Automatic Speech Recognition Thesis

Artificial Intelligence

Karan Chand s1033357
in collaboration with Mel Phan s1082649

Testing Robustness of Multilingual Emotion Recognition Models on Film Data

---

July 7, 2023

Abstract

Emotion recognition models are continually updated, though evaluation of them in diverse domains is regularly (insufficiently) unexplored. This paper aims to evaluate state-of-the-art (SOTA) multilingual emotion recognition models on their robustness in a film domain. Languages spoken include: original English, dubbed Italian and dubbed Spanish. Evaluation is based on accuracy and properties of confusion matrices in emotion and sentiment classification. The sentiment analysis model was not very robust to positive sentiments in any language, as the TPR neared chance level (∼50%). The pretrained emotion recognition model achieved above chance performance for all languages, though there seems so be a bias in the model and/or data. The pretrained models are thought to have insufficient features to make a correct classification, as speech features are excluded. A combined text and acoustic model is constructed for all languages and achieved substantial increases in performance for both the Italian and Spanish utterances (max accuracy of 52% and 57%, respectively). Classification on English utterances kept performance on the same level as the acoustic model. SOTA performance is reached on the EMO-film emotion recognition task using this combined model.

Contents

1 Introduction **3**

2 Related Work **4**
2.1 Textual Emotion Recognition 4
2.2 Combined Model 4
2.3 Previous Performance 4

3 Methodology **6**
3.1 Development 6
3.2 HuggingFace 6
3.3 Pretrained Sentiment Analysis Model 6
3.4 Pretrained Emotion Recognition Model 7
3.5 Audio-to-Emotion Model 7
3.6 Audio-to-Text 8
3.7 Model Emotions $\neq$ Data Emotions 8

4 Experiments **9**
4.1 EMO-film dataset 9
4.2 Combined Model Training Settings 9
4.3 Robustness to Whisper 9

5 Results **10**

6 Discussion **13**
6.1 Robustness of Pretrained Sentiment Analysis Model 13
6.2 Robustness of Pretrained Emotion Recognition Model 13
6.3 Robustness of Newly Trained Emotion Recognition Model 14
6.4 Comparison to Scotti et al. 14

7 Conclusion **16**

Chapter 1

Introduction

There is constant development in the field of Natural Language Processing (NLP). Models such as automatic speech recognition, emotion recognition, text classification, summarization and text generation are some popular real-world applications (Kołakowska et al., 2014). However, there is a need for evaluating these models. New data is collected every single day, which can be used to (further) investigate state-of-the-art (SOTA) NLP models. This paper will focus on SOTA emotion recognition models and how they perform on film data. Emotion is investigated in the forms of basic emotion and sentiment. Where emotion consists of the sub-categories: anger, disgust, happiness, fear, and sadness. Sentiment will either be positive or negative.

Understanding how and why emotion recognition models achieve certain results can provide essential information on how to improve models. It increases our understanding of how computation is done and how the trained models interpret our language. Emotion recognition models that have audio as input also take speech features into account. The evaluation of these models can thus provide information on how intonation and other speech properties are processed and understood. The dataset used here, called the EMO-film dataset, contains audio files taken from films in original speech and their dubbed over counterparts. This dataset will be used for evaluation. A newly constructed emotion recognition model is trained on this dataset, which implements the linguistic and acoustic features.

In this paper, the robustness of multilingual emotion recognition models to different languages in the film domain will be investigated. It will be investigated if the same emotion is conveyed in dubbed audio of the same scenes and if the emotion recognition models will be able to distinguish the emotions in the same way. The evaluation data has not been used before on the considered models, so results should provide useful insights and prove valuable in the field.

Chapter 2

Related Work

2.1 Textual Emotion Recognition

Textual Emotion Recognition (TER) is the classification of character-based input into emotions. This field has been attracting attention, as it has the potential to be of commercial and scientific use. Many papers have been showing challenges related to TER and how to overcome them. Advice for users of sentiment analysis has been given by Thakur et al., providing information on the importance of the combination of features extracted, type of classification algorithm used and dataset integration (Thakur et al., 2018). In emotion recognition, Deng et al. provided suggested use for word embedding, architecture, and training levels (Deng and Ren, 2021). This information is relevant to the research performed here, as sentiment analysis and emotion recognition are both implemented. This previous research will aid in efficient integration of the pretrained and to-be trained models.

