Multimodal emotion recognition for video content

In collaboration with RTL Nederland

Lyuba Polyanskaya

Daily supervisor: Dr. Daan Odijk
First examiner: Dr. Ben Harvey
Second examiner: Dr. Chris Janssen

Master thesis

Utrecht University

Artificial Intelligence
Utrecht University
The Netherlands
2019

Abstract

There is a huge body of research dedicated to automatic emotion recognition from facial expressions and from texts. Although previous work has shown that combining the information learned from these channels improve the quality of predictions, this combination is underrepresented in the bimodal emotion recognition domain. Our research is aimed to close this gap by creating an emotion recognition model that joins facial and textual emotion classifiers. Two building components for this bimodal model are 1) the high-performing convolutional neural network mini-Xception which is responsible for the facial emotion recognition and 2) the BERT model of word embeddings fine-tuned on the textual emotion classification task. Firstly, we investigated if engaging the textual and video modalities would improve the quality of emotion classification. Secondly, we evaluated the high-performing models that were employed for the facial (mini-Xception) and textual (BERT) emotion predictions on a new dynamic data source which was the Dutch soap opera *"Goede Tijden, Slechte Tijden"*. Our result showed that the performance of these two models on the new data is not high: Mini-Xception gained 0.17 macro F1-score, fine-tuned BERT - 0.26 macro F1-score. As far as the bimodal model is concerned, fusing BERT with mini-Xception did not improve the classification: the bimodal model (Random Forest) turned out to perform worse (macro F1-score 0.22) than the textual emotion classifier, but slightly better than the facial emotion classifier. All in all, this research demonstrated the necessity of cross-dataset evaluation for high-performing deep learning models. Although the bimodal model did not outperform unimodal models as expected, joining the modalities can still be an efficient approach for automatic emotion classification. In our set-up there were some obstacles for the bimodal model that may have resulted in poor performance. First, the task of annotating soap opera shots with emotions was formidable even for human annotators. In addition, the bad predictions of the unimodal models resulted in the low bimodal performance.

Key words: Multimodal emotion recognition, Deep learning, Cross-dataset evaluation

Contents

1 Introduction

In 2000, a group of scholars and practitioners shared their vision on the future of artificial intelligence (Simon et al., 2000). One of their predictions stated that AI research would be directed towards recognizing and understanding different instantiations of human emotions. Now, this forecast comes to life. Sentiment analysis and emotion recognition in texts (Chopade, 2015), automatic facial emotion recognition (Fasel and Luettin, 2003; Sariyanidi et al., 2015), and emotion recognition in speech (El Ayadi et al., 2011) are becoming well-defined research areas.

Systems that are able to recognize human emotions are not only interesting from the scholarly perspective. Such systems also have various applications in everyday life. They can improve human interaction with robots (Alonso-Martín et al., 2013), provide users of automatic tutoring systems with more personalized experience (Litman and Forbes, 2003), monitor car-racing drivers to prevents accidents (Katsis et al., 2008). In addition, emotion recognizing artificial intelligence is relevant for the entertainment industry. For example, the flow of video games can be linked to player's emotional responses (Yannakakis, 2012), and movie recommendations can be adjusted to a user's mood (Ho et al., 2006).

Due to the way people express emotions, creating an automatic emotion recognition algorithm is a complex task. Emotions are not homogeneous in the way they are portrayed. Cowie et al. (2001) described them as co-occurrence of different types of events. These events are realized through different modalities: gestures, body postures, movements, verbal messages, non-verbal voice characteristics (prosody, intonation), and facial expressions. There is an extensive body of research dedicated to recognizing emotions within one of these modalities. In the unimodal emotion recognition domain, speech and facial expressions received the most attention from researchers. There are numerous algorithms and techniques that analyze the textual information (Chopade, 2015), acoustic characteristics of voice (El Ayadi et al., 2011), detect faces and their expressions (Fasel and Luettin, 2003; Ko, 2018; Sariyanidi et al., 2015). There are, however, other types of signals that can be employed for the emotion recognition task: physiological signals received through EEG (Liu et al., 2018), body gestures and postures (Noroozi et al., 2018).

Since humans are usually exposed to all of the above-mentioned modalities while analyzing other people's emotions, it is logical to assume that each channel has its unique contribution to emotion recognition. De Silva et al. (1997) conducted an experiment in which humans had to identify an emotion expressed in a video clip either by means of a video or an audio cue of the clip. It turned out that some emotions, such as sadness and fear, were better recognized with the audio cue, while other emotions, such as anger and happiness, were better recognized with the video cue.

Busso et al. (2004) adapted the findings of De Silva et al. (1997) and further assumed that fusing different modalities into one emotion recognition system would produce better results than unimodal emotion recognition systems. They conducted research in which the video clips of an actress expressing three emotions (sadness, happiness, anger, plus a neutral state was included) were analyzed by three different systems. One of the systems extracted geometrical markers on the face and translated them into emotion labels, the second one took acoustic features of the actress's voice, and the third one combined the above-mentioned classifiers into one. The bimodal approach had the highest absolute performance across different emotions (89% accuracy averaged across emotion classes). It also successfully classified pairs of emotions that were often confused in unimodal systems.

With the advances in computational power and the flourishing of machine and deep learning algorithms, combining different modalities into one emotion recognition system is gaining more and more popularity in industry and academia. This research is dedicated to developing such a bimodal emotion detection algorithm that fuses the textual and facial emotion recognition

models. This bimodal model, along with its unimodal components, is tested on the video and subtitle content of the Dutch soap opera *"Goede Tijden, Slechte Tijden"*.

In the following part of the introduction we will show how this research contributed to the automatic emotion recognition domain. We will overview different types of facial, textual, and multimodal emotion recognition models, discuss the basic principles of their architectures, state-of-the-art performance metrics reported for these types of classifiers, and their current limitations.

1.1 Psychological foundations of automatic emotion recognition

In order to create an automatic emotion classifier, it is first necessary to define what emotion is. This research adapts the definition of *emotion* provided by Jeon (2017). Emotions are defined as physiological responses of the brain and body to stimuli from the environment. Emotions have a distinct cause, are brief in duration, and hence are relatively intense.

There are different approaches to describing emotions. One of them is through discrete categories (Zeng et al., 2009). The most popular categorical description is created by Ekman (1971). His cross-cultural study proved that there are six basic emotion categories (happiness, sadness, fear, anger, disgust, and surprise) that are universally recognized by humans. As well as having defined what these basic emotions are, Ekman and Friesen (1978) came up with the Facial Action Coding System (FACS), the comprehensive, anatomically based descriptors of facial movements associated with different emotions. Due to its elaborate descriptive scheme and easy adaption of discrete classifications to pattern mining algorithms, Ekman's classification of emotions received much attention in automatic emotion recognition in the computer vision research field from its dawn (Zeng et al., 2009).

Despite the prevalence of discrete labelling frameworks (such as Ekman's classification) in automatic emotion recognition, there are some limitations in this approach. Cowie et al. (2001) state that in rich contexts of everyday communication it can be hard or even inappropriate to put a label on people's facial expressions. There are dimensional approaches to emotion description that are aimed to overcome this downside by describing an emotion through different scales (Zeng et al., 2009). For example, the emotion theory of Russell (1980) defines an emotion through two coordinates: how positive/negative it is and whether a person is going to act or stay passive in this affective state. Although continuality of dimensional approaches is an advantage, there is only a small amount of data sources available for training an emotion recognition model in this framework (Mollahosseini et al., 2017).

In addition, even though Ekman and Friesen (2003) stressed the universality of basic emotions, they also admitted that the culture influences the way these emotions are displayed via *cultural display rules*. In the research (Ekman and Friesen, 1969) it was shown that Japanese participants tended to hide negative feelings (such as disgust, anger, fear, and sadness) with a smile in the presence of an experimenter, while American participants continued to express these emotions. In this study, we assume that the way Dutch people portray emotions are similar to the way people in English-speaking parts of the world do. Thus, methods and data derived for automatic facial emotion recognition from the latter group can be used for the facial emotion analysis of the Dutch content.

In addition, the deception of emotions can also be an issue for emotion recognition along with the ambiguity of displayed emotional signals: lowered eyebrows, for instance, may be an indicator of concentration as well as anger. Hence, Cowie et al. (2001) claim that engaging different modalities to automatic emotion classification should help resolve some ambiguities in signal and enhance the performance of emotion recognition systems. This study follows this recommendation.

Despite the pinpointed downsides, in this study we adapt the Ekman's classification of emotions for our multimodal emotion recognition model. Ekman's classification is by far the most detailed when it comes to automatic emotion recognition, as it covers such a wide range

of emotions. Joining the textual and visual modalities should help in resolving the problems and ambiguities in emotion recognition that we discussed above.

1.2 Automatic facial emotion recognition

Since many multimodal approaches for automatic emotion recognition in video (together with this research) register emotions portrayed in faces, overviewing the domain of facial emotion recognition is required.

In the recent review of facial emotion recognition (FER), Ko (2018) divided the algorithms developed for this task into two large groups: conventional and deep-learning based approaches. The term *"conventional"* there is used as an opposition to deep learning.

The conventional algorithms tend to have a common architecture. They consist of 1) registration of faces and facial components; 2) extracting features; 3) training the classifier on the extracted features. As far as the first stage of the pipeline is concerned, the Viola-Jones algorithm (Viola et al., 2001) is frequently used due to its low computational cost and good performance (Wang, 2014).

In comparison to deep learning algorithms, in conventional approaches the features are handcrafted. They can be geometric or appearance based. The geometric features track the distance between predefined geometric landmark positions (i.e. the tip of the nose, the forehead centre). Appearance based features are extracted using the textural variation of face images with the help of different descriptors, such as Principal Component Analysis (PCA), Gabor filters, Local Binary Patterns (LBP) (Yu and Liu, 2015).

The features can be utilized both for static and video-based FER. Static FER relies only on the static facial features obtained from a frame (image), while video-based FER captures the variations between features in consecutive frames.

Finally, the derived features can be inputted to a training algorithm. In conventional FER Support Vector Machines (SVM) is frequently used (Ko, 2018).

Nowadays state-of-the-art results in automatic emotion recognition and other fields of computer vision, such as object detection, face recognition, and scene understanding, all belong to deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

Processes analogous to face extraction, feature extraction, and classification arguably underlie the performance of deep network implementations. The distributed nature of CNNs and RNNs representations can, however, obscure the processes involved into a black box. Hence, in some implementations of deep learning models some parts of the analysis pipeline are substituted with explicit and more intuitive algorithms. For example, in the model that is used in this research (Arriaga et al., 2017), the Viola-Jones algorithm is applied for the face registration step.

In the model that is implemented for facial emotion recognition in our research, feature extraction and classification steps are performed by a CNN. CNNs contain three different types of layers: convolutional layer(s), pooling layer(s), fully connected layer(s). Convolutional layers take an image or a feature map as an input and multiply the input values with a filter (or a set of filters) in a sliding manner. Filters used in CNNs tend to work like edge detectors (Ng, 2018). After the convolution, the result can be inputted into a pooling layer aimed at reducing the dimensionality of the feature space, hence, improving computational costs and smoothing local variations, making the features more robust (Ko, 2018). Pooling layers take the highest (max pooling) or an average value among neighbouring values in a sliding manner. Usually a fully connected layer or layers are the last step of a CNN. After the last fully connected layer the probability of an image belonging to a particular class is computed.

Since a CNN analyzes static, frame-based data, hybrid approaches for covering temporal variations were developed. The combination of a CNN for retrieving spatial features from individual frames and an RNN with long short-term memory units (LSTM) for analyzing temporal

features in consecutive frames received much attention in computer vision (Ko, 2018). Since deep learning networks, such as RNNs, require extensive amount of labelled data for training (which is not the case in our research) and the representations they learn may be not interpretable, heuristical approaches (Lin et al., 2013) for covering temporal variations in data were developed.

As this research employs the CNN-based model together with a heuristical approach for analyzing facial expression and capturing dynamical changes in video, RNN architecture is not covered in this overview. For more information see Ko (2018). The model's architecture used in this research for facial emotion recognition and the heuristical approach used for adapting the model's frame-based predictions to video data are explained in Sections 2.4.1 and 2.4.2 respectively.

1.3 Automatic emotion recognition in texts

Besides facial emotion recognition, our research also includes deriving emotion labels from textual data. The task of automatic emotion recognition in the textual domain is similar in its essence to the task of sentiment analysis. Emotion recognition, or emotion mining, however, is a more complicated task, as it covers a more fine-grained classification of data. Depending on the theoretical framework on which the emotion recognition systems are based different emotion categories are used. Popular models adapt Ekman's set (or a subset) of basic emotions (Becker et al., 2017). Assigning these emotion categories can be formulated as a typical pattern mining problem. Depending on the availability of training data, supervised learning, such as SVM or Naive Bayes (Becker et al., 2017), can be performed. If there is no labeled data, researchers use unsupervised learning algorithms, such as latent Dirichlet allocation (LDA) (Lin et al., 2011), or heuristical approaches based on emotion lexicon (Mohammad, 2016).

Emotion recognition in texts bears some challenges (Becker et al., 2017). First, there is no consensus on the definition of emotion (Munezero et al., 2014) and, hence, which emotion model should be adopted. Each emotion requires its own resources (e.g. labeled texts, emotion lexicon) that can differ drastically depending on the psychological framework in which an AI researcher decides to work.

The second large challenge for the task is the drastic underrepresentation of non-English resources. In comparison to the task of facial emotion recognition where datasets with labelled facial expressions (especially the ones with a variety of ethnicities presented, such as FER-2013 Goodfellow et al. (2013)) can be adapted to new image and video data from a different culture, the task of textual emotion recognition usually requires to have language specific emotion lexicon or labelled corpus.

This problem was highlighted in several overviews of the emotion and sentiment analysis domains (Becker et al., 2017; Cambria et al., 2017; Mohammad, 2016). Researchers usually combat this problem by translating the target data into English or translating labeled emotion lexicon from English to the target language (Dashtipour et al., 2016). The latter approach has been used to adapt the NRC Word-Emotion Association Lexicon, which was primarily annotated for the English language (Mohammad and Turney, 2013), to Dutch. This translated emotion lexicon is the only Dutch source for automatic emotion recognition in open availability.

A more recent trend in the domain of textual emotion recognition/sentiment analysis is focused on employing *word embeddings* pre-trained on a large corpus (Becker et al., 2017). *Word embeddings* are numerical representations (vectors) of words (Jurafsky and Martin, 2018). Given a corpus of texts a word embedding can be learned from other words surrounding a given word. By means of such a pre-trained embedding model, the textual data is transposed into numerical feature space that can be further used for training a machine learning classifier on any specific NLP task, such as emotion recognition. In such a set-up, word embeddings represent a model of a language which, given a sufficiently big size of the corpus from which the embeddings were retrieved, should produce better and more generalizable predictions for the given NLP task

(Arora et al., 2016). There are several models that produce word embeddings for English and other languages: word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017). All these models have been used to produce word embeddings for Dutch.

Recently Google published[1] a new open-source model for obtaining word embeddings which is called Bidirectional Encoder Representations from Transformers (the BERT model). This model is the first model that creates *bidirectional* word embeddings (Devlin and Chang, 2018). Bidirectional word embeddings reflect not only the right or the left the context of a word, but the neighbours of the word from both sides. This bidirectional context was hard to achieve in existing embedding models (Devlin and Chang, 2018). In addition, Google released a multilingual word embedding model that included Dutch.

In the paper where BERT was introduced (Devlin et al., 2018), the researchers fine-tuned BERT's English model for eight language understanding tasks, such as question answering, sentence inference, textual similarity, acceptability judgment, and, finally, sentiment analysis, the task that is the closest to our research. The size of the datasets varied from 392,000 classification objects to 2,500. The reported performances of BERT improve the existing state-of-the-art models in all NLP tasks where BERT was tested. Despite BERT's successful performance, six out of eight tasks that Devlin et al. (2018) tested BERT on were binary classification tasks. The sentiment classifier fine-tuned on the binary task (positive/negative labels) achieved 94.9% accuracy. The performance of BERT embeddings on a multiclass problem and, moreover, for non-English data is still an open question.

Since one of our goals is to test this newly developed high-performing model BERT on a multiclass problem and on a new non-English data source, in Section 2.4.3 a detailed description of the model architecture and how it is employed in our research is given.

1.4 Types of multimodal emotion recognition models

Fusing textual information with facial expressions is one of the configurations of the multimodal emotion recognition task. In this subsection, this and other possible modality combinations of multimodal emotion recognition research are discussed.

Multimodal emotion recognition systems usually take audio and video cues as an input. As far as the particular features of these cues are concerned, extracting emotions depicted by facial expressions is the most popular task for the video cue (Haq and Jackson, 2010). Combining acoustic features from an audio signal with emotion recognition from facial expressions is by far the most popular combination in the multimodal emotion recognition domain (Pérez-Rosas et al., 2013; Poria et al., 2017, 2016).

Moreover, there is an annual competition "Emotion Recognition in the Wild"[2] (EmotiW) held for the audio-visual emotion recognition task. Every year, one of the challenges in the competition consists of assigning one emotion category from Ekman's classification to a video clip derived by means of visual (facial expressions) and acoustic (non-verbal) channels. The labeled corpus for training and testing the competing models consists of short video clips cropped from TV-shows and movies (in English). In the last three years (2016 - 2018) around 100[3] teams participated in this competition each year. The performance (accuracy) of the winning algorithms increased from 41% in 2013 (Kahou et al., 2013) to 60.34% in 2017 (Hu et al., 2017).

