The Many Faces of Anger: A Multicultural Video Dataset of Negative Emotions in the Wild

by

Roya Javadi

B.Sc., Sharif University of Technology, 2015

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Fall 2021

Declaration of Committee

Name:	**Roya Javadi**
Degree:	**Master of Science**
Thesis title:	**The Many Faces of Anger: A Multicultural Video Dataset of Negative Emotions in the Wild**
Committee:	**Chair:** Maxwell Libbrecht Assistant Professor, Computing Science
	Angelica Lim Supervisor Assistant Professor, Computing Science
	Mo Chen Committee Member Assistant Professor, Computing Science
	Steve DiPaola Examiner Professor, Interactive Arts and Technology

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

The portrayal of negative emotions such as anger can vary widely between cultures and situations, depending on the acceptability of expressing full-blown emotions rather than suppression to maintain harmony. The majority of emotional datasets collect data under the broad label "anger", but social signals can range from annoyed, contemptuous, angry, furious, hatred, and more. In this thesis, we curated the first in-the-wild multicultural video dataset of emotions, and deeply explored anger-related emotional expressions by asking culture-fluent annotators to label the videos with 6 labels and 13 emojis in a multi-label framework. We provide a baseline multi-label classifier on our dataset, and show how emojis can be effectively used as a language-agnostic tool for annotation.

Keywords: Affective Computing; Human Emotion Recognition; Multi-Label Classification

Dedication

To my family and
my grandmothers, in loving memory.

Acknowledgements

First and foremost I am extremely grateful to my supervisor, Prof. Angelica Lim for her invaluable advice, continuous support, and patience during my Masters study. I would also like to thank Prof. Mo Chen for his feedback on my thesis. I would like to thank all the members in the ROSIE Lab, especially Emma Hughson for her part in data collection. It is their kind help and support that have made my study and life during COVID-19 pandemic a wonderful time. Finally, I would like to express my gratitude to my husband Hamid Ramazani, my parents, my siblings, and my grandfather. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Anger is a basic emotion suggested to be found around the world. Despite the universal theories that associated prototypical expressions such as bared teeth to anger, depictions of anger come in a variety of social signals and arousal levels [1]. Furthermore, some research [8, 28, 10] suggests that more collectivist cultures tend to suppress negative emotions to maintain harmony. These reasons imply that correctly recognizing anger in videos is highly dependent on the samples fed into models for emotion recognition tasks.

Automatic emotion expression recognition for anger is relatively poorly studied and deserves more investigation. In [29] for example, the authors found that using both AffectNet and FER+, resulted in accuracy between 45% and 54% for contempt, anger, and disgust on AffectNet dataset, while happiness had a 77% accuracy rate. Similarly, in [18] where images were multi-labeled with 26 categories, after 'aversion' with an average precision of 7.48, anger was the worst-recognized basic emotion with an average precision of 9.49, followed by fear (14.14), surprise (18.81), sadness (19.66) and happiness (60.69). Hence, a deeper understanding of the anger emotion is necessary.

Most of the work in this field is on images, but dynamics of face and head are also important. Therefore, we decide to focus on dynamic representations of emotions, such as videos. A major challenge in emotion recognition in videos is to find a high-quality dataset, which includes both high-quality videos and their corresponding labels. In particular, it is important to gather in-the-wild datasets, as in-the-lab collected datasets such as [31] may not well represent the variety of possible expressions. Among in-the-wild datasets, GIFGIF+ [5] introduced by MIT is a relatively massive and realistic video dataset that consists of more than 23000 animated GIFs over 17 emotions. The authors' dataset collection method is semi-automatic, meaning that they used both human labor and clustering techniques to label GIF samples. Aff-Wild2 [17] is another in-the-wild video dataset containing 260 videos that is used for estimation of the valence and arousal emotion dimensions. A variety of subjects, movements, and context is also another key point in video datasets. The Affectiva-MIT Facial Expression Dataset (AM-FED) database [21] contains 242 facial videos of people watching Super Bowl commercials using their webcam. It has frame-by-frame annotations

of 14 Action Units (AU), head movements, and facial landmarks, but there is not much variance in head pose and subjects, as all participants are reacting to videos while sitting in front of a computer.