2.2 Combined Model

An architecture that proposes a combined model for emotion recognition is implemented here. The architecture consists of a text model and an acoustic model, which are combined to increase performance of emotion recognition (Tripathi and Beigi, 2018). Higher performance is achieved for the combined models, compared to when they are evaluated separately. The authors of the paper describing this model actually integrate multiple modalities in classification, such as facial expressions and hand movements. The research covered here only considers text and speech features, as features in other modalities are not available.

2.3 Previous Performance

Previous work has been done on the EMO-film dataset by Scotti et al., who implemented a combined text and acoustic model to classify emotion (Scotti et al., 2021). The proposed model is evaluated on multiple datasets, including the EMO-film dataset. A convolutional neural network was the chosen architecture, which achieved state-of-the-art (SOTA) performance on the EMO-film dataset (accuracy of 39.4%, 37.7% and 37.0% for the respective languages English, Italian and Spanish). However, happiness anger and sadness are the only emotions studied. Disgust and fear are left out of classification, so there is a gap to be filled. Performance on the EMO-film dataset was significantly lower than other datasets used, such

as the IEMOCAP dataset, who's lowest accuracy was 70.0% using their model. This difference in accuracy was attributed to the errors made by the automatic speech recognizer. This will be investigated further in this paper. There were no other relevant studies performing multilingual emotion recognition on the EMO-film dataset that was used in this research.

Chapter 3

Methodology

3.1 Development

Programming was done using Python. Importing, modifying and exporting models was done using HuggingFace and the datasets library. Preprocessing of the data is done using the pandas library. The full model, including the model output is stored on GitHub: https://github.com/KaranChand/ASR_Sentiment

3.2 HuggingFace

HuggingFace (HF) was used to provide the SOTA audio-to-text, text-to-sentiment and text-to-emotion model. HF is a platform where the community can upload, train, test, and download models which were uploaded by others. The models used are not trained or fine-tuned specifically on the film data, as that domain has not been explored or published by the community yet. It is important that the models are recent, as emotion recognition models are continually updated and improved upon. HF include reputable companies, such as Google and Facebook. HF also provides a useful library called 'datasets', which allows for accessible modification and conversion of datasets. HF is mainly based on transformers, though they provide many other architectures as well. Many types of data can be found on HF, though the dataset chosen here was not found, nor any that were useful and had similarities to it. The dataset thus needed to be converted to a HF dataset before use.

3.3 Pretrained Sentiment Analysis Model

The multilingual sentiment analysis model chosen from HF is the pretrained twitter-xlm-roberta-base-sentiment model by cardiffnlp (Barbieri et al., 2021). This model takes text as input and outputs a sentiment from the set: positive, negative and neutral. This sentiment analysis model is very popular and well documented. Since the dataset contains emotions, it is required to map the emotions to their respective binary sentiment. This mapping was done as seen in table 3.1. From this, it can be seen that the dataset is not balanced. Secondly, the dataset does not contain neutral sentiments, which need to be removed from model output. This was done using the same methods for emotion output, explained in 3.7.

Emotion	Sentiment
Anger	-
Joy	+
Fear	-
Disgust	-
Sadness	-

Table 3.1: Shows the sentiment mapped from dataset emotions

3.4 Pretrained Emotion Recognition Model

The multilingual emotion recognition model chosen from HF is the pretrained emotion-english-distilroberta-base model by j-hartmann (Hartmann, 2022). This model takes text as input and outputs an emotion from the set: anger, disgust, fear, joy, neutral, sadness and surprise. This is a very popular emotion recognition model on HF, as it covers the basic emotions. The model was trained on 6 diverse datasets, which makes this model a decent performer in multiple domains. The datasets consist of texts from Twitter, Reddit, student self-reports, and utterances from TV dialogues. TV dialogues are comparable to the EMO-film dataset used here, which makes this model interesting to investigate. Generalizability, popularity and pretrained properties are the main reasons behind choosing this model.