A less extensive body of research is dedicated to emotion recognition systems that combine video, audio, and textual features. In this area, many papers are dedicated to sentiment classifiers. The multimodal sentiment classifiers, such as (Ellis et al., 2014; Poria et al., 2015a, 2016; Rosas et al., 2013), output a polarity (positive or negative, neutral state is included only in (Ellis et al., 2014)) in a video fragment. As far as emotion recognition is concerned, there

[1] https://github.com/google-research/bert

[2] https://sites.google.com/view/emotiw2018

[3] https://drive.google.com/file/d/1-mVVbabm8ePTMJKwO0itdMXB3j5vEw7h/view

are two models that combined textual, video, and audio features according to the review of Mollahosseini et al. (2017). The model of Poria et al. (2015b) employed the combination of hand-crafted geometrical facial features, lexicon-based features for textual data, and acoustic features fused on the feature level by means of SVM[4]. The accuracy of the model across 7 emotion classes (anger, disgust, fear, happiness, neutrality, sadness, or surprise) is 87.95%.

In a more recent research Poria et al. (2016) compared three unimodal emotion (sadness, happiness, anger, neutrality) classifiers trained on textual, video, and audio content of different video corpora. The performance of all the bimodal combinations (video + audio, video + text, text + audio) was better for each emotion than the performance of the respective unimodal models. The performance of the bimodal models in terms of accuracy was 73% (v + a), 74% (v + t), 65% (t + a). The model that combined all three modalities in turn was better than bimodal models: the average accuracy metric of this model was 77%. The researchers deployed different types of neural networks for feature engineering in text, acoustic, and video (facial) data. The final tri-modal model was joined on the feature level by means of multiple kernel learning (MKL).

Although the system that combined the visual (facial) information with the textual content turned out to have the highest performance among the bimodal models in the research of Poria et al. (2016), fusing textual and facial data into one emotion recognition classifier is not wide spread in the multimodal emotion recognition domain. Poria et al. (2016) were not able to find another model with such a modality combination to make the baseline comparison in performance.

High performance of the text + video (facial expressions) combination suggests that speech content (which is language specific) and facial expressions (which are more universal according to Ekman (1971)) is the combination that is the easiest for a classifier to recognize emotions from. This assumption, however, may not be robust due to the way the ground truth labels were collected. Each video clip of the dataset USC IEMOCAP (Busso et al., 2008), which was used by Poria et al. (2016) was annotated by people who were exposed to both audio and video channels. Thus, they were able to see actors facial expressions and hear verbal and non-verbal content of the audio signal. The models developed by Poria et al. (2016) used the same ground truth labels for all modality combinations which may have induced bias in the performance: people were exposed to more sources of information than the models. In our set-up, the channels the human annotators were exposed to were as similar as possible to the information that the models received.

To sum up, this research is dedicated to developing and testing a multimodal emotion detection pipeline which combines textual and facial information. Despite the good performance of text + video modality combination reported in the previous research (Poria et al., 2016), this modality combination turned out to be underrepresented in the domain of multimodal emotion recognition according to our overview. Thus, by filling this void our research can make a valuable contribution to this domain. To limit the scope of our research, the audio (non-verbal) cue was not included in our multimodal emotion recognition model. As mentioned above, the fusion of the textual and facial emotion classifiers is an interesting combination to investigate and evaluate.

1.5 Fusion methods

There are three different approaches on how to join unimodal emotion recognition systems into multimodal ones. The modalities can be fused on the feature, decision, or model levels (Wu et al., 2014). The feature-level fusion is performed by concatenating feature vectors from different modalities and inputting one common vector into one classifier. The feature-level fusion has its

[4]The description of different types of models for facial emotion recognition, emotion recognition from texts, and description of fusion methods are given in the Sections 1.2, 1.3, and 1.5 respectively

downsides. A high-dimensional feature set may suffer from data sparseness, hence, it can be harder for a classifier to find patterns in data. In addition, feature-level fusion does not take interaction between variables into account.

To eliminate some of these disadvantages, many researchers utilized decision-level fusion. In decision-level fusion each modality is trained with separate classifiers and the models' outputs are fused for the final prediction. The decision-level fusion does not suffer from a curse of dimensionality in comparison to the feature-level fusion. Nevertheless, by treating each modality separately, decision-level fusion assumes the independence between modalities. This is a strong assumption, since different modalities (gestures, intonation, facial expressions) usually complement each other in daily communication.

The model-level fusion aims at creating such models that explore the correlation between features of different modalities. The model-level fusion was reported to have a tendency to overfit (Wu et al., 2014).

Despite of the pointed downside, decision-level fusion is performed in this research, since it is easy to implement, especially on complicated architectures of deep learning models which are used in our research. There are different approaches to obtaining the final prediction with decision-level fusion. Simple arithmetical approaches include: majority voting (usually applied to the systems with more than two modalities included), means of the scores, maximum of the scores (Vielzeuf et al., 2017). Other popular approaches include SVM, Multiple Kernel Fusion, and weighted means (Vielzeuf et al., 2017). Weighted mean is the method used by the EmotiW 2016 winners (Fan et al., 2016). It is executed by weighting the score of each model and summing the scores up for the final prediction. The weights are derived by cross-validation on the validation set. The research by Vielzeuf et al. (2017) showed that sophisticated methods for decision-level fusion, such as the ModDrop method or Score Trees, perform worse on unseen data than simpler ones, such as weighted mean.

The fusion techniques that are presented in this research include: means of the scores, maximum of the scores, weighted mean, Random Forest, and SVM.

1.6 Current limitations of multimodal emotion recognition

As discussed in Section 1.4 the combination of text and video cues is underrepresented in the bimodal emotion recognition research, even though it was reported to have good performance in comparison to other modality combinations.

Moreover, the vast majority of the algorithms developed for the multimodal emotion recognition task are based on the corpora for the English language. There are a few studies which explore multimodal emotion recognition for languages other than English: Spanish (Poria et al., 2016), Chinese (Li et al., 2016), Russian (Perepelkina et al., 2018), French (Ringeval et al., 2013). As far as the Dutch language is concerned, we did not find any academic papers on multimodal emotion recognition for this language. There is one video-based Dutch corpus with emotion annotations collected in controlled lab environment in 2008 and reported by Chiţu et al. (2008). This corpus is, however, not publicly available.

Despite the progress being made in adapting emotion recognition systems to non-controlled environments, there is still room for improvement. The vast majority of the performance metrics in the emotion recognition domain are produced by assessing the models on the same source of data that was used for training by means of cross-validation. This evaluation technique may not be a robust performance indicator, especially, for the visual data preprocessing. Visual datasets vary drastically in light conditions, camera positions, the color scheme. These variations are known to influence the performance of deep learning classifiers. Cross-dataset assessment of state-of-the-art deep learning solutions for facial emotion recognition showed that when the classifier is tested on data different from the one it was trained on the performance is close to guessing (Avots et al., 2018). Hence, evaluating state-of-the-art models on new, unseen data from a different source is a relevant research task.

1.6.1 Configuration of this research and its contribution

This research covers three of the above-mentioned gaps in the multimodal emotion recognition domain: an uncommon combination of modalities (facial expressions and textual data) is used here for bimodal emotion recognition, the pipeline is tested on non-English (Dutch) data, and the emotion recognition pipeline is also tested on a new data source recorded in a non-controlled noisy environment.

In order to realize these contributions the following set-up is organized. Our aim is to develop an emotion classifier which would fuse the textual content and facial expressions retrieved from the visual modality. The source data that is used for training and testing this bimodal model is the subtitle and video content of the Dutch soap opera *"Goede Tijden, Slechte Tijden" (GTST)* provided by the media corporation RTL Nederland. Since one 20-minute episode of the soap opera carries a wide palette of emotions, the classification is done on shot basis. A shot, which lasts (on average) 30-40 seconds, is assigned with one of the following emotion labels: anger, disgust, fear, happiness, neutrality, sadness, or surprise. These emotion list is adapted from Ekman's classification of basic emotions.

For the facial emotion recognition task the high-performing open-source model mini-Xception (Arriaga et al., 2017) is employed. The detailed explanation of the model architecture is given in Section 2.4.1. This model is reported to handle well the variations in positions of faces, faces with occluding objects, and other conditions typical for the natural environment. The model performance is 66% (accuracy) across seven classes (anger, disgust, fear, happiness, neutrality, sadness, or surprise). Since this metric was produced by cross-validating one source dataset, testing it on highly variant unseen data of the soap opera is an important evaluation test for the mini-Xception model.

As far as the textual data is concerned, due to underrepresentation of languages other than English in this domain, it is hard to find a dataset with any type of emotion annotation for Dutch. Hence, we created the emotion corpus on the basis of GTST subtitle data. The BERT model released by Google in November 2018 (Devlin et al., 2018) is used for training the model (the BERT architecture is explained in Section 2.4.3). BERT represents the collection of word embeddings for 104 languages (including Dutch) which can be used to train models for *any* natural language processing task. Since the model was published very recently and its performance on non-English data has not been yet reported in academic papers, testing BERT on Dutch textual data contributes to the domain of emotion recognition in texts.

The facial emotion recognition and text emotion recognition models are joined on decision level. Five different fusion techniques are tested: maximum score, mean score, weighted mean, Random Forest, and SVM.

1.7 Research questions

This research is designed to answer the following research questions:

Q1 *Does a model that combines textual data and facial expressions classify emotions better than unimodal (textual and facial) emotion classifiers?*

Q2 *What is the performance of the state-of-the-art facial emotion recognition model (the mini-Xception network) and the state-of-the-art NLP model (BERT) on the emotion recognition task in a soap opera?*

2 Method

The set-up of this research can be divided into three steps: preparing video and subtitle data for the annotation tasks (see Section 2.2 for more information), setting up the annotation environment for people that will create the ground truth labels for the classifiers assessment

(Section 2.3 bears a more detailed description of this step). The final step is running the facial and textual emotion recognition models and creating a bimodal fusion. The description of the unimodal models are provided in Sections 2.4.1 and 2.4.3 respectively and the fusion strategies are explained in Section 2.4.4. The overall set-up is depicted in Figure 1. A more detailed overview of these three steps is provided below. Metrics and statistical tests that are used for evaluating and comparing the perfomance of different emotion recognition models and their configurations are described in Section 2.5.

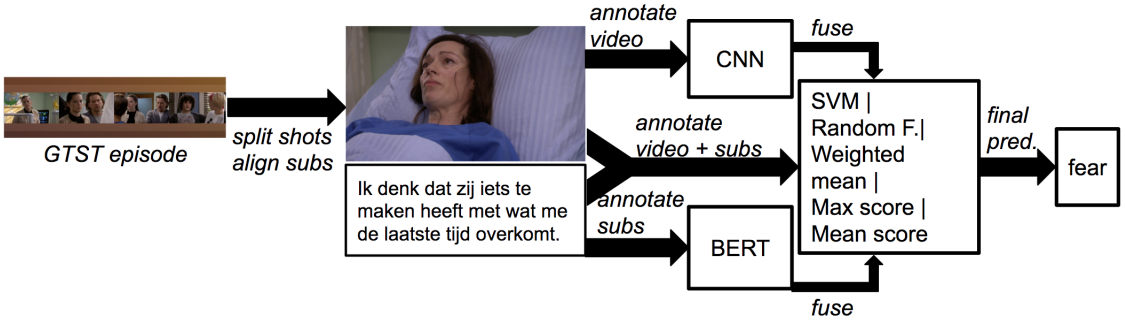


Figure 1: The visualization of the project pipeline.

The input from RTL for our research consists of soap opera video files along with subtitle files with time stamps (for a more detailed description of the provided data see Section 2.1). Since one soap opera episode contains a range of emotions depicted by different actors, we formulated the classification task as the shot-based emotion recognition. Shots are video clips, (usually) brief in duration, recorded from a single camera which constitute building blocks for scenes and episodes. They are physical entities separated by shot boundaries (Rui et al., 1999). One GTST episode is separated into shots by means of a Python library PySceneDetect[5] (more detailed description of this library is given in Section 2.2). When the shots are retrieved, they are cleaned from noise. Noisy shots are considered to be such fragments that contain no faces or multiple faces appearing on the screen simultaneously or one after another (more detailed description of this preprocessing step is given in Section 2.2).

To provide golden standards for assessing performances of the facial emotion classifier (mini-Xception), the textual emotion classifier (BERT-based), and the classifier which combines these modalities, three labelling tasks were organized. These labelling tasks were set up in such a way that human annotators did the same job as the machine learning algorithms in as similar settings as possible. The first group of annotators received a set of video shots with no subtitles and no audio. The second group annotated the corresponding set of subtitle data (same shots, but only the textual content). In addition, the subtitle annotation task had an extra batch of texts to be annotated. This batch was needed for fine-tuning the BERT model (more detailed explanation of BERT architecture and the definition of the fine-tuning step is given in Section 2.4.3). The last group of annotators labelled the same set of video shots where the video content was presented together with the subtitle data (the audio cue was muted). As a result, three labelled datasets were created. The V dataset was retrieved from the first labelling task, the S dataset - from the second labelling task, and VS - from the third one.

When the annotation was complete, the V dataset was inputted to the facial emotion classifier (CNN-based mini-Xception network). The extra batch of the S dataset was used for fine-tuning BERT which was instrumental for developing the textual emotion classifier. When the classifier was fine-tuned, the rest of the S dataset, which included the same shots that were used for testing the facial emotion classifier (the V dataset), was used to test the performance of the fine-tuned textual emotion classifier. Finally, the VS dataset was used for building and

[5] https://pyscenedetect.readthedocs.io/en/latest/

testing the bimodal emotion recognition model.

2.1 Data description

Our project involves using two different types of data: video content of "Goede Tijden, Slechte Tijden" episodes and text subtitles of the same soap opera. In the subtitle files one line corresponds with one utterance (which in turn is a sentence or a series of sentences in 87% cases) and its start and end points. For example:
"00:07:1.560; 00:07:3.760; Dat is natuurlijk een belangrijk evenement." The timings are given in the format "hh:mm:ss.nnn".

2.2 Data preprocessing

The first step of data preprocessing is transforming a soap opera episode video into a series of shots. This step is performed automatically by means of the Python module PySceneDetect. This module detects shot boundaries by comparing the HSV color schemes of two consecutive frames. When the difference is equal or greater than a user defined threshold, the break is registered. During a series of experiments we established this threshold which was sensitive enough to register the changes in the camera positions and not too sensitive to minor changes in the colour schemes, such as movements of an actor's head.

After the shot boundaries are established, the video cues of the shots are aligned with the subtitles. The alignment algorithm compares the time stamps from the subtitle file with the start and end time positions of the shots. If a subtitle or a series of subtitles are within the shot segment, they are assigned to that shot.

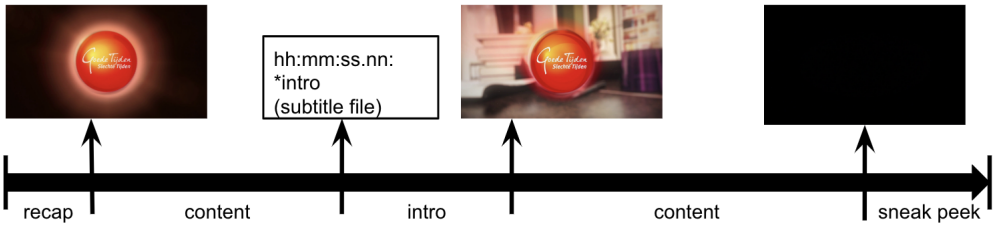


Figure 2: The structure of the GTST episode.

Each GTST episode contains a recap of the previous episode, opening credits, and a sneak peek into the next episode. The shots that fall into one of these three categories are deleted from the annotation set, as they do not carry any episode specific information that is relevant for the emotion annotation. These fragments are detected automatically by a heuristical approach.

Since each episode has a homogeneous structure in terms of where and how recaps, open credits, and sneak peeks are introduced in the subtitle and on the screen, our algorithm compared the video frames in a specific time window to the prototypical frame that always punctuates the end of the recap, the start of the sneak peek, or the end of opening credits (these prototypical frames are depicted in Figure 2). For example, recaps always appear in the very beginning of the episode and last approximately from 30 seconds to 2.5 minutes. Recap always ends with a "Goede Tijden, Slechte Tijden" logo appearing on a black screen (see the first frame depicted in Figure 2). The frame extractor is set in the time window where the intro is expected to end. It compares the typical recap frame to the frames in this time window. When the difference in mean squared error of images' pixels is less than our threshold, the recap boundary is set. The value of this threshold was established experimentally. The same procedure applies to the sneak peek detection: the sneak peek usually takes the last minutes of an episode and it always starts with a last scene fading out. Thus, the frame that punctuates the start of the sneak peek is a black screen. As for the opening credits detection, their start point is always indicated in the

subtitle file (**intro* line states that the intro song is about to begin). We get the approximate starting point of the opening credits from a time stamp in such a subtitle line. We know the average length of an intro song, as it is always the same song for each episode. Hence, we establish the approximate end of the opening credits. The opening credits always end with a GTST logo on a colourful background (the background of such a frame is always the same for each episode). Hence, we can verify our estimation by comparing the frames in the selected time window to the prototypical end of the opening credits (see the second frame in Figure 2). The general structure of GTST segments is depicted in Figure 2.

In order to conduct a valid comparison between unimodal (facial and textual) emotion recognition models and the model that combines these two modalities, the test dataset of shots and corresponding subtitles should be the same for all three experiments. Since the soap opera data was produced in a dynamic, non-controlled environment, we firstly cleaned the shot data from the noise that may confuse human annotators and skew the labelling process. The scenarios that we consider as noise in our research are described in the following paragraph.

As this project concerns modelling emotions from textual data and facial expressions, there should be no conflict between these modalities. Thus, a prototypical shot that is acceptable for testing the models is the shot where there is one face presented through the whole duration of the fragment and this face produces the textual cue. Hence, the cleaning process included eliminating shots with no face presented, with multiple faces presented, shots where there were different people shown on the screen, and shots where an actor shown in the shot was not the speaker, but the listener.