Psychological research indicates that people express emotions differently depending on their age, gender or culture [19]. A few datasets have focused on underrepresented groups in emotion recognition, namely EmoReact [24], which contains videos of children in the wild reacting to objects and answering questions about them, and ElderReact [20] where older adults react to videos and express their opinion. Both of them collected in-the-wild videos from the same channel on YouTube, containing videos of people seated at a desk.

Annotation schemes in affective computing also have room for improvement. For instance, the broad emotion label of anger is problematic for several reasons. Firstly, there is a range of anger-related phenomena which could be more precisely recognized with a more specific word. Similar to how the word "mammal" can encompass a wide range of animals, more specific words such as "cat" or "dog" can provide better labels. Here, we consider labels such as annoyed, furious, hatred, contempt, and disgust. Secondly, emotion "readings" can be a highly subjective theory of mind exercises, which can depend on the annotator or their own internal state. An alternate scheme is to provide objective behavioral descriptions of facial expressions (e.g. frowning, rolling eyes), similar to AU recognition but at a higher level. Here, we explore representations of social signals using emojis, which in addition are language-agnostic. Using emojis for labeling has gained popularity recently. Saheb Jam et al. [27] used emojis to annotate emotional expressions in videos of interaction between a human and a robot. Vemulapalli et al. have used emojis to build an embedding space for facial expressions [32]. Herein, we used 13 emojis related to 6 emotion labels of 'annoyed', 'anger', 'fury', 'hatred', 'disgust', 'contempt' in order to investigate whether mapping emojis to their emotion category can improve language-agnostic annotations. Thirdly, emotional expressions can be mixed [7]. While fields such as text, speech, and music emotion processing have readily accepted the multi-label paradigm, video-based approaches still assume one label per sample, throwing out data that does not have a high inter-rater agreement or use single-label classifiers on datasets that have multi-labels for each sample. As noted by [16], "Ambiguous, subtle expressions of emotion, which often obtain no majority agreement from human annotators, are prevalent in the real world". Since we are working on in-the-wild data, a multi-label baseline is an important component in this thesis.

To summarize, the main contribution of this thesis is collecting a dataset from underrepresented cultures in the anger category, as well as annotating it by people from the same culture. The majority of datasets usually focus on Western-Caucasian subjects and we rarely see other cultures or ethnicities such as East-Asian or West-Asian. In this research, we tried to bridge this gap by collecting data from underrepresented cultures from Middle East, which to the best of our knowledge has never been done before. A few works such as [2] designed feature extractors and classifiers on a multi-cultural (East-Asian and

Western-Caucasian) image dataset. Khanh et al. also collected a Korean emotion dataset from Korean movies and used a Multi-Layer Perceptron to classify them into basic emotions. They showed that training the model using Korean videos and testing on English videos and vice-versa yielded the worst result [14]. Another contribution of ours is a collection of social signals for each video in the dataset, aiming to identify culture-dependent facial expressions in emotions.

In this thesis, we first describe the process of raw data collection. Then, we outline the emotion and emoji annotation procedure. Finally, we present dataset statistics and a baseline classification for future benchmarking.

Chapter 2

Related Work

2.1 Overview

There have been many datasets on emotion and emotion recognition, most of them for the image modality. Video datasets, due to the challenges of collecting them, are more limited and much smaller. For example, some image datasets such as AffectNet [22] contain over 1 million images, while GIFGIF+, which is one of the massive video datasets in this field, contains 23000 videos. Therefore, enhancing emotion video datasets, especially with underrepresented groups, can be very beneficial to the affective computing field. The main limitation of GIFGIF+ is that only around the top 100 GIFs for each of anger and disgust have a relatively high confidence in their single-label paradigm.

In the Introduction, we briefly mentioned some related publications. In this chapter, we elaborate further on the most related ones.