3.5 Audio-to-Emotion Model

The third model used for evaluation consists of three separate audio-to-emotion models, as a pretrained multilingual model was not available. Each model was trained on a single language, though the structure remained the same. The English text model makes use of GloVe word embeddings, which are pretrained and open-source (Pennington et al., 2014). This is used for measuring the linguistic and semantic similarity of words. The GloVe embedding input length was set to 10, with a vector size of 300. Wav2Vec embeddings were used for Italian and Spanish models, as GloVe embeddings were not available. Wav2Vec embeddings had windows of length 10, with a vector size of 100. These embeddings originate from the CoNLL 2017 Shared Task, which automatically segmentated, tokenized and annotated raw texts in 45 languages (Ginter et al., 2017). Long short-term memory (LSTM) is used, containing 256 units and a dropout of 20% per layer. An acoustic model is developed to be combined with each text model. This model integrates acoustic features using the Mel Frequency Cepstral Coefficient (MFCC), Mel-spectrogram and Chromagrams. They are then combined to increase recognition accuracy. The architecture of the combined model is shown in figure 3.1. This architecture is based on a previous study that combined the two models in this way (Tripathi and Beigi, 2018). Like the paper, text and acoustic models are combined at the final layer.

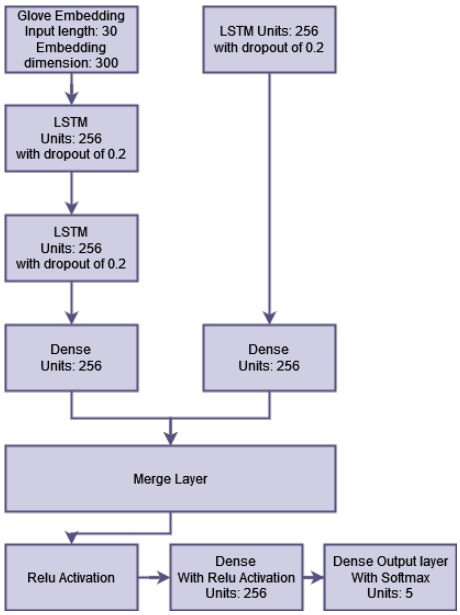

Figure 3.1: Shows the architecture of the combined model

3.6 Audio-to-Text

A pretrained Whisper model is used to provide the text features from the audio (Radford et al., 2022). The Whisper model is chosen for extracting textual features, as it is a multilingual model and automatically outputs punctuation. Punctuation is of importance in emotion recognition, as it can convey useful information. This textual information is used for all models, though the combined model makes use of speech features as well. It should be noted that the pretrained models only use the output from the Whisper model to classify the emotions.

3.7 Model Emotions $\neq$ Data Emotions

It can be seen from the outputs of the emotion recognition model that they do not match the emotions of the dataset fully. The EMO-film dataset emotion set is a subset of the emotion recognition model output. The design choice was made to reduce the model output to only output the subset, as it would not be comparable to the dataset. This was done by sorting the emotion classification output on probability and removing the subset from the model output. Remapping the model output using an additive probability could have been done as well, though this was not implemented due to time restrictions. For example, surprise is not in the dataset, so instead of removing the probability output, it can be spread over the remaining outputs in an intelligent manner. This manner should be based on the literature, though this literature was not found and it was thus not investigated. Remapping the model output would lead to less information loss and might improve results.

Chapter 4

Experiments

4.1 EMO-film dataset

The EMO-film dataset was used for evaluating the models. This dataset has not been used before in the context of the SOTA models that have been chosen. This is a multilingual emotional speech corpus consisting of 1115 audio instances of 43 films and includes transcriptions. Genres include: comedy, drama, horror, and thriller. Languages covered are: original English, dubbed Italian and dubbed Spanish. The mean length of audio instances is $\sim$3.5 seconds. The emotions included are: anger, happiness, sadness, disgust and fear. Notable here is that there is not a neutral emotion in any of the instances. This dataset needed to be converted into a format that is readable by the datasets library. This was done using the pandas library in Python. Preprocessing of the real transcriptions was done by removing the notations for non-verbal sounds denoted by '[non-verbal]'. Ellipses were removed as well, as the Whisper model does not output this and will never classify this type of punctuation correctly. Other punctuation is kept, as they can be compared to the output of Whisper. The output of the Whisper model will be compared to the real transcriptions, to see how it will affect emotion recognition and sentiment analysis.