The cleaning step was produced in a semi-automatic manner. The shots with multiple or no faces were detected by means of Python implementation[6] of the Histograms of Oriented Gradients (HOG) algorithm for human detection (Dalal and Triggs, 2005). Then the subset where one face was registered was manually checked. Shots where actors on the screen changed and where the listener was depicted were manually excluded together with the shots with multiple faces.

Figure 3: A shot with two people presented in the video cue and the subtitle.

There were borderline cases that were not excluded from the annotation set. If there were several people depicted in one shot, but there was only *one* person's face visible through the whole fragment, the shot was preserved for the annotation. If there were cues from different people presented in such a shot, the shot was still preserved for the annotation job. If there

[6] https://github.com/ageitgey/face_recognition

were multiple non-neutral emotion states presented in such a fragment, annotators had to choose a corresponding option of the questionnaire. The shots with multiple emotions were later excluded from the research. The full list of the noise options that were available to annotators for reporting and how we used annotator's feedback for the final dataset selection is discussed in Section 2.3.

In Figure 3, an example of a borderline case is represented. In this shot two actors appear on the screen, but annotators can only see one person's face through the whole fragment. The first sentence of the subtitle (*Heel zeker? Really sure?*) belongs to the actress whose back is visible. Since a scenario where there were multiple speakers in the subtitle turned out to be quite a common case, we decided not to exclude such shots from the annotation tasks. Our motivation is as follows. As long as there is one face visible through the shot, the shot depicts a speaker, not a listener, and the annotators did not report multiple emotions presented in the shot and the subtitle, then the line of a non-visible person should be short and non-significant to the overall emotion perception, as in the example from Figure 3. This shot was kept for the annotation tasks.

2.3 Data annotation

The fair comparison between human performance and the performance of three machine learning emotion classifiers requires to set up the research in such a way that the nature of information that a machine and a human annotator receive for judgment is as identical as possible. In addition, to compare the performance of the model that joins textual and facial emotion data to the unimodal classifiers, the set of labelled shots should be the same for three models.

In order to fulfill the above mentioned requirements, three annotation tasks were conducted. The set of 1,005 pre-selected shots were given to three groups of annotators. The first one was to assign the emotion label based on the video content of the shot (V), the second labelled the subtitles (S), and the third one had to watch a video and read the corresponding subtitle to assess emotions (VS). 2,000 extra subtitles were added for training the textual emotion classifier.

Despite the similarity between the set-ups for human judgments and models predictions, our approach yields three different datasets which might have different quality of labels. Annotators who are assigned to the VS set-up are exposed to a larger amount for information in comparison to V and S annotation tasks, and, hence, should make a more informed emotion assessment.

We chose Labelbox[7], an open-source customizable resource for distributing annotation tasks for internal and external teams, for our annotation tasks. Labelbox rendered pre-defined interfaces for the shuffled dataset and equally distributed classification objects among the employees of RTL Nederland who volunteered for the annotation tasks. The project annotation interfaces are depicted in Figure 17. The results of the annotation tasks are provided in Section 3.1.

There were 39 annotators overall. Five people were assigned to the video annotation, four people were assigned to the VS annotation. Since the subtitle task carried 3,005 annotation objects, the remaining 31 people were distributed to this task. The annotators could only contribute to one of the three labelling tasks. Since annotation can be tiring and annoying for humans and the batches are quite big, annotators were instructed to spend as much time as they wanted, they could split the task into several sessions if it was convenient for them. As a result of such a set-up, the level of contribution was different per each annotator. Among 31 subtitle annotators only 11 contributed to labelling significantly (more than 70 labels per person). The descriptive annotation statistics is provided in Section 3.1.

Besides choosing a label for a fragment, annotators could report the noisy conditions that could skew the models performance. They could choose corresponding options if there was no face or multiple faces presented in the video fragment. If the video was extremely short or of low quality, the interface could not play it, hence, the annotators had to tick the corresponding

[7] https://labelbox.com/

box. If multiple sentences in one subtitle were in shuffled order (e.g. *"I liked it. I saw a movie yesterday."*) due to the errors in the subtitle file or in our preprocessing pipeline, the annotators were supposed to report it. If the emotion label was easy to derive in spite of the mixed order, the annotators were instructed to assign the label. The subtitles with such a problem were then manually fixed. The most important type of noise that needed to be registered is when the shot/subtitle depicts different non-neutral emotion states. Multiple people talking in the subtitle does not represent an issue as long as there is only one emotion that can be retrieved. The *"Can't decide"* category was added to the emotion labels as a label for the shots with the problems discussed above and for the shots depicting emotions outside of the list, such as trust. To sum up, the shots and/or subtitles where no face, different faces, multiple emotions, no text in the subtitle, problems with the video playback were reported by the annotators were excluded from the final dataset. In addition to these categories, shots and/or subtitles with the *"Can't decide"* label were also excluded.

2.4 Models description

The architectures of the facial emotion recognition model (mini-Xception), the textual emotion recognition model (BERT), and the way these two models are joined for the bimodal emotion recognition system are described in the following part of the paper.

2.4.1 Mini-Xception

The model that is used in this research for facial emotion recognition on the visual modality is an open-source[8] model called mini-Xception (Arriaga et al., 2017). The model is based on a convolutional neural network (CNN), but its architecture has some specific features aimed at reducing computational costs, improving robustness of features, and model's generalizability. The model uses: residual blocks, depthwise separable convolutions, global average pooling, and batch normalization. Before applying all the above-mentioned operations the visual data is pre-processed. As the network was initially trained on the grayscale dataset FER-2013 (Goodfellow et al., 2013), the image is turned to grayscale. The face boundary is detected by means of the Viola-Jones algorithm.

The Viola-Jones algorithm (Viola et al., 2001) is a widely-used technique for face detection. It has low computational cost, hence, it is useful for application where the processing speed matters (Viola et al., 2001). This algorithm, however, was reported to have a poor performance rate in noisy scenarios (Yang et al., 2016). In comparison to modern deep learning solutions, the Viola-Jones algorithm is worse at detecting occluded faces, small-scale faces, and faces in atypical positions (tilted, turned). These characteristics might affect the mini-Xception performance at a very early stage: the actor's face in the frame might not be registered.

When the facial borders are detected, the algorithm can perform emotion classification. We are not going to describe the principles of convolutional neural networks here (for this information see Section 1.2 and Ko (2018)). In the following part, the modification of CNN (residual blocks, depthwise separable convolutions, global average pooling, and batch normalization) that are implemented in mini-Xception are briefly discussed.

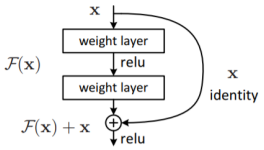


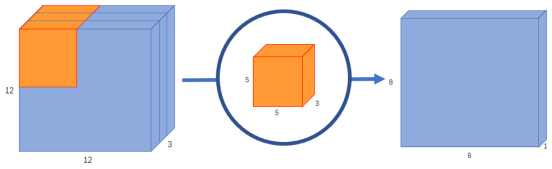
Figure 4: A residual block visualization.

Residual blocks were first introduced in the Residual Learning architecture (ResNet) (He et al., 2016). In plain networks the layers are stacked together in a consecutive manner. It was previously assumed that making a network deeper would improve its performance. The optimization of performance, however, turned out to be not that straightforward. The problem of *degradation* was discovered: with

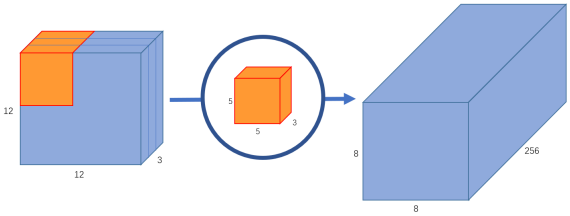
[8] https://github.com/oarriaga/face_classification

the increasing number of layers, the training error stops decreasing at some point (He et al., 2016). ResNets were proposed to solve that problem by adding a 'skip' connection between two non-consecutive layers (see Figure 4). A segment of a network with such a connection is called a *residual block*.

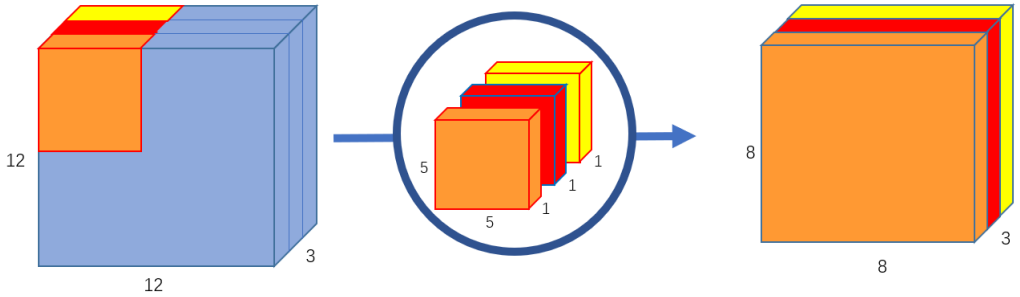
The convolutional layer in mini-Xception consists of depthwise separable convolutions (Howard et al., 2017). The comparison between a standard convolution and a depthwise separable one is depicted in Figure 5. Imagine that we have a coloured RGB image of the size $12 \times 12 \times 3$. After applying one standard filter (5a) the image is flattened to the one-dimensional 8×8 image. After applying 256 filters each result of the convolution is stacked into the $8 \times 8 \times 256$ volume. Depthwise separable convolution divides the operation of applying n standard filters to an image into two steps. Depthwise convolution (5c) applies *one* filter per *one* channel of an image. Pointwise convolution (5d and 5e), firstly, flattens the output of the depthwise convolution with the of a $1 \times 1 \times N_{channels}$ filter, secondly, adds the volume by applying several (256 here) filters of such size. The size of the resulting matrix after applying the standard convolution and depthwise separable convolutions is the same, but the computational cost of depthwise separable convolutions is significantly less. In order to produce the standard convolution for the sample image the computer has to perform 1,228,800 multiplications, while for the depthwise separable convolution this number is 53,952[9].



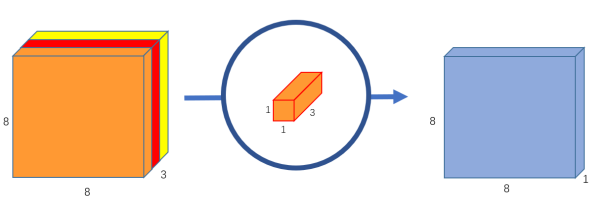
(a) Standard convolution with one filter.



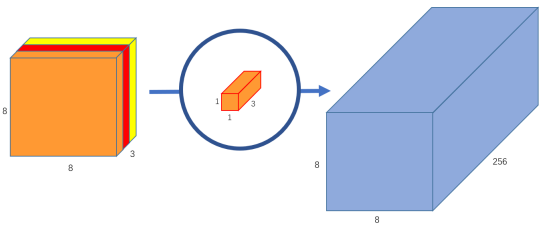
(b) Standard convolution with 256 filters.



(c) Depthwise convolution.



(d) Pointwise convolution with one filter.



(e) Pointwise convolution with 256 filters.

Figure 5: The comparison between standard convolutions (5a - 5b) and depthwise separable convolutions (5c - 5e).

Batch normalization (Ioffe and Szegedy, 2015) normalizes the mean and the variance of

[9]For more detailed explanation of the computational cost see the source article: Wang (2018).

neurons' values in a layer. By applying batch normalization to hidden layers of a neural network the training process can be accelerated.

Global average pooling was developed as a strategy to improve the overall generalizability of a network (Lin et al., 2013). Fully connected layers which are usually placed in the network right before the final prediction is made are known to overfit. Therefore, Lin et al. (2013) proposed to replace the traditional fully connected layers in the CNN with global average pooling as a last step alternative. This method takes n feature maps (where n is the number of classed in the task) at the end of the training process and averages each feature map to one value. The vector of averages of lengths n can be inputted to the softmax layer for a final prediction. The authors of the technique listed three advantages of global average pooling over fully connected layers: it enables correspondence between feature maps and classes. Secondly, there is no parameter to learn and optimize, as it is with fully connected layers, and, hence overfitting is combated. Lastly, it is more robust to spatial variations in the input image.

The mini-Xception network incorporated the modifications mentioned above. It was trained on the grayscale images (FER-2013 dataset) depicting human faces of different age and race in various poses and positions relative to the camera. The dataset is labeled with Ekman's basic emotions (including neutrality). The published performance of the mini-Xception model in the facial emotion recognition task was 66% (accuracy) on average.

We took the mini-Xception model with the architecture described above and queried it for emotion predictions. Given an image, mini-Xception does all the preprocessing and analysis computations discussed above and outputs a vector of length seven which depicts the probability distribution across seven emotion classes. Mini-Xception is a frame-based model, so we investigated how to transfer its output to a video (shot) based classification. Our solution to this problem is described in the following section.

2.4.2 Capturing temporal changes

Since mini-Xception was trained to recognize emotions in static image data, this model needs some adaptation for dynamic video data of the soap opera. In order to incorporate temporal variation of the shot, a modified heuristical approach of Vielzeuf et al. (2018) is used in this research. The approach is originally called *max-average pooling*. In max-average pooling each of the frames is, firstly, inputted into the mini-Xception model. The output of the model is a vector of length seven where each value is a probability of a frame belonging to a particular emotion group (the vector sum is 1). The frames of one shot is divided into n groups. Vielzeuf et al. (2018) set the n value to 16. Then one frame with the highest score in one of seven categories is chosen from each of n groups. At the end the values of n chosen frames are averaged. This averaged vector is the final facial modality vector that will be inputted for the bimodal classifier.

Vielzeuf et al. (2018) showed that using max-average pooling as a technique for capturing temporal variance in combination with ResNets turned out to have higher performance on the facial emotion recognition task than using a more complex LSTM model: 67.1% (max-average pooling) and 58.2% (LSTM) weighted accuracy.

Since the number of groups that the frames are divided into is completely arbitrary in max-average pooling, we decided to modify this approach. We propose three ways of deriving the overall probability vector and an emotion label per video shot. The first approach is to count how many times among the frames of one shot a particular class has the highest probability. The final vector is the count statistics over seven classes (e.g. there were 8 frames in a shot with the highest value of *'angry'* in comparison to other class probabilities of a frame) and the final label is the one class which "wins" the most. This method will be called *max count* here. The second approach is about finding the most representative frame in a collection of shot frames. The most representative frame is such an image with the highest class probability among all the other frames (e.g. the classifier is 99% sure that the frame depicts an angry face). Thus, the shot vector is the vector of such a frame and the shot label is the class with this

highest probability. We call this method *max pooling*. The last heuristic is to average the class probabilities of all frames and associate a shot with such an average probability vector. From now on this method is referred to as *averaging* in this paper.

2.4.3 BERT description

BERT's (Devlin et al., 2018) abbreviation stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformer. BERT is a deep learning model that belongs to the family of Transformer models. The Transformer was firstly introduced by Vaswani et al. (2017) and was reported to be a more computationally efficient model with a better performance for natural language processing tasks in comparison to recurrent and convolutional neural networks.

Typically, neural networks deal with textual information by generating vector representations for each word of the corpus vocabulary. The neural network part which deals with this task is called an *encoder*. The encoder in a classical configuration of an RNN works in left-to-right order, scans one word at a time and passes the weight associated with this word to the next word. This sequential forward scanning continues until the end of the input sequence. The weaknesses of such an architecture are that it uses *only* previous content for producing word representations and that its sequential nature has a big computational cost (Uszkoreit, 2017). The Transformer solves both of these problems by introducing the notion of *attention*. The Transformer executes a small number of steps per each input: when computing the representation of a word only its nearest neighbours (the words within one sentence) regardless of their order (the predecessors and successors of the word are involved) are taken into consideration.

The BERT model is an encoder. The contribution of BERT's encoding algorithm to deep learning is its bidirectional, contextual, and unsupervised nature. BERT is a completely unsupervised deep language representation that was trained on plain Wikipedia texts (Devlin and Chang, 2018). In addition, in its recent open source version BERT includes vector representations for 104 languages (including Dutch).

In order to make the required prediction (e.g. translate the sentence from English to French, answer the question), both RNNs and Transformer models use a *decoder*. The decoder takes the vector representations derived by the encoder and trains the model to make a desired prediction. The decoding step is not included into the BERT architecture. This system, however, can be fine-tuned to perform *any* natural language processing task given a small labeled corpus.

The representations created by encoders can be *context-free* or *contextual* (Devlin and Chang, 2018). Context-free representation models, such as word2vec (Mikolov et al., 2013), generate a single vector per each word in the corpus vocabulary. Hence, the context-free model will give the same vector representation of the word *"band"* as in *"a gold wedding band"* and as in *"a rock band"*. Contextual models like BERT give a word representation depending on the other words in a particular sentence (i.e. they need context to produce **and** to output a word embedding).

For the word *"band"* in the phrase *"I saw a cool rock band at Coachella"* a *unidirectional* contextual model will compute the embedding based on *I, saw, a, cool, rock*, but will not include *at, Coachella*. The bidirectional model like BERT will take both preceding and succeeding words into account starting from the very bottom of the neural network.

In order to compute word embeddings, the encoders are usually trained to predict the context of the word or the word given its context. In this task, the bidirectionality is hard to implement, as simple conditioning on the left and right context would lead the target word to accidentally "see" itself in the process of learning and this would lead to skewed embeddings. The BERT model overcomes this problem by masking several random words in the input. The training task of the model then is learning the masked words. The model uses each word of an input with conditions on its predecessors and successors in a sentence to solve this task.