2.2 In-the-Wild Emotion Datasets

Due to its subjective nature, annotating a large amount of affective data can be very complicated and cumbersome. Authors in [5] have proposed a combination of manual and automatic labeling. Initially, the GIFGIF dataset contained about 6000 gifs, labeled with 17 emotions by Internet users on their website using a gamification approach. The authors of [5] trained a multi-task learning model (3D CNN with transfer learning) to enable efficient labeling of a large dataset of GIFs, and enhance the number of labeled GIFs to more than 23K.

2.2.1 Under-Represented Groups Datasets

Another way of improving current datasets is to focus on underrepresented groups such as minority race, ethnicity or age. A few papers have tried to collect in-the-wild videos from such groups, including children, older adults, or people from different racial or ethnic backgrounds.

The EmoReact[24] dataset contains 1102 videos collected from React channel on YouTube, exclusively focusing on children. These videos are annotated for 17 affective states, including six basic emotions (happiness, sadness, surprise, fear, disgust, and anger), neutral, valence, and nine complex emotions including curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment, and frustration, however, the publicly available dataset contains only a small subset of these emotions. Each video was annotated by three independent workers for seventeen labels using 1-4 Likert scale. The authors compared unimodal and multimodal emotion recognition using Naive Bayes, linear SVM, and Radial Basis Function kernel SVM as their baseline models.

ElderReact [20] is another dataset collected from React channel, aiming to collect expressions of basic emotions in older adults. It contains 1323 video clips of 46 unique individuals. They allowed workers to rate multiple emotions for each video clip. The authors used OpenFace to extract visual features and COVAREP [6] to extract audio features. They summarized each video clip by calculating the mean and standard deviation of all features for all frames. Similar to EmoReact, these features were then fed to baseline classifiers: Radial Basis Function kernel Support Vector Machine [23], Gaussian Naive Bayes and XGBoost [4]. They trained the models using unimodal and multimodal features.

Although both datasets mentioned here had multi-labels for each video, none had used multi-label classifiers for predicting all possible labels and performed classification for each class separately.

2.3 Emotion Recognition in Videos

Over the past decade, some emotion recognition tasks in videos used various models of deep learning like [15] and [13], included but not limited to Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and transfer learning techniques. Kaya et al. [13] have used transfer learning on extracted faces from in-the-wild video dataset of EmotiW 2016. They used CNN and pre-trained network of VGG-FACE, and then fine-tuned the weights using FER 2013 [9] dataset. Their approach is illustrated in Figure 2.1.

AffWild2 [17] is another in-the-wild video dataset. It consists of 260 videos collected from YouTube with more than 1,400,000 frames. The researchers tried to cover a wide range of subjects, contexts, illuminations, and etc. Four annotators labeled videos for valance and arousal according to the 2D emotion model in Fig. 2.3. The annotation was continuous, that is, for a single video, annotators could assign different values in [-1, +1] interval for each subset of frames. After data collection, authors applied a CNN-RNN with attention architecture on face images extracted from frames as illustrated in Fig. 2.2. The CNN network was pre-trained on existing models including VGGFACE, ResNet-50 and DenseNet-121. The model with VGGFACE CNN network and GRU cells in RNN had the best performance.