4.2 Combined Model Training Settings

Training data was randomly chosen using a 75:25 train:test split on the available data. Batch sizes of 128 were created. The model trained for 10 epochs using the full training data every epoch. The Adam optimizer was used, in combination with categorical cross-entropy loss. Due to time and resource restrictions, training was not extended further.

4.3 Robustness to Whisper

To show how much the Whisper automatic speech recognition affects emotion recognition, classification on original transcriptions is compared to classification on Whisper output. These results could also be interpreted as how robust the models are to noise in textual data.

Chapter 5

Results

The Word Error Rate (WER) and Character Error Rate (CER) of the Whisper model is shown per language in table 5.1. Whisper has the most trouble recognizing the Italian audio with a WER and CER of 0.30 and 0.13, respectively. This is relatively high compared to the WER and CER of the recognized English audio (0.17 and 0.07).

Language	**WER**	**CER**
English	0.17	0.07
Italian	0.30	0.13
Spanish	0.25	0.09

Table 5.1: Errors made by the automatic speech recognizer: Whisper. WER and CER are shown for all 3 languages.

The sentiment analysis model output is shown in figure 5.1. The output of the Whisper model does not seem to have a great impact on the classification of the sentiment analysis model, as overall accuracy stays around 70%. The EMO-film dataset is unbalanced in positive and negative instances, so True-Positive Rates (TPR) and True-Negative Rates seem like more useful metrics. The real transcriptions and Whisper transcriptions lead to a TPRs and TNRs found in table 5.2.

Language	**TPR**	**TNR**	**TPR with Whisper**	**TNR with Whisper**
English	0.56	<u>0.73</u>	<u>0.59</u>	0.72
Italian	0.44	<u>0.77</u>	<u>0.51</u>	0.76
Spanish	0.42	<u>0.79</u>	<u>0.45</u>	0.78

Table 5.2: Performance of the classification, evaluated using TPR and TNR for all languages. Performance on real transcriptions is compared to when Whisper transcriptions were used. Underlined rates represent the best performance.

Figure 5.1: Confusion matrices: Classification of the pretrained sentiment analysis model. The top row contains classifications of the real transcriptions, while the bottom row contains classifications of the transcriptions given by the Whisper model.