To fine-tune BERT to a specific classification task, a labeled corpus should be provided. The embeddings learned from BERT are treated as features: depending on the nature of the

task, word or document embedding is fitted into a classifier. Since the BERT model produces embeddings on the word level, for the document-level fine-tuning tasks, such as emotion or sentiment classification, some modifications should be applied to go from the word-level embeddings to the document-level embeddings. Two libraries for fine-tuning BERT implement two different approaches. Native TensorFlow implementation of the fine-tuning step provided by the creators of BERT[10] uses the [CLS] embedding which represents the whole sequence and which is created at the beginning of each training example automatically (Devlin et al., 2018). The Flair[11] library which also utilizes BERT word embeddings pools the average value from the words of the document. This averaged vector is a representation of the document. Both in the native fine-tuning model and in the Flair model a simple combination of linear and softmax activation functions (analogous to logistic regression) on the feature vector produce the class probability vector. The size of the output can be adjusted depending on the number of classes required by the task.

The extra batch of 2,000 subtitles (described in Section 2.3) was used to fine-tune BERT on the emotion recognition task. In this batch, the neutral category turned out to be over-represented. Since class imbalance is known to affect many classification algorithms, including deep learning models (Buda et al., 2018; Wang et al., 2016), we downsampled the neutral category, so that the number of neutral samples is almost the same is the next biggest emotion class. Both native TensorFlow and Flair implementations for fine-tuning were used for this task and later their performance was compared on the S test set (the batch that contained the same shots in the V, S, and VS annotation tasks). The configuration of the Flair model that we used included the combination of BERT and FastText word embeddings being trained in a linear classifier on 34 epochs with 0.01 learning rate. As for the TensorFlow classifier the learning rate was 0.000001, it only utilized BERT embeddings, and it used 10 epochs to train.

2.4.4 Fusing mini-Xception and BERT

When the output of mini-Xception is adapted to the video-based emotion classification and when BERT is fine-tuned on the multiclass emotion recognition problem, the bimodal fusion can be made.

For the bimodal fusion two probability vectors each of length seven derived from the mini-Xception network on the V dataset and fine-tuned BERT on the S dataset were joined on the decision level for the final prediction. In order to produce one output label, several models and heuristical approaches were tested. The most simple ones include choosing a maximum probability out of two vectors (*maximum of the scores*); summing the corresponding probability values of two vectors and choosing the maximum value in the sum (*means of the scores*).

Weighted mean is the modification of the means of the scores technique. Before summing up the vectors, each of them is multiplied by the model weight. These two weights represent our trust in the models' outputs. They reflect the models' performance. In order to derive these weights we divide our annotated VS dataset into training (428 shots) and test (184 shots) sets. These two weights are derived by grid searching the optimal combination on the training set. The sum of the models' weights should always equal to 1.

In addition, two unimodal predictions, each of length seven, concatenated into one feature vector were inputted to the Support Vector Machines (SVM) and Random Forest classifiers. The reported performance metric is F1-score derived on cross-validated dataset.

2.5 Evaluation methods

The following part of the paper concerns the quantitative methods for assessing the performance of the machine learning models tested in this research (Section 2.5.1). In addition, we ran some

[10] https://github.com/google-research/bert/blob/master/multilingual.md#fine-tuning-example
[11] https://github.com/zalandoresearch/Flair

statistical tests to evaluate the difference between models or to estimate the effect of the emotion class on the classifier performance. Depending on the set-up, we divided the statistical tests into three groups. In Experiment 1 (Section 2.5.2) we compare the difference between the bimodal model performance and mini-Xception/BERT performance. In addition, this experiment concerns the comparison of two different fine-tuning architectures (Flair and native TensorFlow) for the textual emotion classifier. This experiment is also used to compare the performance of the max count heuristic applied to the mini-Xception output and the performance of a Random Forest classifier trained on the max count statistics for the facial emotion classification. Experiment 2 (Section 2.5.3) is dedicated to comparing the performance of mini-Xception on the soap opera data to its reported performance. Experiment 3 (Section 2.5.4) evaluates the effect of emotion class on the classifiers performance. Another use of Experiment 3 is to evaluate if there is any influence of the emotion class and the model type (facial or textual classifiers) when it comes to the agreement rate between human labels and unimodal predictions.

2.5.1 Metrics

The metric that is used in our research for evaluating the performance of machine learning classifiers is F1-score. This metric was chosen because it incorporates two basic machine learning performance metrics, precision (P) and recall (R), in a balanced way. F1-score is a harmonic mean of precision and recall (Sasaki et al., 2007):

$$F1 = \frac{2PR}{P+R}$$

Since the emotion recognition task in our research is multiclass, the overall model performance characteristic is derived from the metrics (F1-scores) calculated per each emotion category. The class specific F1-scores can be averaged in different ways for the overall estimation of the classifier performance. *Micro* average uses global numbers of true positives, false positive, and false negatives across all classes to calculate the metrics, for example, F1-score. *Macro* average calculates F1-scores specifically for each class and then averages the numbers for overall performance evaluation. *Weighted* average does the same as *macro*, but it uses weights that are proportional to the number of class objects [12]. Hence, *weighted* averaged can be biased to the largest category presented, especially, when there is class imbalance and the classifier overfits. Since F1-score incorporates the precision and recall scores and *macro* averaging does not favour any particular class, macro F1-score is used in our research as an assessment metric.

Besides the macro F1-score, the accuracy metric is used when we compare the performance of the state-of-the-art models (BERT and mini-Xception) reported in the source papers to their performance on the new data of the soap opera. Accuracy is the proportion of class instances that the model classified correctly to the total number of classification objects. Although the accuracy score has many downsides, for example, it does not provide class specific evaluation (Sokolova et al., 2006), this metric was reported as a performance metric for mini-Xception in (Arriaga et al., 2017) and for some NLP tasks on which BERT was fine-tuned (Devlin et al., 2018). Thus, in order to make a fair comparison of these models tested on our data, accuracy was used.

2.5.2 Experiment 1

In order to answer one of our research question *"Does combining two modalities improve shot-based emotion classification?"* we need to compare the bimodal model to the facial emotion classifier, and to the BERT-based textual emotion classifier. To conduct such a comparison, the following experimental set-up was established. Suppose that there are two sets of emotion

[12] https://scikit-learn.org/stable/modules/model_evaluation.html#multiclass-and-multilabel-classification

predictions derived from models A and B. Each of these prediction sets was divided into 10 parts. For each part macro F1-score was calculated. As a result, we have the F1-scores distributions for models A and B. These two distributions are checked with a t-test. The null hypothesis of the t-test is that the difference in samples means (F1-scores distributions of A and B) equals to zero (Field et al., 2012, Chapter 9). If the p-value is low (< 0.05), we can reject this hypothesis claiming that the performance of models A and B have a significant difference.

There are different scenarios where this experiments is used: comparing the performance of Flair and TensorFlow fine-tuning models on the S dataset and the bimodal - unimodal models comparison. When the human annotation labels that were used for calculating F1-scores are the same for each of the two models in question, the dependent t-test is used. The dependent t-test is used in the Flair-TensorFlow comparison and in the bimodal-unimodal comparisons when the VS dataset labels where used as ground truths for both bimodal and unimodal models. The independent t-test is used when the F1-scores are produced from different ground truth datasets. The independent t-test is conducted for mini-Xception (ground truth from the V dataset) - bimodal (ground truth from the VS dataset) comparison and for BERT (ground truth from the S dataset) - bimodal (ground truth from the VS dataset) comparison.

2.5.3 Experiment 2

When we compare the performance of mini-Xception on a new data source of the soap opera to the accuracy score reported by Arriaga et al. (2017), we used the one sample t-test (Field et al., 2012, Chapter 9). The one sample t-test compares the mean of our sample to the true mean that is expected to be found in a population. In this experiment we split the mini-Xception prediction into 10 batches and for each batch calculate the difference in accuracy scores between our predictions and the reported prediction of 66%.

We set the expected population difference to 0. Hence, if there is a certain level of confidence ($p < 0.05$) in the one sample t-test, we can accept a two-tailed alternative hypothesis which is the true mean of score difference is not 0. If this assumption turns out to be correct, then we can say that the difference between the performance of mini-Xception on our data and reported performance is significant.

2.5.4 Experiment 3

When comparing the ability of a particular model to classify specific emotion classes, the combination of one-way ANOVA and Turkey Honest Significant Difference test (Turkey's test) is conducted. In order to prove that there is a *significant* difference in the ability of the models to distinguish emotion classes, we set up the following statistical experiment. The model predictions are divided into ten batches, then we calculate class specific F1-scores for each batch. Hence, we have seven emotion specific F1-score distributions. One-way ANOVA can test whether the means of the emotion specific F1-score distributions are different and how significant this difference is (Field et al., 2012, Chapter 10). The F-ratio reported by ANOVA can answer this question. F-ratio compares the systematic variance of the model to its unsystematic variance (error). In this paper we report F-ratio and the p-value. The p-value in this test indicates the probability of an observed effect to be reported when in reality there is no effect.

The one-way ANOVA only indicates whether there is an effect of the independent variable (emotion class) on the dependent variable (F1-score). To investigate *how* the performance of a classifier differs depending on emotion, i.e. which emotion is classified better or worse, a post-hoc analysis should be conducted. The Turkey's test is such a post-hoc analysis. The Turkey's test reports the difference in means and the level of significance of this difference for all possible pairwise combinations of emotions. The pairs that have a significant ($p < 0.05$) difference are reported in this research together with the difference value and p-value.

Besides the emotion specific classification analysis, one-way ANOVA in combination with the Turkey's test is used to investigate whether there is any effect of 1) emotion classes and 2) the model type (facial or textual) to the human-model agreement rate.

To sum up, the metric that is used for assessing the performance of our models is F1-score (macro averaged for the overall assessment). For comparing mini-Xception performance on a new data source to the performance reported by Arriaga et al. (2017), accuracy is calculated. The difference in models performance is tested using different types of t-tests (Experiment 1 and Experiment 2). As for the inference on emotion specific performance, the combination of one-way ANOVA and the post-hoc Turkey' test is used (Experiment 3). The same combination of one-wat ANOVA and the Turkey's test is utilized for assessing the effect of emotions and modalities on human-model agreement rate. In all the statistical experiments the confidence level of 95% was used. Hence, the significance value for rejecting the null-hypothesis is $p < 0.05$.

When it comes to reporting the models performance, usually[13] macro F1-scores derived from the *whole* test set are reported. When for some statistical tests we divide the dataset into ten batches, the average F1-score (M) from these batches is reported together with the standard deviation (SD).

3 Results

In this section, the results of our research are provided. The distribution of the labels assigned by humans in three annotation tasks is given in Section 3.1. The performance of the mini-Xception network on the soap opera video data is discussed in Section 3.2. The evaluation of the textual emotion classifier fine-tuned on BERT embeddings is presented in Section 3.3. Finally, different approaches for bimodal fusion and their performance are discussed in Section 3.4.

3.1 Annotation results

Three different groups of people participated in three labelling task. Five people annotated the V dataset, 20 people labelled the S dataset, 4 people - the VS dataset. Such an uneven distribution of human resources was chosen because of the amount of subtitle data that needed to be annotated: three times more objects than in the other two tasks. The overview of the annotators contribution is provided in Table 1. As people were instructed to annotate as many objects as possible and they could take breaks in the process, the contribution of each annotator for three datasets varied drastically. The standard deviation in column 3 illustrated this variation: while people annotated, for example, 168 subtitles on average, some of them did three or four times more than this number. One person annotated 910 subtitles which is 27% of the S dataset.

Task	**# people**	**Mean (*SD*) # labels per person**	**Largest contr. (labels)**	**# objects**	**# clean objects**
video	5	201.4 (132.9)	361	1007	840
subs	20	168.0 (206.9)	910	3360	2,675
video + subs	4	253.3 (123.2)	429	1013	831

Table 1: Results of three annotation tasks (video, subtitles, video + subtitles).

The shots which were annotated as having multiple or no faces, plus shots and subtitles with multiple emotions were deleted from the datasets. Table 2 shows that the most common noise

[13]Except for the cases where machine learning classifiers (Random Forest or SVM) were trained on our data. Then the average macro F1-score from five-fold cross-validation is reported.

type in V and S annotation tasks is more than one emotion expressed. The combination of these two channels helped to resolve this ambiguity which led to the drop of this noise reported in the VS dataset. In the S dataset, the amount of subtitle with multiple speakers presented is eight times more that the number of subtitles which annotators labelled as having multiple emotions. Only 27 subtitles with multiple speakers were annotated as having multiple emotions in the S annotation task.

Task	**Noise type**						
	>1 face	**No face**	**No video**	**>1 emotions**	**No sub**	**Shuffled text**	**>1 people talking**
video	3	0	21	84	-	-	-
subs	-	-	-	85	15	36	686
video + subs	8	0	38	27	1	-	15

Table 2: The number of different noise types across the annotation tasks.

After dropping noisy data, there were 840 labelled videos, 2,675 labelled subtitles, and 831 labelled combinations of video and subtitle left. Since the noisy shots reported in these sets were not completely identical, the intersection of these cleaned datasets resulted in 636 common V, S, VS objects (see Figure 6). These 636 labelled objects were to be used to test mini-Xception and BERT models and built and test a bimodal classifier. The $2,675 - 636 = 2,039$ subtitles were used for fine-tuning BERT.

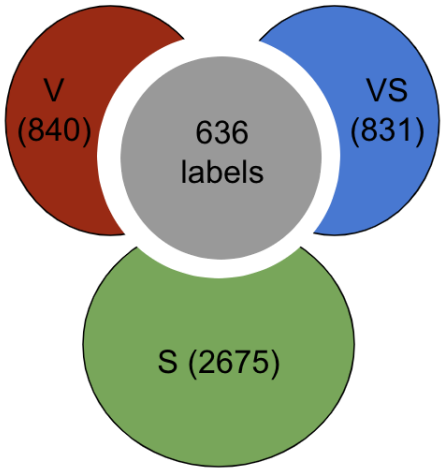
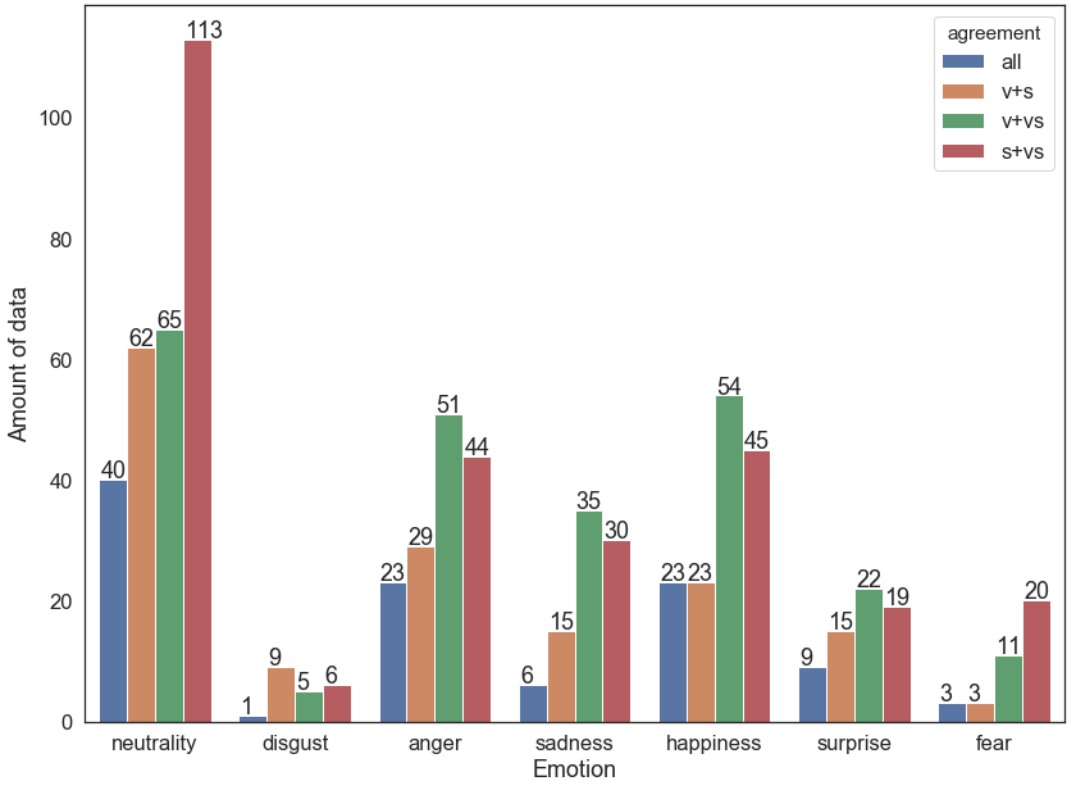


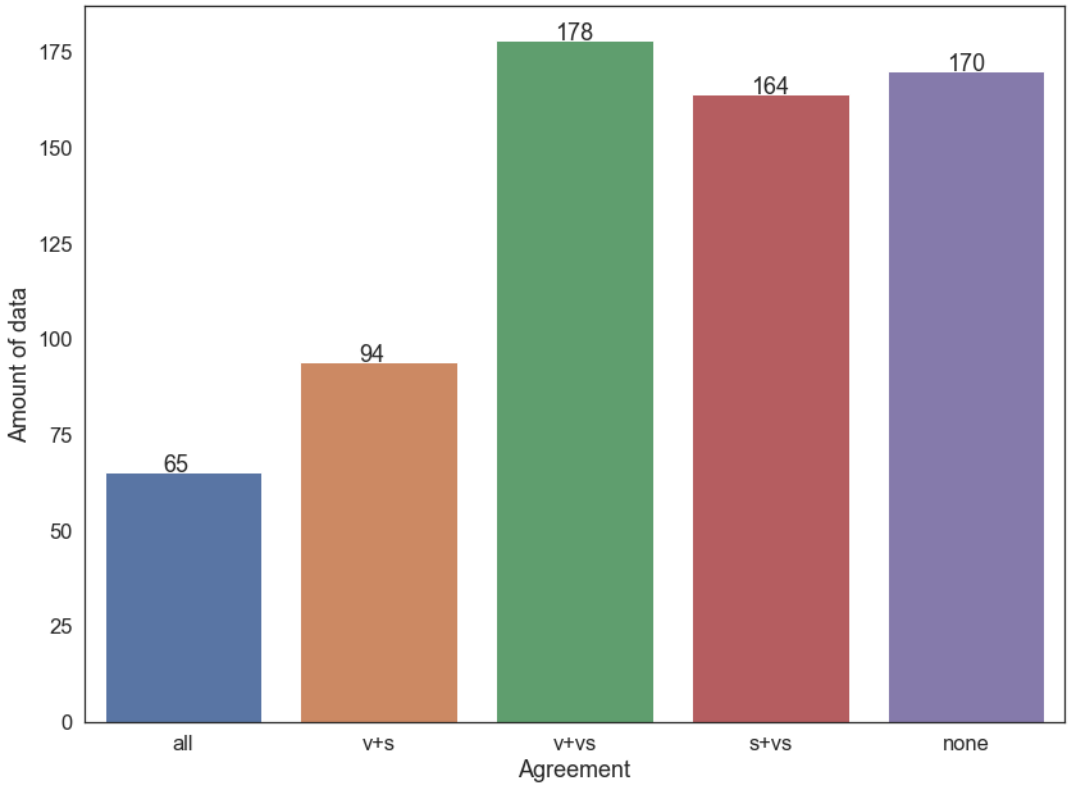
Figure 6: Number of non-noisy shots in three annotation tasks.