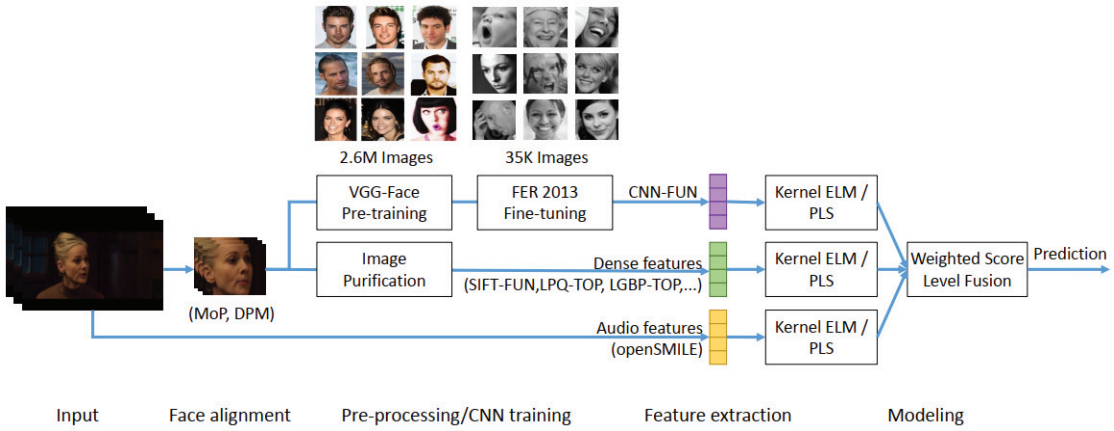


Figure 2.1: Proposed model in [13] using pre-trained networks on VGG-FACE(Copyright © 2017, Elsevier).

In this first step, we did not intend to use deep learning models on our dataset for two main reasons: First, they are black boxes and extracted feature embeddings are harder to interpret (e.g. compared to AUs), secondly, our dataset is much smaller than the aforementioned datasets and even transfer learning may lead to overfitting on training data.

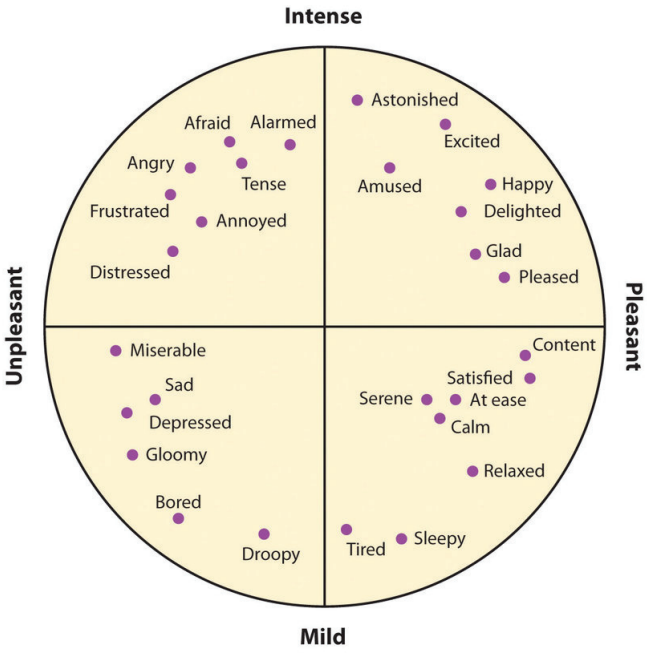


Figure 2.2: 2D Emotion Wheel (Circumplex model) used for annotation in [17] (arXiv preprint arXiv:1811.07770, 2018).

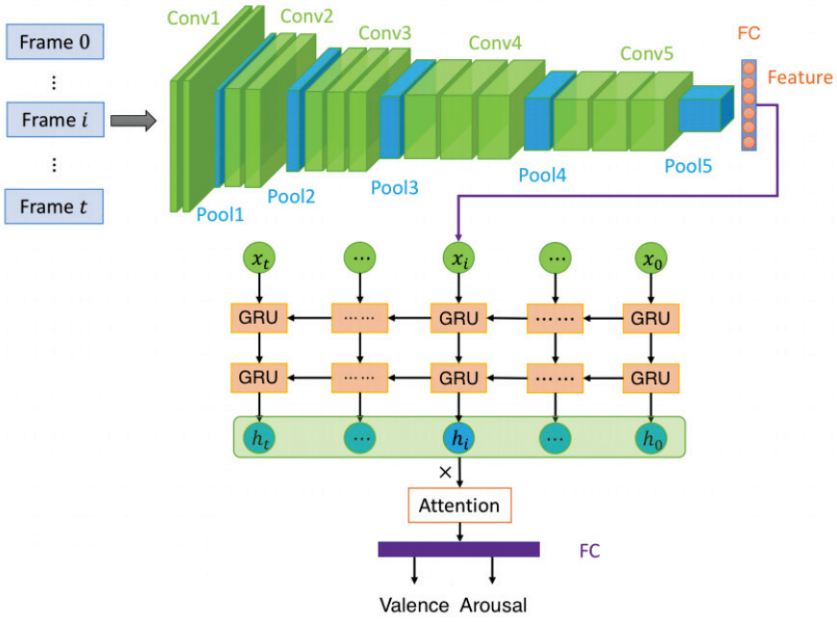


Figure 2.3: CNN-RNN architecture with pre-training in [17] (arXiv preprint arXiv:1811.07770, 2018).