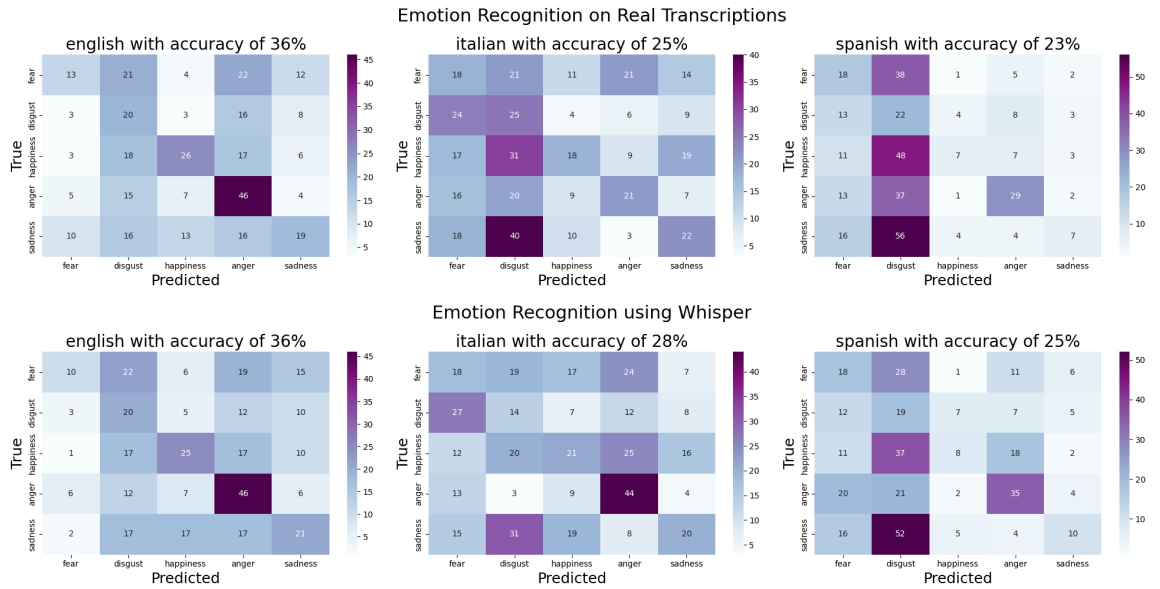

Figure 5.2: Confusion matrices: Classification of the pretrained emotion recognition model. The top row contains classifications of the real transcriptions, while the bottom row contains classifications of the transcriptions given by the Whisper model.

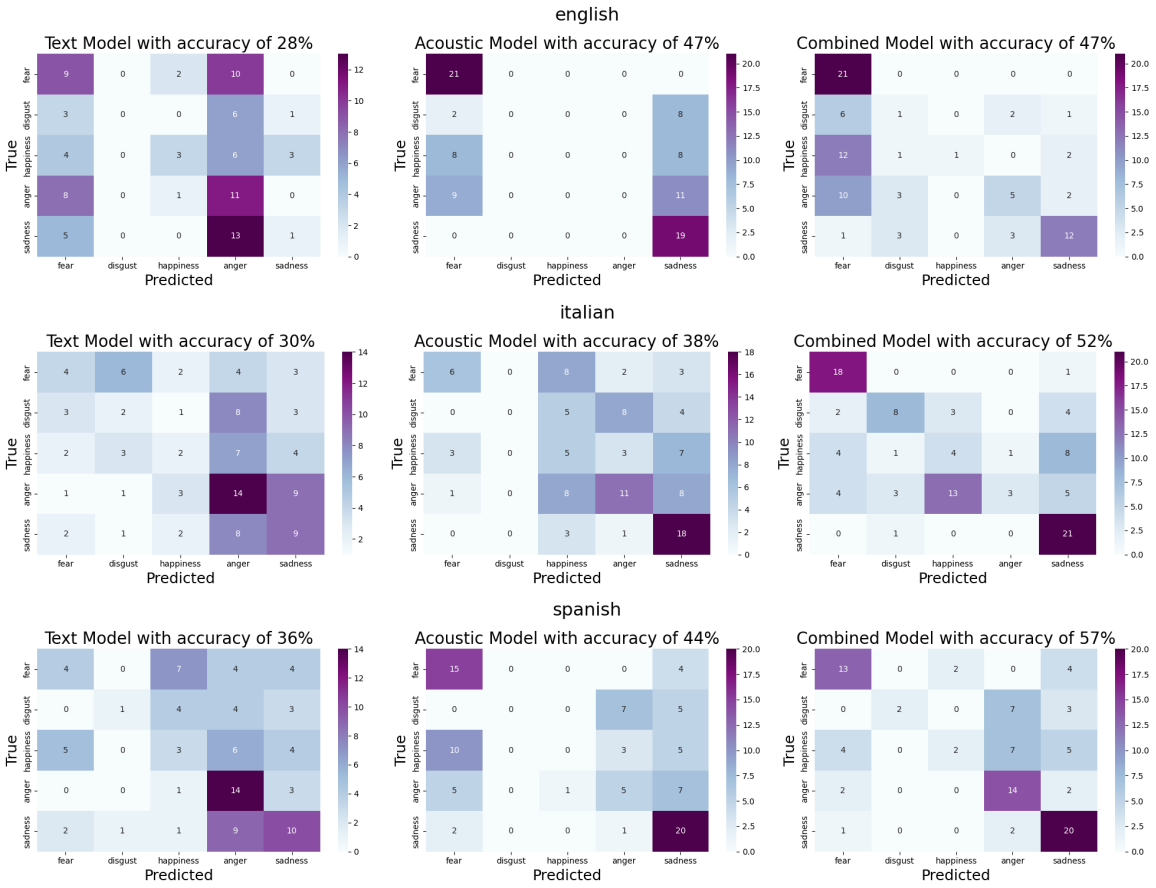

Figure 5.3: Confusion matrices: Classification of the newly trained combined emotion recognition model, consisting of a text model and an acoustic model. Each row shows evaluation on a different language, respectively representing English, Italian and Spanish. Each column of confusion matrices respectively represents the performance of the models individually: text model, acoustic model, and text + acoustic model