In Figure 7a the agreement rate across seven emotion classes and three datasets which were used for testing the algorithms is shown. The largest agreement rate belongs to the neutral category (especially between S and VS datasets). As for the non-neutral emotion states, strong emotions such as anger, sadness, and happiness are more agreed upon in comparison to disgust, surprise, and fear. This trend can be caused by the fact that the latter emotions are quite rare in our dataset in comparison to sadness, happiness, and anger. In addition, disgust, surprise, and fear may be less recognizable for participants especially on the video (facial) data. The V dataset, however, has a big contribution at the agreement rate for anger, sadness, and happiness (see V + VS agreement rates for these emotions).

If we eliminate the neutral category (Figure 7b) 27% of the test set have no agreement between V, S, and VS annotation tasks. This number shows the complexity of the task that annotators were given.



(a) Agreement between three annotation tasks across different emotion classes (with agreement tasks specified).



(b) Label agreement between three annotation tasks (neutral state excluded).

Figure 7: Overview of the annotators' agreement.

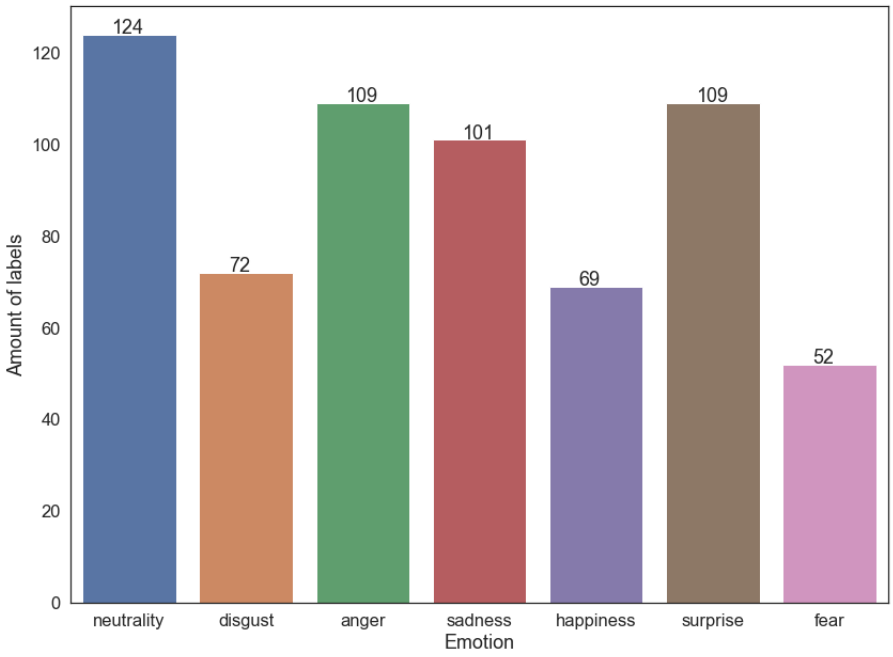


Figure 8: Label distribution in cleaned video dataset.

3.2 Mini-Xception performance

As a result of the video annotation task, the V dataset (636 labels) was created. The label distribution of the V dataset in Figure 8 shows that the emotion class distribution in our dataset is uneven.

The labelled shots from the V dataset were divided into frames and inputted into the mini-Xception pipeline. The face detection step which was done by the Viola-Jones algorithm failed to register any faces for 24 shots out of 636. Since in these shots no emotions could be registered, the performance of the network there could not be directly compared to the BERT-based classifier. In addition, there were no probability vectors to represent visual modality in this 24 shots for the bimodal model. Hence, we excluded these 24 shots from testing the performance of mini-Xception, BERT-based emotion classifier, and bimodal models. The size of the test set shrank from 636 to 612 shots.

As mentioned in Section 2.4.2, in order to turn the frame-based predictions of mini-Xception to one video-based label, we applied three different heuristics for capturing temporal variations: max count, max pooling, and averaging.

As discussed in Section 2.5.1, the main metric that we use for assessing the performance of the emotion classifiers is F1-score. The averaging that is used here for going from class specific performance to the overall model performance is *macro F1-score*. All three heuristical approaches for retrieving shot labels from mini-Xception have macro F1-score of 0.17 on the V dataset.

In addition to heuristical approaches for capturing temporal variation in the frames, we trained a Random Forest classifier on max count statistics (where each shot label is represented by a vector with the counts of the winning classes). Random Forest cross-validated across five folds gave almost the same result as after merely applying the heuristics. Macro average F1-score across five folds was 0.19. A dependent t-test showed that there is no significant difference in Random Forest trained on the max count heuristic output and max count predictions ($t(9) =$

$0.69, p = 0.5$). The test result demonstrates that there is no additional signal in preprocessed mini-Xception predictions.

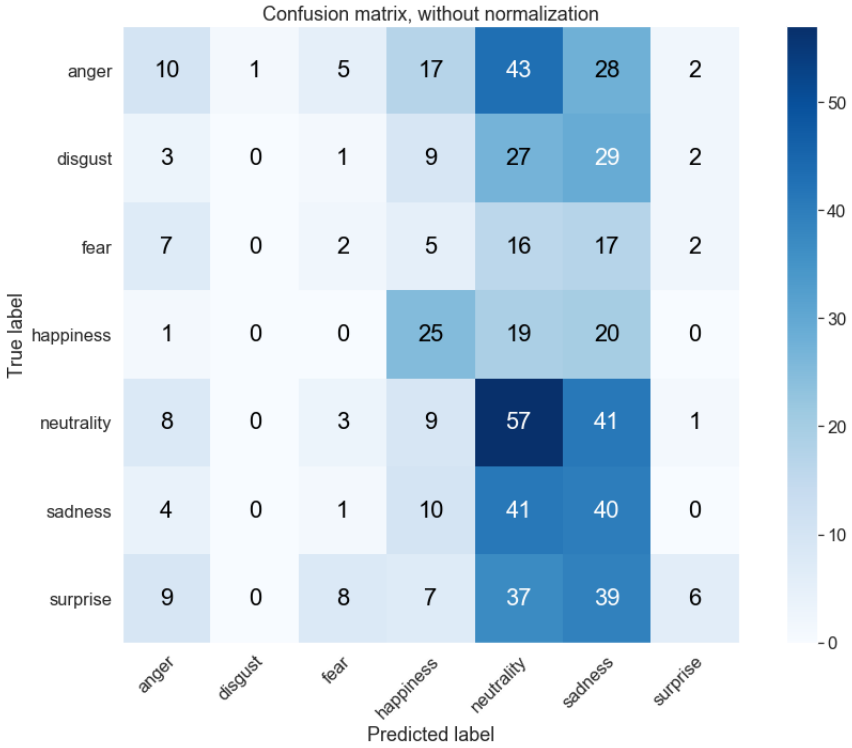


Figure 9: Confusion matrix for the max count heuristic.

Since the macro F1-score of all three heuristical approaches is the same, we randomly chose one of them, max count, to look into the details of mini-Xception classification. In order to assess how well mini-Xception classified different emotions, first, we need to prove that the difference between emotions in the model performance is significant. The one-way ANOVA (the detailed explanation of this step is provided in Section 2.5.4) was conducted to compare the effect of emotion class on the model performance (F1-score). There turned out to be a significant effect of emotion on the performance for seven emotion classes [$F(6, 63) = 16.69, p < 0.001$)]. In addition, post-hoc analysis using the Turkey test was carried out. In Table 3 the results of this analysis is summarized. The class 'happiness' turned out to have a significantly ($p < 0.05$) higher mean of F1-scores distribution in comparison neutrality, disgust, anger, and fear.

classes	**difference**	***p*-value**
happiness-neutrality	0.19	0.002
happiness-disgust	0.32	$p < 0.001$
happiness-anger	0.27	$p < 0.001$
happiness-fear	0.21	$p < 0.001$

Table 3: Part of the Turkey HSD test related to the happiness label.

In Figure 9 the confusion between human and predicted annotation of the shots is plotted. 88% (536/612) of the shots were assigned with 'happiness', 'neutrality', or 'sadness' labels. Happiness had one of the lowest number of labels in our dataset. This class, however, had a significantly higher performance than four other classes (neutrality, disgust, anger, and fear). If we take a look at the emotion distribution of the FER-2013 dataset (Connie et al., 2017) on which mini-Xception was trained (see Table 4), we see that happiness, neutrality, and sadness have the biggest representation in this dataset. Although Arriaga et al. (2017) managed to

combat overfitting and get 66% of accuracy across 7 classes, the network does not generalize well on the new dataset and still reflects the distribution of the data on which it was trained.

happiness	**neutral**	**sadness**	**fear**	**anger**	**surprise**	**disgust**
8989	6198	6077	5121	4953	4002	547

Table 4: Distribution of the emotion labels of the FER-2013 dataset.

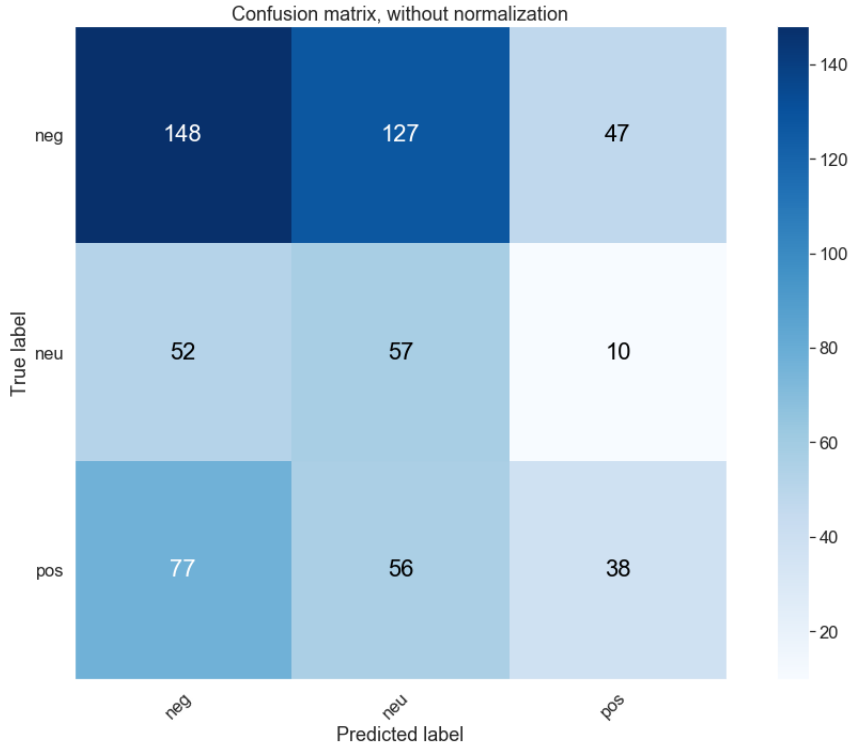


Figure 10: Confusion matrix for the sentiment classification.

Given the class imbalance of the dataset that was used for training the network, the complexity of the task for the annotators (all the annotators of the V dataset admitted that it was quite a formidable task for them to assign an emotion label from a video cue with muted sound), and the noisy nature of the shots, we decided to evaluate network's performance on a simpler classification task. In order to do so, we changed the experimental set-up from fine-grained emotion classification to assessing the polarity of the facial emotion (whether it is a positive, a negative, or a neutral facial expression). Seven emotion classes were mapped to these three polarities. The mapping has been conducted according to the research of Yuvaraj et al. (2014) who plotted categorical emotions on the valence-arousal dimensional scheme, an approach to emotion description developed by Russell (1980). According to Yuvaraj et al. (2014), surprise and happiness get positive valence, disgust, sadness, anger, and fear are matched with negative valence, and neutrality remains neutral.

Mapping the results of max count heuristic to the ternary sentiment labels gives the following results. Macro average F1-score 0.37. A one-way ANOVA demonstrated that there is a significant difference in F1-scores between the sentiment classes $[F(2, 27) = 14.06, p < 0.001]$. The post-hoc Turkey test showed that there is a significant difference between the performance of the model on the negative and neutral class ($p = 0.001$). The negative class on average had a better F1-score by 0.18 points. The negative class also had a significantly better performance in comparison to the positive class ($p < 0.001$). The difference between F1-scores there was on average 0.22.

When it comes to comparing the performance of mini-Xception on a new soap opera data source to the reported performance (66% accuracy), the one sample t-test (described in Section 2.5.3) was conducted. The difference between the accuracy scores derived from our data and the reported accuracy score turned out to be significantly different from 0. On average the reported performance had 51.3 ($SD = 4.3$) higher accuracy than accuracy derived from the soap opera data. The accuracy of mini-Xception on the V dataset was 23%

To sum up, the performance of the network on a new data source is not nearly as good as it was reported in the source paper (Arriaga et al., 2017). The class imbalance in our test dataset, differences in the data that the network was trained on and our test data were an obstacle for the network performance.

3.3 BERT performance

After saving 612 subtitles[14] for training and testing the bimodal emotion recognition model and BERT model, there were 2,039 subtitles left for fine-tuning BERT. Among these 2,039 subtitles 813 belonged to the neutral category. We downsampled the neutral category, so that the number of neutral samples is almost the same is the next biggest emotion class. The class distribution for fine-tuning BERT is provided in Figure 12. There is still some class imbalance in the training dataset: surprise, fear, and disgust are underrepresented. The reason why we did not reduce the amount of neutral, happy, and angry subtitles to the category that had the lowest amount of data (disgust) is that the size of the training dataset might not be sufficient for fine-tuning BERT. The dataset with the lowest amount of training samples that was presented in the BERT original paper had 2,500 objects. Hence, we are using the training dataset of 1,549 subtitles and we note that class imbalance present in our data is a limitation.

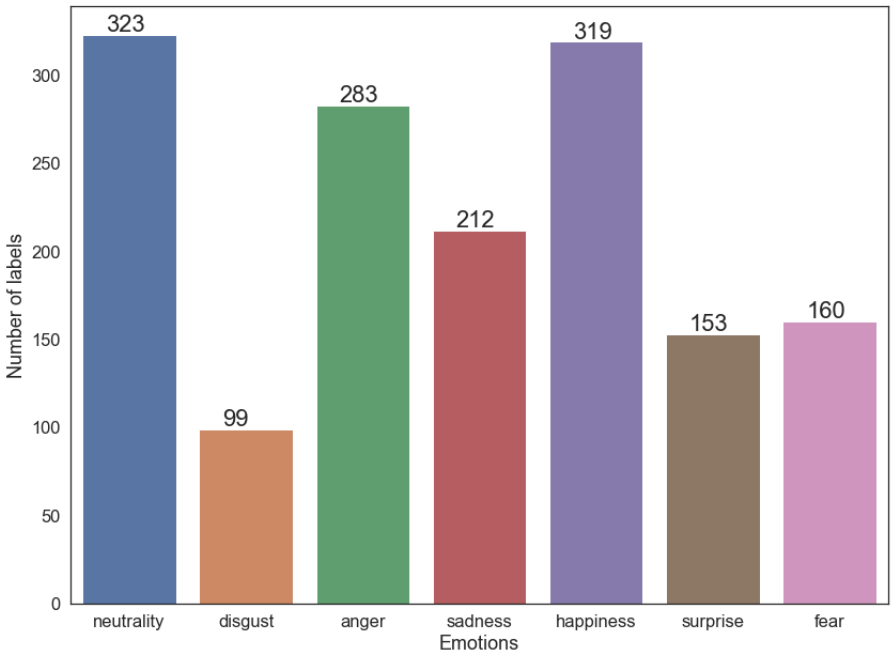


Figure 11: Class distribution for training BERT.

When the models were fine-tuned, the performance of the Flair and the TensorFlow classifiers

[14]We deleted 24 subtitles associated with 24 shots where the visual emotion classifier did not register any faces from the original 636 test set.

were later tested on the S test set with 612 subtitles. The macro F1-score of the Flair model was 0.25 on the S dataset, and the macro F1-score of the native TensorFlow model was 0.26.