Chapter 3

Data Collection

The massive amount of videos on YouTube is a great asset for researchers. A major challenge that remains to be addressed is extending videos with in-the-wild or close-to-in-the-wild emotions. We collected more than 200 videos on YouTube from Persian and North American (NA) cultures. All Persian videos were collected from TV series and movies accessible on YouTube, and NA videos included movies, "vlog" style content, reality television shows (e.g., Dance Moms), and talk shows (e.g., The Late Show with James Corden). Since some videos were too long for our purpose (e.g. containing more than one emotion) or had multiple actors, we trimmed or split, or removed them. In the end, there were 97 videos for Persian and 104 videos for NA culture, each lasting between 1 to 10 seconds.

Our main challenge in label collection was that crowd-sourcing platforms such as Amazon Mechanical Turk (AMT) do not provide a facility to choose the culture of annotators. Thus, we decided to design and implement a web interface and collect culturally fluent annotators in Canada. Annotators were 19+ years old residents of Canada who were fluent in English, as well as Farsi if annotating Persian videos.

We designed an interface using Flask[1] and an SQLite database. Each annotator registered on the website and accepted a Research Consent form. We only asked for general information about the culture, language and perceived individualism of people. No directly identifiable information was collected during the study. The participants were allowed to withdraw from the study at any time. An Amazon gift card code corresponding to approximately the minimum wage in Canada was emitted upon completion of or withdrawal from the study, and this thesis was approved by the university research ethics board. In total, we recruited 10 people from each culture via social media. Audio was removed from videos. Each user annotated half of the videos for their culture, resulting in 5 annotations per clip.

[1] https://flask.palletsprojects.com/

3.1 Multi-Cultural Annotation Interface

We designed an interface using Flask [2] and an SQLite database. Each annotator registered on the website and accepted a Research Consent form. We only asked for general information about the culture, language, and perceived individualism of people. No directly identifiable information was collected during the study. The participants were allowed to withdraw from the study at any time. An Amazon gift card code corresponding to approximately the minimum wage in Canada was emitted upon completion of or withdrawal from the study, and this thesis was approved by the university research ethics board. In total, we recruited 10 people from each culture from social media. Each user annotated half of the videos for their culture. Fig. 3.3 and 3.4 show the questionnaire panel that users interacted with.

3.2 Emotion Labels

Inspired from the anger category in the emotion wheel by The Junto Institute in Fig. 3.1, we defined 6 negative emotions for labels: Contemptuous, Annoyed, Anger, Hatred, Furious, and Disgusted. We also added a "None" option. Annotators were allowed to select more than one label and leave their idea or thoughts in a comment box. The annotations were done independently and blindly, meaning that annotators did not have any information about each other's labels.

3.3 Emoji Labels

In this thesis, we also included 13 emoji annotations to extract underlying subtle social signals of each emotion. The emojis that users could select is presented in Fig. 3.4. Pilot annotations among researchers suggested difficulty in obtaining unanimity in labeling; in many samples, people selected multiple emotions. This gave us a clue that allowing multiple labels for videos would be better than restricting the labels to only one. We accumulated all the emotion and emoji labels belonging to each video and did majority voting to derive final labels. If two or more labels had equal votes, we attached all those labels to the video. For example, a video that has 2 votes for annoyed, 2 votes for anger, and 1 vote for contempt, is assigned both anger and annoyance. Fig. 4.2 shows the distribution of the number of labels.