Chapter 6

Discussion

6.1 Robustness of Pretrained Sentiment Analysis Model

It can be seen from table 5.2 that the speech recognition done by Whisper does not cause emotion recognition to drop substantially compared to classification on real transcriptions. Even a WER of 0.30 for the Italian dataset affects the emotion recognition minimally. The pretrained sentiment analysis model is thus very robust to Whisper output. The cause for this is most likely the core of sentiment analysis: The sentiment of a sentence can most of the time be trivially computed, as a negative or positive word will weigh the corresponding classification heavily. The Whisper model thus needs to correctly recognize the emotionally charged words correctly to determine sentiment. Following this logic, emotion recognition should be harder to perform. The languages achieve similar TPR and TNR, seen in table 5.1, so the pretrained model is seemingly balanced in evaluating various languages.

Interesting to note is that inclusion of the Whisper model increases the TPR for all languages. The cause for this is unknown, though it is speculated that the Whisper model adds emotional punctuation to the transcription that was not in the real transcription. The real transcription is subjectively and possibly inconsistently annotated by a human. The Whisper model might be better at expressing emotion in text than humans. This effect is also seen in table 5.2, as performance using Whisper transcriptions is either the same or greater than using the real transcriptions. This was not further investigated due to time restrictions.

The pretrained sentiment analysis model is not a very good estimator of the true sentiment. Negative sentiments are classified correctly only around 75% of the time, while positive sentiments are wrongly classified on around chance level. A reason for this could be that no finetuning on the dataset was performed, which may have led to the low performance. A second reason for low performance might be caused by disregarding speech features. The Whisper model only provides text features, which are likely to be insufficient for correct classification. The emotional information integrated in tone and flow of speech is ignored.

6.2 Robustness of Pretrained Emotion Recognition Model

Similar to the pretrained sentiment analysis model, the pretrained emotion recognition model is rather robust Whisper output. Overall classification accuracy on Whisper transcriptions seems to either be at the same level or even higher than real transcriptions. Recognition on

the English dataset is substantially higher than the other languages, so the pretrained model seems to have an affinity to the language. This could be due to greater/longer training or even more available training data for this language.

An interesting feature of the Spanish classification is that there seems to either be a bias in the model or in the data. 'Disgust' is predicted in a greater amount than the other emotions, as is seen in figure 5.2. The model could have been trained insufficiently on Spanish emotions, leading to the model classifying the utterance as 'disgust'. Or this could be due to the emotion mapping that was applied, removing the 'neutral' and 'surprise' emotions. The model might have chosen for a removed emotion, if it could. This would still have been wrong given the utterances, though the type of emotion it would have classified could have been different. This was not investigated further. A bias in the dataset could be that the Spanish dubbed utterances contain less emotion than the other languages. This would make it more difficult for the multilingual emotion recognition model to correctly classify this language. This hypothesis is not further explored due to time restrictions.

The model performs above chance level in every language, though performance can be increased substantially from this point. Just as in the sentiment analysis model, text features alone might not be sufficient for a high performing model.

6.3 Robustness of Newly Trained Emotion Recognition Model

The confusion matrices seen in 5.3 have a similar feature seen in the Spanish classification in figure 5.2. Some models (English text and acoustic, Italian text and Spanish acoustic model) seem to have a bias in choosing a certain emotion. As the emotion chosen is not consistent in the different models, the effect is thought to be caused by insufficient training. As training was only done for 10 epochs, convergence is unlikely. Due to time and resource restrictions, training could not be extended or improved and is left to further research.

However, it is still interesting to see how well the combined model performs compared to the models individually. Except for the English language, the combined model performs substantially better when the text model and acoustic model are combined. Though the English combined model has the same overall accuracy as the acoustic model, the confusion matrix of the combined model is more centered on the diagonal, reducing the bias of the model. This is found in the other languages as well, though the Italian and Spanish models perform substantially better when combined. This result was not expected, as the Italian and Spanish languages are dubbed data. Original English utterances were expected to achieve greater classification performance, as original film speech was thought to be the best carrier of emotion. As shown through limited training, this is not the case. A cause for this might be that actors that dub the utterances might exaggerate emotions, though no further investigation was done to confirm this.