When the test set was split into ten batches according to the set-up of Experiment 1, the mean F1-score of the Flair classifier turned to 0.23 ($SD = 0.07$) and the mean F1-score of the TensorFlow classifier became 0.25 ($SD = 0.05$) The difference in models performance was measured with the dependent t-test (the procedure is described in Experiment 1 from Section 2.5.2). The dependent t-test showed that there is no significant difference between the models in terms of performance ($t(9) = -0.45, p = 0.67$)

We chose the native TensorFlow implementation of fine-tuning to take a closer look at emotion specific performance of this model[15]. In order to infer if there is a significant difference in emotion specific performance of the classifier, a one-way ANOVA was conducted on the prediction of the native TensorFlow model. This test showed that the type of emotion had a significant effect on the classifier performance [$F(6, 63) = 16.21, p < 0.001$)]. The post-hoc Turkey test (Table 5) showed that the performance on the disgust label is significantly ($p < 0.05$) worse than on happiness, neutrality, surprise, sadness, and fear. The performance on the anger class is significantly worse than on surprise, fear, sadness, happiness.

classes	**difference**	***p*-value**
neutrality-disgust	0.15	0.09
happiness-neutrality	0.26	< 0.001
sadness-disgust	0.31	< 0.001
happiness-disgust	0.41	< 0.001
surprise-disgust	0.29	< 0.001
fear-disgust	0.27	< 0.001
sadness-anger	0.30	< 0.001
happiness-anger	0.40	< 0.001
surprise-anger	0.27	< 0.001
fear-anger	0.26	< 0.001

Table 5: Part of the Turkey HSD test with the significant differences in emotion pairs.

Since the performance of BERT-based classifiers turned out to be quite moderate on the multiclass emotion classification problem, we trained the TensorFlow fine-tuning model on the downgraded sentiment classification problem. The set-up that was used is adapted from the mini-Xception result section 3.2. The assumption behind this experiment is the same as there: does fine-tuning BERT on a simpler classification task improve the performance of the model?

The macro average F1-score of this ternary classification is 0.42. A one-way ANOVA proved that there is a significant difference in the ability of our classifier to classify different sentiments [$F(2, 27) = 14.64, p < 0.001$)]. In addition, according to the Turkey's test, the negative class is classified significantly better than the neutral class (difference in mean F1-scores = 0.18, $p < 0.001$) and negative emotions were better classified than positive ones (difference in mean F1-scores = 0.21, $p < 0.001$).

The negative class had significantly better performance than neutral and positive classes, which can indicate that the subtitles associated with anger, fear, disgust, and sadness have more distinctive vocabulary than the other two classes.

Even though there is no NLP task presented in BERT's source paper (Devlin et al., 2018) with the same classification problem as in our research, an indirect comparison of one of BERT's reported performance and our ternary sentiment classification can be made. BERT creators fine-tuned the English embeddings on the binary set-up of the sentiment analysis task and reported 94.9% accuracy which is much higher than the results that we obtained on the ternary sentiment analysis task which is 45%.

[15]The classification report of this model derived from the test set S is provided in Table 8.

The performance of the algorithm when fine-tuned on BERT embeddings turned out to be low on a multiclass problem and non-English data. The TensorFlow implementation gained 0.26 F1-score. The Flair model (0.25 F1-score) which incorporated FastText embeddings to the BERT embeddings turned out to have no significant difference with the native TensorFlow implementation. The ternary sentiment analysis task turned out to have 0.42 macro F1-score and 45% accuracy. The low performance can be explained by the quality of the labelled dataset and the size of the training set (1,549 samples is still quite a small amount in comparison to the data sources reported in the BERT paper (Devlin et al., 2018)).

3.4 Joining the modalities

Since we tried different heuristics for capturing temporal changes in the mini-Xception network output and different configurations for fine-tuning classifiers on BERT embeddings, we chose two of these configurations for bimodal fusion. As far as the visual modality is concerned, the vector with the maximal value (max pooling) is chosen to represent this modality. As for the textual data, a subtitle vector of one shot is represented by the class probability distribution of the native TensorFlow fine-tuning classifier. These two vectors each of length seven were stacked together for a bimodal prediction.

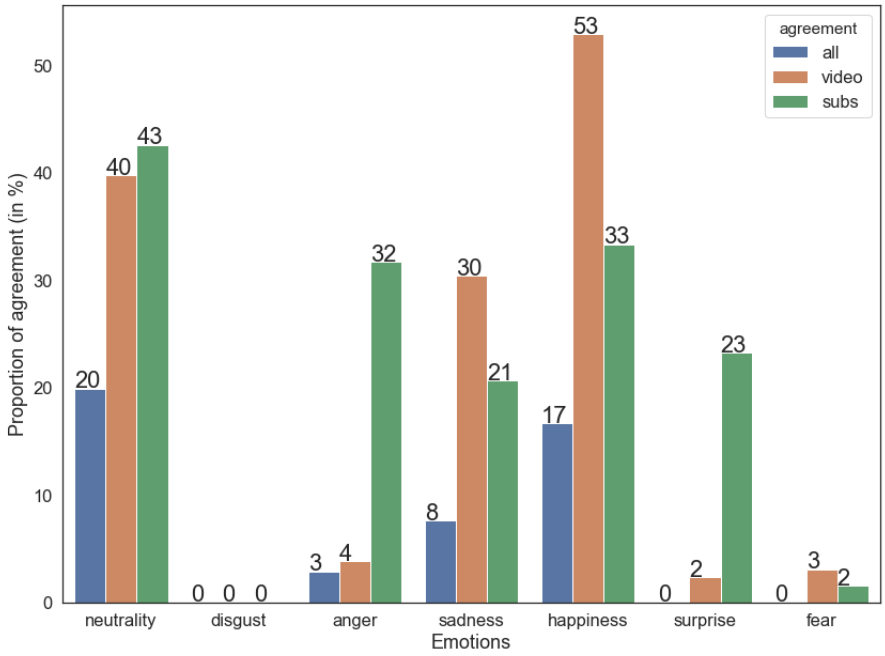


Figure 12: Agreement between predictions of mini-Xception (*video*), BERT-based model (*sub*) and the human VS annotations. *All* bar shows when the human annotation, video, and subtitle predictions coincide.

First, let's overview how the predicted values of BERT classifier and mini-Xception network agree with human annotation in the VS dataset. In Figure 12, the agreement rate between the predicted labels and human annotation is depicted. Overall, for more than half of the shots (343 out of 612) there is neither agreement between predicted V and VS datasets, nor predicted S and VS. As for the shots where there is some agreement, a one-way ANOVA showed that the emotion class $[F(6, 14) = 4.03, p = 0.01)]$ has a significant effect on the agreement rate. As far as the type of the modality (facial or textual classifier) is concerned, the effect is not significant

$[F(2, 18) = 1.69, p = 0.21)]$. Running the Turkey's test on the emotion class showed that there is only one pair of emotions that had a significant difference in the agreement rate. Happiness is agreed upon more (p = 0.048) than disgust (the difference in the agreement rate is 34%).

As far as the bimodal classification is concerned, the three heuristical approaches, such as the average value of two modality vectors, choosing the maximal value, and weighting the outputs of two models (weighted mean), show the following macro averaged F1-scores: 0.18, 0.18, 0.19 respectively.

As for the weighted mean heuristic, we split the 612 set into train (428) and test (184) datasets and looked for the best combination of weights on the training set. The macro F1-score reported above was derived from the test set of 184 shots. As for the other two heuristical approaches, the macro F1-score reported is from the whole VS dataset.

The combination of scores in the weighted mean heuristic that gave the highest performance on the training set is 0.3 × visual modality + 0.7 × textual modality. We argue that the weight values do not show that the subtitle vector is in larger agreement with the human annotation, as the majority of shots do not have any agreement with the mini-Xception and BERT predictions and the subtitle prediction does not have an overwhelming number of agreement with the human annotation in comparison to the visual modality. The high weight of the subtitle prediction is rather explained by the distribution of probabilities in mini-Xception vectors (Figure 13) and in BERT vectors (Figure 14).

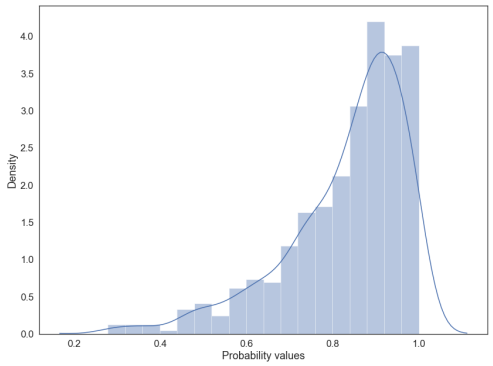


Figure 13: The distribution (density) of the maximal probability values in mini-Xception predictions.

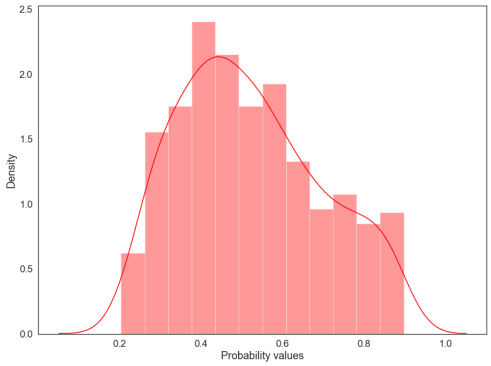


Figure 14: The distribution (density) of the maximal probability values in BERT-based subtitle emotion classifier.

The class probabilities that mini-Xception outputs are usually much bigger than the probabilities of the BERT-based classifier. For example, when using the maximal score heuristic the value from the visual modality vector is chosen in 563 cases. Hence, the weighted mean heuristic tries to balance the probability distribution by "punishing" the visual modality vector with the lower score.

Besides applying the heuristical approaches, we tried to train Support Vector Machines and Random Forest classifiers on the joined class probability vectors. The result are as follows: 0.15 and 0.22 F1-score averaged on five folds of cross-validation. Training a dummy classifier that outputs a label in accordance with a class distribution on the training set of 428 shots gives 0.15 macro F1-score. Thus, the performance of SVM is not better than an informed random guess.

We compared the performance of one of the bimodal fusion strategies on the VS dataset to the performance of the facial and textual emotion classifiers, which were used for the bimodal model, on the V and S datasets respectively. This comparison was conducted in accordance with Experiment 1 (Section 2.5.2). The chosen bimodal model (Random Forest) has performance of 0.19 ($SD = 0.05$) macro F1-score on 10-fold cross-validation. The native TensorFlow

implementation of BERT fine-tuning model gave 0.25 (SD = 0.05) macro F1-score on the S set split into ten parts, and applying max pooling heuristic on the mini-Xception output has 0.17 ($SD = 0.06$) macro F1-score. The bimodal fusion slightly outperformed the facial emotion classifier, but it did slightly worse than the textual emotion classifier. Running the independent t-test to compare bimodal and mini-Xception models showed that there is no significant difference in these models performance ($t(18) = 1.10, p = 0.29$). As for the BERT-bimodal comparison, the independent t-test demonstrated that the difference in performance reported above is significant ($t(19) = -2.20, p = 0.04$). Bimodal system did not outperform both unimodal systems which contradicts our initial hypothesis that joining the modalities would boost the performance of the automatic emotion recognition system.

Since the comparison reported above included the metrics that were derived on three different datasets (V, S, and VS) that represented the ground truth, this analysis can seem as flawed. In three labelling tasks people were exposed to different types or combination of modalities which, on the one hand, made the human judgments as close to the judgment of the corresponding classifiers, but, on the other hand, lack of information in video labelling and subtitle labelling tasks might have produced low quality emotion annotation that was subsequently treated as the ground truth in the assessment of the unimodal emotion classifiers. People who were assigned to label the combination of video shots and subtitles were exposed to the largest amount of information in our experiments. Thus, the labels that were produced for the VS dataset should have better represented the emotions in soap opera fragments in comparison to the other two datasets. Consequently, we cannot claim whether the level of significance reported by the independent t-test purely characterizes the models performance. A more rigorous interpretation of, for example, the bimodal-BERT comparison is that the combination of labels created in the S annotation task and the BERT predictions is significantly different from the combination of VS dataset labels and bimodal predictions.

Due to the reasons mentioned above, we investigated how all three models performed on the VS dataset. The bimodal model (Random Forest) has performance of 0.19 ($SD = 0.05$) macro F1-score on 10 fold cross-validation. The native implementation of BERT fine-tuning model gave on average 0.20 ($SD = 0.04$) macro F1-score, and applying max pooling heuristic on the mini-Xception output has on average 0.16 ($SD = 0.04$) macro F1-score across ten batches. The bimodal fusion slightly outperformed mini-Xception and had worse performance than BERT. Two dependent t-tests showed that the differences in bimodal and mini-Xception models scores ($t(17) = 1.50, p = 0.15$) and in bimodal and BERT scores ($t(15) = -0.37, p = 0.72$) are not significant.

This approach to answer our research question has its own downsides. Even though we compared the models on *one* set of ground truth labels that should have been more accurate than the other two sets, this approach can make the performance of the bimodal model better than it is in reality. For example, the BERT-based emotion classifier was trained on the subtitle dataset where the annotators based their judgment purely on the textual channel which might have resulted in a lower performance on the VS dataset (average 0.20 F1-score on the VS dataset in comparison to average 0.25 F1-score on the S dataset).

To sum up, to answer our research question *"Will the model that joins textual and visual modalities perform better than unimodal models on the task of emotion classification"* we compared the F1-scores of facial, textual, and bimodal emotion recognition models derived from their own datasets and from the VS dataset. In the first set-up BERT-based textual classifier (in combination with the labels created in the S annotation task) turned out to have a significantly higher performance than the bimodal model. The bimodal model in turn on average had a higher performance in comparison to mini-Xception, but this difference is insignificant. When testing all models on VS dataset, the BERT system also outperformed the bimodal model and mini-Xception, but the difference between mini-Xception-bimodal and BERT-bimodal performance turned out to be not significant.

The possible explanation of such a small difference in unimodal and bimodal metrics lies in the nature of labels that were collected. Extracting emotion labels from a soap opera with an audio cue muted turned out to be a difficult and ambiguous task for the human annotators. 170 shots out of 636 turned out to have *no* agreement in human labels assigned in V, S, and VS annotation tasks. In addition, the predictions that were made by facial and textual emotion models did not correspond well (no agreement in 343 out of 612 shots) with the VS annotation tasks which is logical to expect when even the human labels do not have much agreement.

4 Discussion

The goal of our research was to answer the following research questions:

1. *What is the performance of the state-of-the-art facial emotion recognition model (the mini-Xception network) and the state-of-the-art NLP model (BERT) on the emotion recognition task on a new data source of soap opera?*
2. *Does a model that combines textual data and facial expressions classify emotions better than unimodal (textual and facial) emotion classifiers?*

In order to answer these research questions we asked RTL employees to annotate three datasets with video shots of the soap opera (V), subtitles (S), and the combination of those two (VS). The performance of mini-Xception, BERT-based textual classifier were tested on V and S datasets respectively. The performance of mini-Xception in the shot-based emotion prediction task was 0.17 F1-scores for all the heuristical approaches for capturing temporal variation tested (max count, max pooling, averaging). The accuracy score of mini-Xception was 23%. The difference between our accuracy and reported accuracy (66%) turned out to be significant. Thus, we can claim that the performance of mini-Xception on the new soap opera data source is significantly worse than Arriaga et al. (2017) reported when they introduced this model. As far as the BERT-based classifier is concerned, the performance of the native TensorFlow implementation of the fine-tuning step gave 0.26 F1-score on the S test set. Fusing these two unimodal models for the bimodal prediction did not significantly improve the classification quality: Random Forest performance on the whole VS dataset was 0.22 macro F1-score.

In the following subsections we discuss the results of the mini-Xception performance (Section 4.1), BERT performance (Section 4.2), and the bimodal fusion performance (Section 4.3) in more detail. The limitations of our research and future directions are provided in Sections 4.4 and 4.5.

4.1 Reflecting on the mini-Xception performance

As summarized above, the performance of mini-Xception turned out to be less impressive than it was reported in the original paper (Arriaga et al., 2017). Arriaga et al. (2017) reported 66% average accuracy, while on our data the accuracy was only 23%. Such a huge difference in performance resulted from the difference in the data sources. Originally the network was trained on the static image data. We had to apply some heuristics to transform static frame-based predictions into shot-based predictions. More importantly, the images that were used for training the classifier were retrieved by querying emotion words in Google (Goodfellow et al., 2013) and, hence, the representations that the network learned were quite distinct and prototypical instantiantions of emotions. Actors acting in the dynamic settings of a soap opera did not necessarily include these prototypical realizations of basic emotions.

In Figure 15, the confusion matrix of emotion classification is presented with some examples of the actors faces that the network analyzed. The faces that are represented in the plot were chosen from such frames where the probability of the class predicted by mini-Xception was the

highest. The class predictions from max count heuristic were used here. From this figure, it can be concluded that the Viola-Jones algorithm produced some false positives, which marked that the part of the frame that was detected as a face was not a face (see the fragment 1).

As the network was trained on prototypical instantiations of emotions, the network learned prototypical portrayals of emotions. The part of the plot with the true label *'sadness'* (line 2) depicts a man whose emotions were misclassified with happiness and neutrality. The male actor does not exhibit prototypical emotion markers associated with sadness, such as lowered corners of the mouth. His rheumy eyes rather than mouth or eyebrows portray the sad expression.

In addition, the dynamic nature of the data provided to the network contributed to misclassifications. Since actors were talking through the shots, some frames depicted actors in the middle of an utterance, which involved open mouths, smiles, associated with some emotions such as surprise and happiness (see 3 and 4).

Furthermore, the quality of acting in this soap opera might have its own influence on the performance of the network. There were a lot of shots where the emotion could only be derived by the subtle expression in the eyes and there were no facial movements. This trend was present in women's faces in particular: compare the movement of the eyebrows of the male and female actors. Actresses, such as 5 and 6, just slightly moved the eyebrows or did not moved them at all. A possible reason for that is heavy camera make-up or beauty procedures applied for combating wrinkles.

