Sánchez, José Manuel Ferrández-Vicente, and Eduardo Fernández. 2020. Affective robot storytelling human-robot interaction: Exploratory real-time emotion estimation analysis using facial expressions and physiological signals. *IEEE Access*, 8:134051–134066.
- [Wang et al.2016] Yaqi Wang, Shi Feng, Daling Wang, Ge Yu, and Yifei Zhang. 2016. Multi-label chinese microblog emotion classification via convolutional neural network. In *Asia-Pacific Web Conference*, pages 567–580. Springer.
- [Wang et al.2020] Congcong Wang, Paul Nulty, and David Lillis. 2020. A comparative study on word embeddings in deep learning for text classification. pages 37–46, 12.
- [Yeh et al.2020] Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee. 2020. A dialogical emotion decoder for speech emotion recognition in spoken dialog. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6479–6483.