6.4 Comparison to Scotti et al.

The pretrained emotion recognition models seem to have substantially lower performance compared to the trained multilingual model constructed by Scotti et al. (2021). The English dataset has comparable performance, though it is still lower. This is probably due to the fine-tuning that was performed by the authors, which was not done here. The authors claimed that low performance on the EMO-film was caused by transcription errors introduced by the

automatic speech recognizer. Research shown here contradicts this finding, as the accuracy stayed comparably low, even when classification was done on real transcriptions. Even classification on the Whisper model output achieved quite high performance (maximum WER of 0.30 and CER of 0.13). This shows that either the emotion recognition model is insufficient or the dataset used is not of high quality. Given the scarcity of studies done on the EMO-film dataset and the limited training that has been done, an answer to this question is not available.

It is interesting to see that the combined models achieve higher performance than the combined model constructed by Scotti et al. (accuracy of 0.47, 0.52 and 0.57 compared to 0.39, 0.38 and 0.37 for the respective languages English, Italian and Spanish). Overall, performance of the models is not high, though they are substantial increases compared to performance achieved by Scotti et al. A reason for this might be that the combined model trained by Scotti et al. was trained on multiple languages and datasets. This increases the generalizability of the model, though performance on a particular dataset might decrease. Their model is a multilingual model, which is different from what was constructed here. Three models were constructed, each classifying a separate language. Further research should thus combine these three models to fully compare the model to the model made by Scotti et al. Scotti et al. found that the EMO-film was of significant lesser quality than the other datasets used, which might cause the decrease in performance as well. As far as we know, the combined model constructed here reaches SOTA emotion recognition performance on the EMO-film dataset. However, generalizability to other datasets is not investigated, which potentially leads to further research.

Chapter 7

Conclusion

This paper aimed to research the robustness of SOTA multilingual emotion recognition models. A pretrained sentiment analysis model, a pretrained emotion recognition model and a newly constructed and trained combined text and acoustic model were chosen for evaluation. The EMO-film dataset contained English, Italian and Spanish utterances, which were used to measure performance. The Whisper transcription errors affected the classification of emotions positively, contrary to former belief. The sentiment analysis model was not very robust to positive sentiments in any language, as the TPR neared chance level (∼50%). A higher performance was found for TNR (∼75%), though overall performance was lower than expected. The pretrained emotion recognition model achieved above chance performance for all languages, though there seems to be a bias in the model and/or data. The pretrained models are thought to have insufficient modality features to make a correct classification, as speech features are excluded. A combined text and acoustic model is implemented for all languages, though limited training is performed. The combined model achieved substantial increases in performance for both the Italian and Spanish utterances (max accuracy of 52% and 57%, respectively). Classification on English utterances kept performance on the same level as the acoustic model (47%). Substantial increases in classification performance on the Italian and Spanish utterances is hypothesized to be caused by exaggeration of emotion by the dubbing actors. SOTA performance is reached on the EMO-film emotion recognition task using the combined model.

Bibliography

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2021). Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. *arXiv preprint arXiv:2104.12250*.

Deng, J., & Ren, F. (2021). A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*.

Ginter, F., Hajič, J., Luotolahti, J., Straka, M., & Zeman, D. (2017). CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. http://hdl.handle.net/11234/1-1989

Hartmann, J. (2022). Emotion english distilroberta-base.

Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., & Wrobel, M. R. (2014). Emotion recognition and its applications. *Human-Computer Systems Interaction: Backgrounds and Applications 3*, 51–62.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. https://doi.org/10.48550/ARXIV.2212.04356

Scotti, V., Galati, F., Sbattella, L., & Tedesco, R. (2021). Combining deep and unsupervised features for multilingual speech emotion recognition. *International Conference on Pattern Recognition*, 114–128.

Thakur, P., Shrivastava, D. R., & DR, A. (2018). A review on text based emotion recognition system. *International Journal of Advanced Trends in Computer Science and Engineering*, *7*(5).

Tripathi, S., & Beigi, H. (2018). Multi-modal emotion recognition on iemocap with neural networks. *arXiv preprint arXiv:1804.05788*.