Emotion-specific performance of mini-Xception on our data still reflects the initial label distribution of the model's training set. 88% of the labels that mini-Xception outputted for our data are 'happiness', 'neutrality', or 'sadness'. These are three classes with the largest amount of data represented in the FER-2013 dataset on which mini-Xception was trained (see Table 4. This observation together with the low performance of the neural network leads to the conclusion that mini-Xception does not generalize well when it comes to predictions on a different data source. This conclusion supports the findings of Avots et al. (2018) who conducted cross-dataset comparison of deep learning facial emotions classifiers and received almost random results on the new data source.

4.2 Reflecting on the BERT performance

When it comes to the textual emotion model, using the TensorFlow implementation for training our customized emotion classifier on top of BERT embeddings also gave a moderate result. Macro F1-score on the subtitle test set was 0.26. Since in the source paper of BERT (Devlin et al., 2018) the embeddings were fine-tuned on the English data and they were not tested on the task of emotion classification, we cannot directly compare the performance of our classifier to the original paper. Nevertheless, given that BERT was reported to beat every state-of-the-art model in eight different language understanding tasks, such as sentiment analysis, next sentence prediction, recognizing textual entailment, the performance of our classifier is low (average score for different NLP tasks was 82.1 in (Devlin et al., 2018)).

One of the possible reasons of the low performance of our BERT-based classifiers is that the subtitles that were picked for training and testing usually did not have strong opinion words, i.e. compare "I am so *angry* at you" and "Give me back my book!".

Table 6 provides a closer insight into the human annotation and model classification.

#	subtitle	translate	human label	BERT label
1	Tot de dood ons scheidt, dat duurt niet lang, he.	Until death do us part, it won't take long, hey.	sadness	anger
2	Als dat echt zo was dan had je me dit nooit geflikt.	If that were true then you would never have done this to me.	anger	sadness
3	En ze denkt dat ik het nog moeilijker krijg... als ik met hem trouw en hij straks dood is.	And she thinks that it will be even harder for me... if I marry him and he dies.	fear	sadness
4	Ze is even goed mijn dochter als de jouwe.	She is as much my daughter as yours.	anger	neutrality
5	Wij gaan even een biertje drinken.	We are going to drink a beer.	happiness	neutrality
6	Wat kan er allemaal gebeuren?	What can happen?	fear	anger
7	Wat heb jij nou gedaan?	What did you do?	surprise	anger

Table 6: Examples from the subtitle test dataset.

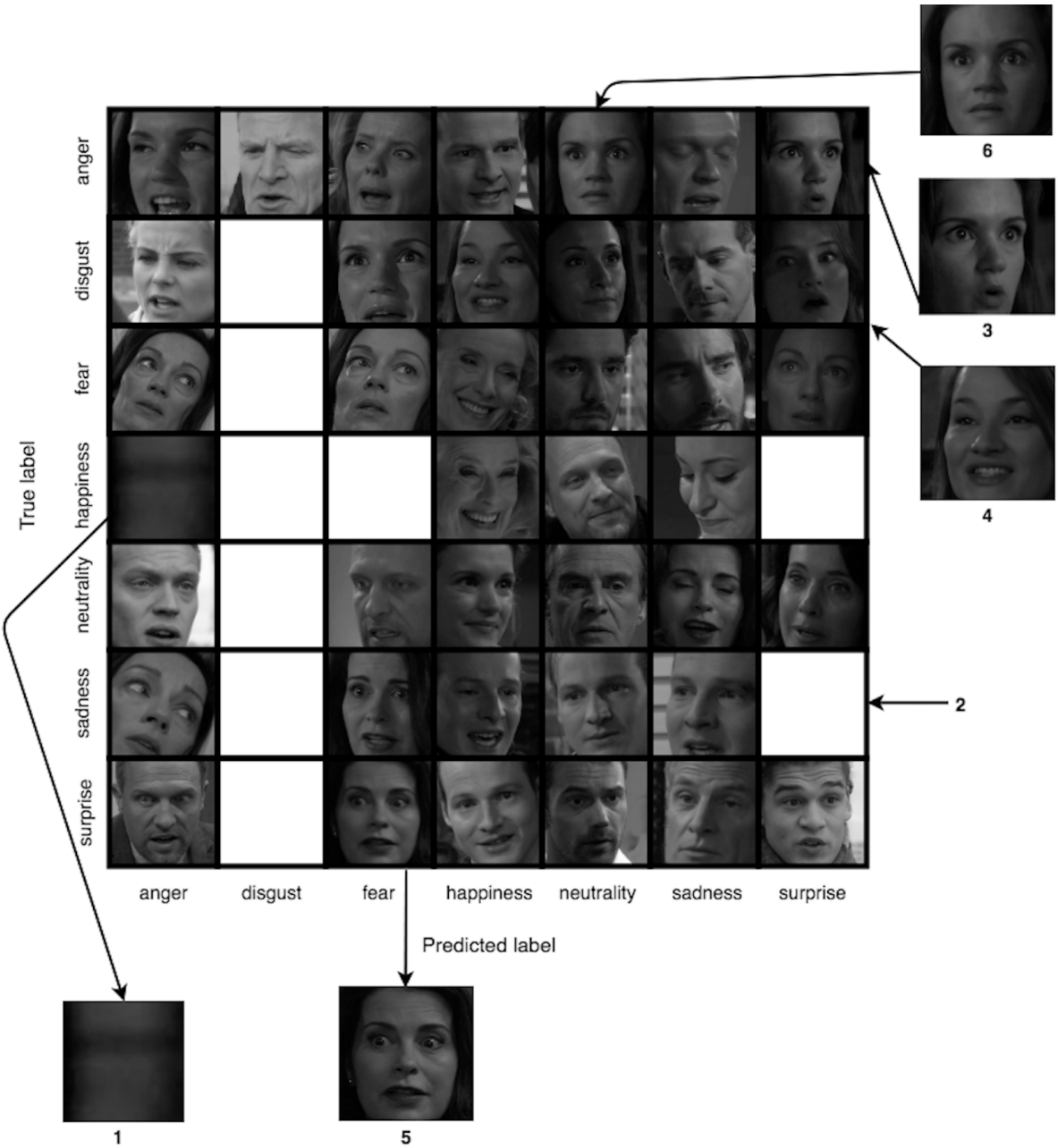


Figure 15: Confusion matrix for the mini-Xception predictions (the max count heuristic) with marked examples.

From Table 6 one can see how ambiguous the subtitle dataset is even for human annotation. In comparison to the data sources collected specifically for the tasks of emotion recognition and sentiment analysis, the subtitle data that was used for this research does not necessarily have strong emotion markers, such as opinion words, in many documents. Subtitles 2, 4, 5, 6, 7 contain common language words and humans still managed to assign them with some emotions.

The other difference between this subtitle data and special sentiment or emotion datasets is a higher level of ambiguity. For example, the subtitles 4, 5, 6 can be neutral or bear an emotion flavour which can also be very different: *'What can happen?'* can be said in a sad, fearful, or happy moment. The human annotators faced this ambiguity in the labelling task and it resulted in the sparse and heterogeneous labels that the classifier treated as the ground truth.

The classifiers that we trained treated word embeddings as the language representation in a high-dimensional space. Given that words, especially the words that do not bear strong emotion connotation in their meaning, can carry different emotions (see the word *dood* in subtitles 1 and 3) depending on the context it would make the task of emotion classification very hard for a model. The contextual and bidirectional nature of BERT word embeddings should have helped to resolve this problem, but the average performance was still not as high as expected. The ambiguity between classes and the multiclass nature of the problem (which by definition is more difficult for a classifier than binary problems) explain low performance.

Our results, however, can still be a valuable contribution to the transfer learning approach and to BERT in particular. Since Dutch word embeddings are inputted into the multilanguage model which consisted of 104 languages, this model is quite sparse and it may be not powerful enough to capture subtle class differences (such as in the task of multiclass emotion classification) in a one specific language. In addition, the vocabulary size that is created for each language may not be sufficient and produce only partial tokens

In the list below there are a couple of examples of BERT tokens:

1. Het sp ##ij ##t me. (Original: Het spijt me. Translation: I'm sorry.)
2. Miss ##chien word ik gear ##reste ##erd. (Original: Misschien word ik gearresteerd. Translation: Maybe I'm being arrested.)

The '##' sign indicates that BERT's tokenizer did not find such a word in BERT's vocabulary and divided the word into subwords which are presented in the vocabulary. BERT lacked some basic words that carry emotion such as *spijt* (sorry) and *gearresteerd* (arrested) in its vocabulary.

The division of words into subwords is a feature of the WordPiece tokenization method (Wu et al., 2016). WordPiece is aimed at automatically generating language vocabulary list which is compact, but yet representative of a language. Out of vocabulary words in this method are decomposed to blocks that are presented in vocabulary. This blocks tend to have some meaning (since they are a part of a vocabulary list): strawberries → ##straw, ##berries (Shukri H, 2019), Johanson → ##Johan, ##son[16].

Even though through the learning process the embeddings *sp, ij, t* can help the classifier produce the right label, we argue that the quality of language representation can be better for the multilingual BERT model. Firstly, the examples that were shown above are not decomposed in accordance with semantics and Dutch morphology (see sp-ij-t and gear-reste-erd).

Secondly, the vocabulary size of the English BERT embeddings is 30,522 (Shukri H, 2019), while in multilingual (104 languages) model the size is 110,000 [17]. Hence, each language in the multilingual model contains on average 1,000 vocabulary words. Even though the researchers used sophisticated weighting algorithms for balancing representation of more and less "popular" languages in Wikipedia, the vocabulary size of each language in the multilingual model differs drastically in comparison to the monolingual English model.

[16]Example from: https://github.com/google-research/bert#tokenization.
[17]https://github.com/google-research/bert/blob/master/multilingual.md#tokenization.

In addition, here is an example of an English document after BERT's tokenization: "i predict , sadly , 2017 will set a new record . you see , the way drug addiction works is like am ##way . one addict has to find new addict ##s to support their habit . the problem grows exponential ##ly ." The tokenizer is able to conduct morphological segmentation there (addicts → addict, s) and to divide a complex word into subwords in a semantically plausible way (exponentially → exponential, ly). There is only three times where out-of-vocabulary words are divided into subwords in a 34 words piece of text. The average length of a subtitle in our dataset is 13 ($SD = 11$), words and the subword division was done by the tokenizer on average 4 ($SD = 4$) times per subtitle.

To conclude, the performance of the BERT-based classifier on the multiclass problem and non-English data is quite low. We believe that the reason for the poor performance lies in the subtitle ambiguities which make the learning process more difficult for a statistical model. In addition, the quality of the model itself has a negative effect: small vocabulary size and poor preprocessing is an obstacle for such a fine-grained and subtle classification task as emotion recognition.

4.3 Reflecting on the bimodal model performance

The predictions of the facial and textual emotion classifiers were used to try different heuristics (average value, maximal value, weighted mean) and machine learning models (SVM, Random Forest) for training and testing the bimodal model. The metrics of the best performing bimodal model were compared to two unimodal models to find out if joining textual and visual channels improves the recognition of emotions in the soap opera data.

Joining the outputs of mini-Xception and BERT-based classifier did not provide an improvement to the performance. When the performance of the three classifiers (facial, textual, bimodal) where compared to their respective ground truth labels (V, S, VS), the following results were obtained. Random Forest trained on the joined probability vectors of mini-Xception and BERT gave 0.19 macro F1-score performance on 10-fold cross-validation which is slightly better than mini-Xception performance (the difference is insignificant: $p > 0.05$) and slightly worse than BERT's performance (the difference is significant: $p < 0.05$).

This result contradicts the research of Poria et al. (2016) who conducted a similar experiment and compared the performance of all possible modality combinations of audio, video, and text. They, however, used one set of ground truth labels for all the experimental set-ups which can be treated as a performance bias: the metric that was reported, for instance, for v + t combination also included audio (non-verbal) channel, since people were exposed to all the modalities while labelling the data. Using the same ground truth labels (VS dataset) for all three models also did not support the findings of Poria et al. (2016). Running the experiment on the same ground truth VS labels for all three models gave the same result as in testing the performance on the different test sets. The BERT-based TensorFlow classifier slightly outperformed bimodal (0.20 F1-score in comparison to 0.19), and the bimodal model did slightly better than mini-Xception (0.19 in comparison to 0.16). The dependent t-tests, however, showed that these differences are insignificant.

We could not confirm our initial hypothesis bases on the work of Poria et al. (2016) that joining the modalities would boost the classification performance in the task of automatic emotion recognition. When tested on different ground truth datasets, the textual emotion classifier did significantly better than the bimodal model which contradicts this hypothesis. We have to, however, take into account that due to the fact the ground truth labels were derived from three different annotation tasks, their quality has direct influence on the evaluation scores that were produced, and, consequently, on the results of the independent t-test. When tested on the same ground truth dataset, the trend in the performance was the same (BERT - bimodal - mini-Xception) as in the previous experimental set-up, but the difference in performance turned out to be insignificant.

4.4 Limitations

In terms of the experimental set-up, our method has higher internal validity than Poria et al. (2016) had in their research. The data that was used for training (textual emotions) and testing facial emotions and textual emotions utilized the labels that were retrieved from video and text channels respectively. The bimodal annotation set-up covered both of the channels. The audio cue was muted for all three annotation tasks. The downside of such an approach is that we had to organize non-overlapping groups of people for collecting three dataset which led to uneven distribution of labels created by people. Some of our annotators labelled around 2-3% of the data, others did 27 - 42% of the assigned task. Hence, the datasets that we collected are largely dependent on the contribution of 3-4 people which leads to subjective labels that are treated as ground truth by our models.

Even though in our research the data that was labelled and inputted to the emotion classifiers is derived from the "in the wild" environment of a soap opera, it is still confined in terms of its reflection of the real world settings. In our preprocessing we eliminated the shots with multiple faces, with multiple non-neutral emotion states, and with no faces. All of these conditions play an important role when people recognize emotions in the real world, but they are hard to reflect in a speaker-oriented one-label classification set-up of emotion recognition.

Excluding the audio cue from the signal is an apparent limitation in our research. The most common feedback from the human annotators that we received is how hard it was for them to pin one emotion label on a fragment, especially for the annotators in the video annotation task and the subtitle annotation task.

The fact that mini-Xception was trained on the grayscale images and it also turned our shot data to grayscale to make emotion predictions is another limitation of our research, as there is a difference in the input perceived by the human annotators and our models.

Our decision to make the classification on the shot level has its own limitations. Since shots are usually quite brief in duration (several seconds) it can be difficult even for human annotators to assess emotions there. Treating shots as if they are independent from the content of the scene can result in ambiguities in labelling. In Figure 16 an example of such a confusing shot is provided. The context of the scene where the shot is as follows: a woman 1 whose hair is visible aggressively asks a woman 2 depicted on the screen with whom she has been talking a moment ago. 2 answers that it was her accountant. 1 later accused 2 of being sneaky and secretly trying to overtake 1's father's company. In the context of the scene this shot is filled with negative emotions. *"Dear, that was my accountant"* is an attempt of 2 to prove her innocence, and, probably the actress is trying to portray fear of disclosure. A video annotator labelled it with the 'happy' emotion, the subtitle annotator - with neutrality, and the VS annotator assigned this shot with the surprise label. The VS annotation is probably the best one, as this line of 2 was the first one after the woman 1 entering the room and starting speaking angrily. The labels of V and S annotators are also justifiable given the amount of information they were provided. The face of the woman depicts a smile (happiness) and the subtitle on its own does not have any opinion words, it is pretty neutral.

In addition, the data of the soap opera can be quite different from the real world when it comes to the portrayal of emotions. Moreover, the quality of acting in the soap opera can differ even from the other (semi-)scripted content, such as films, TV-series, reality shows, and talk shows. The investigation of the mini-Xception performance showed that the way soap opera actors portrayed emotions sometimes was very subtle or confusing.

A matter of cultural differences arise when assessing facial emotions of Dutch actors on the model that was trained on the facial emotions of the English speakers. Even though the difference in the cultural display of emotion should not be that big, as between, for example, Japanese and American people (Ekman and Friesen, 1969), this cultural aspect should be taken into account, especially since we are incorporating the textual emotion classifier which is language specific. We should once again pinpoint that the reported score of the mini-Xception

Figure 16: A shot example. Accompanying subtitle is *"Schat, dat was m'n accountant." ("Dear, that was my accountant.")*.

network was reported from the images that were collected by querying English opinion words.

4.5 Future research

Given the list of limitations described in the previous paragraphs, we see the direction for the future research as follows.

The quality of the human annotation, especially in the unimodal (facial, textual) settings, can be improved by extending the pool of annotators, testing the created labels between different annotators, and calculating the inter-annotator agreement rate. Another step to improve the quality of our test dataset is to enrich it with extra annotated data, so that it will have a balance in representation of emotion categories.

In addition, the quality of the textual emotion classification will probably increase if we add more data for the fine-tuning classifier. The smallest size of a training set that was reported in the original BERT publication (Devlin et al., 2018) contained 2,500 examples, while our dataset contained 1,549 training samples (and we also had class imbalance).

In the following paragraphs we present some ideas for the long-term development. Adding the audio cue to the analysis should also improve the annotation of the textual data. The audio cue would open non-verbal content to annotators such as pitch and prosody. Non-verbal characteristics of an utterance should resolve a lot of ambiguities of a bare textual subtitle. In the research of Poria et al. (2016) the model that incorporated textual, audio, and video data had the highest performance among all the possible modality combinations[18]. Hence, we expect that adding the audio cue would not only improve the quality of the labelled dataset, but also improve automatic emotion classification. Adding other features, besides the non-verbal characteristics of voice, such as breathing patterns and gestures, is also an interesting direction for the future research.

A possible lane of development for our facial emotion classifier is to cover a wider range of possible scenarios that are widely spread both in the real world and in the film/TV scripted content. Such scenarios include group-level emotion recognition (when there are many people presented in one shot), finding a method to evaluate a dichotomy in emotions between a listener

[18]The fact that all the model configurations were trained and tested on one set of ground truth labels put some constraints on this claim though.

depicted on the screen and the textual cue, which does not belong to a person depicted on the screen. Assessing the colour scheme of a shot where there are no faces presented could be another valuable contribution for emotion evaluation in films and soap operas. Since these are art domains, colours prevailing in the scene settings can be deliberately chosen by directors to set a specific mood.

5 Conclusion

We compared newly released deep learning models for facial emotion recognition (mini-Xception) and NLP tasks (BERT) on a new data source of the soap opera which has a lot of unpredictable variations in the video cue and which contains a lot of ambiguous textual lines that were not carefully selected for training an emotion recognition model. Our comparison showed that we need to treat critically high performance metrics which are reported in the papers where the state-of-the-art models are introduced. Cross-dataset evaluation of such models should become a wide spread practice in the machine learning community.

The task of multiclass emotion classification turned out to be particularly difficult for the tested machine learning models. In the dynamic settings of everyday live (which the GTST soap opera mimics) it can be very challenging for a classifier to find robust features to learn and to use for emotion assessment. As it was demonstrated in our research, the same phrases can be said with different emotion, people can hide emotion or portray them with deceptive emotion markers (i.e. an insult can be said with a smile on the face). Beside the models, people themselves tend not to be quite accurate on the task of emotion recognition. When Goodfellow et al. (2013) asked participants to assess emotions on actors faces (the actors were asked beforehand to portray specific emotions), the accuracy was around 68%.

Despite these critical points, we still believe that automatic emotion recognition can be at least improved to the human performance on that task. Incorporating different channels of information, capturing bigger palette of emotion than Ekman's seven basic emotions, testing models on a more 'true-to-life' data are the lanes for development. Emotion recognition models can and will benefit various domains of human life, the entertainment industry in particular. Media companies, such as RTL Nederland, can use such models for assessing viewership metrics, they can incorporate emotion features into the recommendation system for TV-series and movies, and emotion recognition models can be used for assessing new pilots of shows at the focus groups viewings.

6 Bibliography

Alonso-Martín, F., Malfaz, M., Sequeira, J., Gorostiza, J. F., and Salichs, M. A. (2013). A multimodal emotion detection system during human–robot interaction. *Sensors*, 13(11):15549–15581.

Arora, S., Liang, Y., and Ma, T. (2016). A simple but tough-to-beat baseline for sentence embeddings.

Arriaga, O., Valdenegro-Toro, M., and Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*.

Avots, E., Sapiński, T., Bachmann, M., and Kamińska, D. (2018). Audiovisual emotion recognition in wild. *Machine Vision and Applications*, pages 1–11.

Becker, K., Moreira, V. P., and dos Santos, A. G. (2017). Multilingual emotion classification using supervised learning: Comparative experiments. *Information Processing & Management*, 53(3):684–704.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM.

Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A. (2017). *A practical guide to sentiment analysis*. Springer.

Chiţu, A. G., Van Vulpen, M., Takapoui, P., and Rothkrantz, L. J. (2008). Building a Dutch multimodal corpus for emotion recognition. In *Programme of the Workshop on Corpora for Research on Emotion and Affect*, page 53.

Chopade, C. R. (2015). Text based emotion recognition: A survey. *International journal of science and research*, 4(6):409–414.

Connie, T., Al-Shabi, M., Cheah, W. P., and Goh, M. (2017). Facial expression recognition using a hybrid cnn–sift aggregator. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, pages 139–149. Springer.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., and Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771.

De Silva, L. C., Miyasato, T., and Nakatsu, R. (1997). Facial emotion recognition using multi-modal information. In *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, volume 1, pages 397–401. IEEE.

Devlin, J. and Chang, M.-W. (2018). Open sourcing BERT: State-of-the-art pre-training for natural language processing. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html. Accessed on: 12-02-2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Ekman, P. and Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98.

Ekman, P. and Friesen, W. V. (1978). *Facial action coding system: Investigator's guide*. Consulting Psychologists Press.

Ekman, P. and Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Ellis, J. G., Jou, B., and Chang, S.-F. (2014). Why we watch the news: a dataset for exploring sentiment in broadcast video news. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 104–111. ACM.

Fan, Y., Lu, X., Li, D., and Liu, Y. (2016). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450. ACM.

Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: A survey. *Pattern recognition*, 36(1):259–275.

Field, A., Miles, J., and Field, Z. (2012). *Discovering statistics using R*. Sage publications.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer.

Haq, S. and Jackson, P. (2010). Machine audition: Principles, algorithms and systems, chapter multimodal emotion recognition. *IGI Global, Hershey PA*, pages 398–423.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ho, A. T., Menezes, I. L., and Tagmouti, Y. (2006). E-mrs: Emotion-based movie recommender system. In *Proceedings of IADIS e-Commerce Conference. USA: University of Washington Both-ell*, pages 1–8.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, P., Cai, D., Wang, S., Yao, A., and Chen, Y. (2017). Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 553–560. ACM.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jeon, M. (2017). Emotions and affect in human factors and human–computer interaction: Taxonomy, theories, approaches, and methods. In *Emotions and Affect in Human Factors and Human-Computer Interaction*, pages 3–26. Elsevier.

Jurafsky, D. and Martin, J. H. (2018). *Speech and language processing*.

Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R. C., et al. (2013). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM.

Katsis, C. D., Katertsidis, N., Ganiatsas, G., and Fotiadis, D. I. (2008). Toward emotion recognition in car-racing drivers: A biosignal processing approach. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(3):502–512.

Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401.

Li, Y., Tao, J., Schuller, B., Shan, S., Jiang, D., and Jia, J. (2016). MEC 2016: the multimodal emotion recognition challenge of CCPR 2016. In *Chinese Conference on Pattern Recognition*, pages 667–678. Springer.

Lin, C., He, Y., and Everson, R. (2011). Sentence subjectivity detection with weakly-supervised learning. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1153–1161.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Litman, D. and Forbes, K. (2003). Recognizing emotions from student speech in tutoring dialogues. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 25–30. IEEE.

Liu, W., Zhang, L., Tao, D., and Cheng, J. (2018). Reinforcement online learning for emotion prediction by using physiological signals. *Pattern Recognition Letters*, 107:123–130.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*.

Munezero, M. D., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.

Ng, A. (2018). Convolutional Neural Networks. Edge detection example. https://youtu.be/5lvG3FfP0lg. Accessed on: 06-02-2019.

Noroozi, F., Corneanu, C. A., Kamińska, D., Sapiński, T., Escalera, S., and Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *arXiv preprint arXiv:1801.07481*.

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

Perepelkina, O., Kazimirova, E., and Konstantinova, M. (2018). RAMAS: Russian multimodal corpus of dyadic interaction for studying emotion recognition. *PeerJ Preprints*, 6:e26688v1.

Pérez-Rosas, V., Mihalcea, R., and Morency, L.-P. (2013). Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 973–982.

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Poria, S., Cambria, E., and Gelbukh, A. (2015a). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544.

Poria, S., Cambria, E., Hussain, A., and Huang, G.-B. (2015b). Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, 63:104–116.

Poria, S., Chaturvedi, I., Cambria, E., and Hussain, A. (2016). Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 439–448. IEEE.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE.

Rosas, V. P., Mihalcea, R., and Morency, L.-P. (2013). Multimodal sentiment analysis of Spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45.

Rui, Y., Huang, T. S., and Mehrotra, S. (1999). Constructing table-of-content for videos. *Multimedia systems*, 7(5):359–368.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Sariyanidi, E., Gunes, H., and Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133.

Sasaki, Y. et al. (2007). The truth of the f-measure. *Teach Tutor mater*, 1(5):1–5.

Shukri H, M. (2019). How the embedding layers in BERT were implemented. https://medium.com/@_init_/why-bert-has-3-embedding-layers-and-their-implementation-details-9c261108e28a. Accessed on: 12-06-2019.

Simon, H. A., Bibel, W., Bundy, A., Berliner, H., Feigenbaum, E., Buchanan, B., Selfridge, O., Michie, D., Nilsson, N., Sloman, A., et al. (2000). AI's greatest trends and controversies. *IEEE Intelligent Systems and Their Applications*, 15(1):8–17.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer.

Uszkoreit, J. (2017). Transformer: A novel neural network architecture for language understanding. https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html. Accessed on: 11-02-2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Vielzeuf, V., Kervadec, C., Pateux, S., Lechervy, A., and Jurie, F. (2018). An Occam's razor view on learning audiovisual emotion recognition with small training sets. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 589–593. ACM.

Vielzeuf, V., Pateux, S., and Jurie, F. (2017). Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 569–576. ACM.

Viola, P., Jones, M., et al. (2001). Robust real-time object detection. *International journal of computer vision*, 4(34-47):4.

Wang, C.-F. (2018). A Basic Introduction to Separable Convolutions. https://towardsdatascience.com/a-basic-introduction-to-separable-convolutions-b99ec3102728. Accessed on: 12-02-2019.

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., and Kennedy, P. J. (2016). Training deep neural networks on imbalanced data sets. In *2016 international joint conference on neural networks (IJCNN)*, pages 4368–4374. IEEE.

Wang, Y.-Q. (2014). An analysis of the Viola-Jones face detection algorithm. *Image Processing On Line*, 4:128–148.

Wu, C.-H., Lin, J.-C., and Wei, W.-L. (2014). Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing*, 3.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yang, S., Luo, P., Loy, C.-C., and Tang, X. (2016). Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533.

Yannakakis, G. N. (2012). Game AI revisited. In *Proceedings of the 9th conference on Computing Frontiers*, pages 285–292. ACM.

Yu, H. and Liu, H. (2015). Combining appearance and geometric features for facial expression recognition. In *Sixth International Conference on Graphic and Image Processing (ICGIP 2014)*, volume 9443, page 944308. International Society for Optics and Photonics.

Yuvaraj, R., Murugappan, M., Ibrahim, N. M., Omar, M. I., Sundaraj, K., Mohamad, K., Palaniappan, R., Mesquita, E., and Satiyan, M. (2014). On the analysis of eeg power, frequency and asymmetry in parkinson's disease during emotion processing. *Behavioral and brain functions*, 10(1):12.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58.

A Appendix

class	precision	recall	F1-score
anger	0.24	0.09	0.14
disgust	0.00	0.00	0.00
fear	0.10	0.04	0.06
happiness	0.30	0.38	0.34
neutrality	0.24	0.48	0.32
sadness	0.19	0.42	0.26
surprise	0.46	0.06	0.10
Metrics averages across classes			
micro avg	0.23	0.23	0.23
macro avg	0.22	0.21	**0.17**
weighted avg	0.24	0.23	0.18

Table 7: Classification report for the max count heuristic.

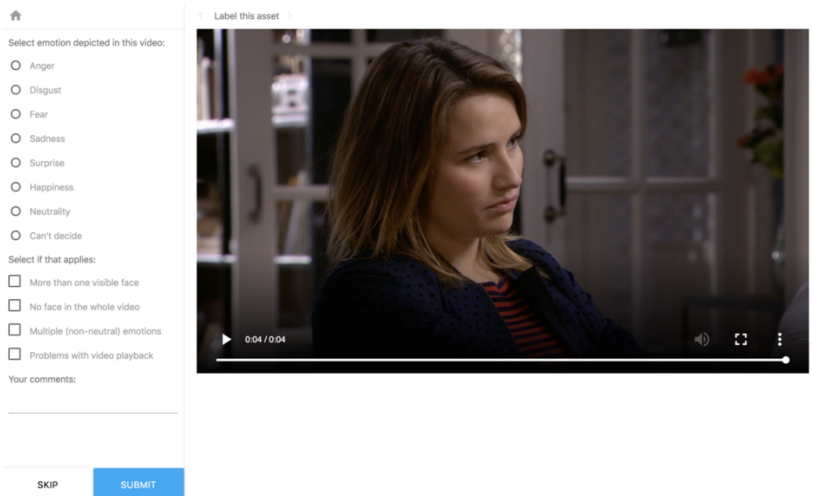
class	***Native TensorFlow BERT implementation***			
	precision	**recall**	**F1-score**	**support**
anger	0.15	0.32	0.20	74
disgust	0.14	0.03	0.04	38
fear	0.17	0.05	0.08	59
happiness	0.33	0.42	0.37	66
neutrality	0.54	0.41	0.46	254
sadness	0.28	0.38	0.32	64
surprise	0.30	0.33	0.32	57
Metrics averaged across classes				
micro avg	0.33	0.33	0.33	612
macro avg	0.27	0.28	**0.26**	612
weighted avg	0.36	0.33	0.33	612

Table 8: Classification report for native TensorFlow module for fine-tuning BERT.

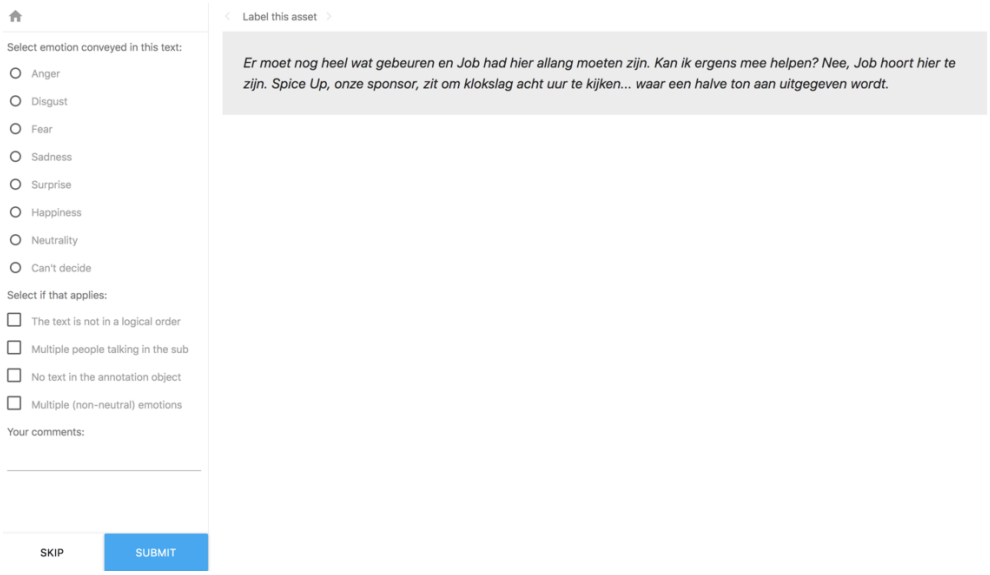
In Table 7 the detailed classification report of the max count heuristic for capturing temporal variation applied on the mini-Xception output is provided. Since all the heuristical approaches (max count, averaging, and max pooling) produced the same macro F1-score of 0.17, the max count classification report was chosen randomly. Class-specific F1-scores in this report show that top performing classes on the V dataset are happiness, neutrality, and sadness. These classes also constituted the largest proportion of the FER-2013 on which mini-Xception was trained.

In Table 8 the classification report of the native TensorFlow implementation of fine-tuning for the BERT model is shown. The macro F1-score across seven emotion classes is 0.26. The lowest performance belong to the disgust class. The post-hoc Turkey's test demonstrated that disgust was classified significantly ($p < 0.05$) worse than happiness, neutrality, surprise, sadness, and fear.

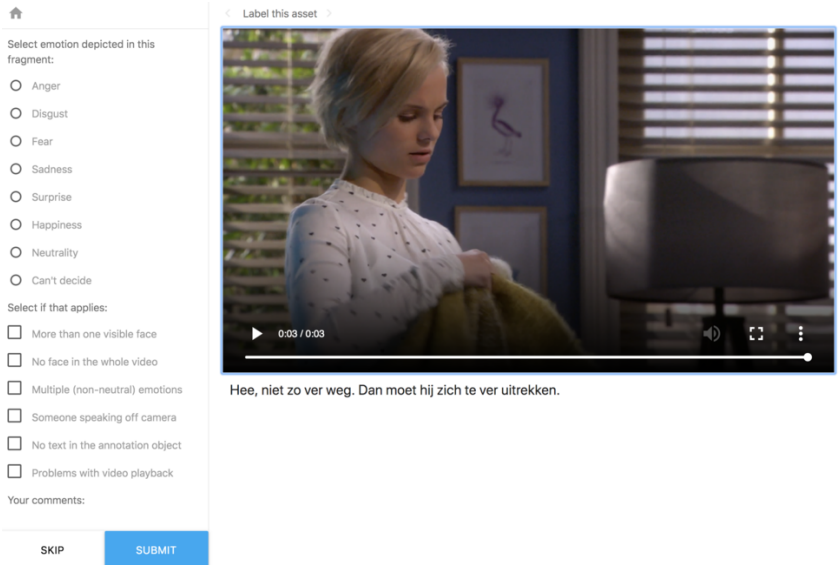
In Figure 17 the interfaces for three annotation tasks are depicted. In the V annotation task (17a) people were to assign an emotion label and to report the noise conditions (if presented) in a video shot with muted audio. In the S annotation task (17b) annotators had to do the same task but for the subtitle line of a shot. In the VS annotation task (17c) annotators were supposed to assign an emotion label based on both video and textual cues. The V, S, and VS tasks consisted of non-overlapping groups of annotators.



(a) The interface for video annotation.



(b) The interface for subtitle annotation.



(c) The interface for video+subtitle annotation.

Figure 17: Set-ups for the video (17a), subtitle (17b), and bimodal (17c) annotation environments.