3.4 Social Signals Annotation

In order to explore what social signals can be found in emotional expressions, one researcher from NA cultural background and one from Persian culture extracted visible face and body expressions for each video. Since social signal annotations such as raised brows, arms crossed,

[2] https://flask.palletsprojects.com/

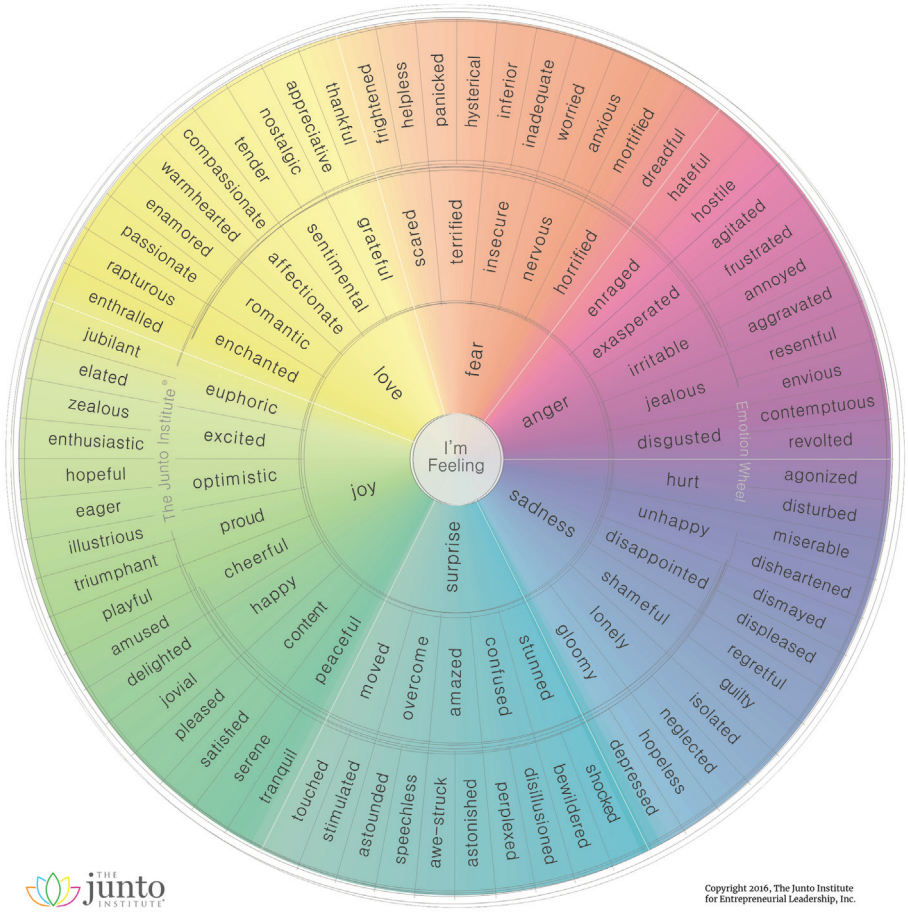


Figure 3.1: The Junto Institute's emotion wheel - (Copyright © 2016, The Junto Institute for Entrepreneurial Leadership, used with permission)

turning head, etc. are relatively objective (compared to emotion labels), both researchers reviewed all videos for this annotation step. Overall, they had consensus for over 90% of videos. The analysis we conducted on these social signals aimed to identify the cultural differences in emotion expression.

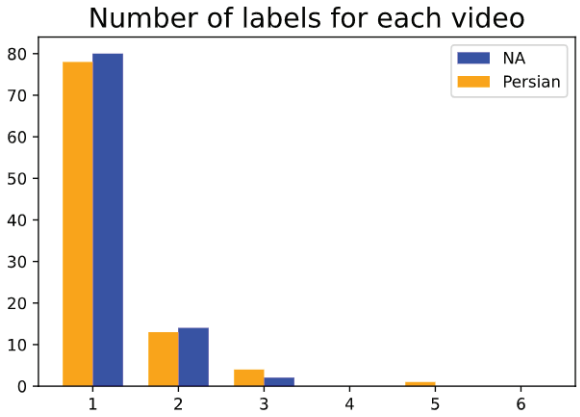


Figure 3.2: Distribution of number of emotion labels

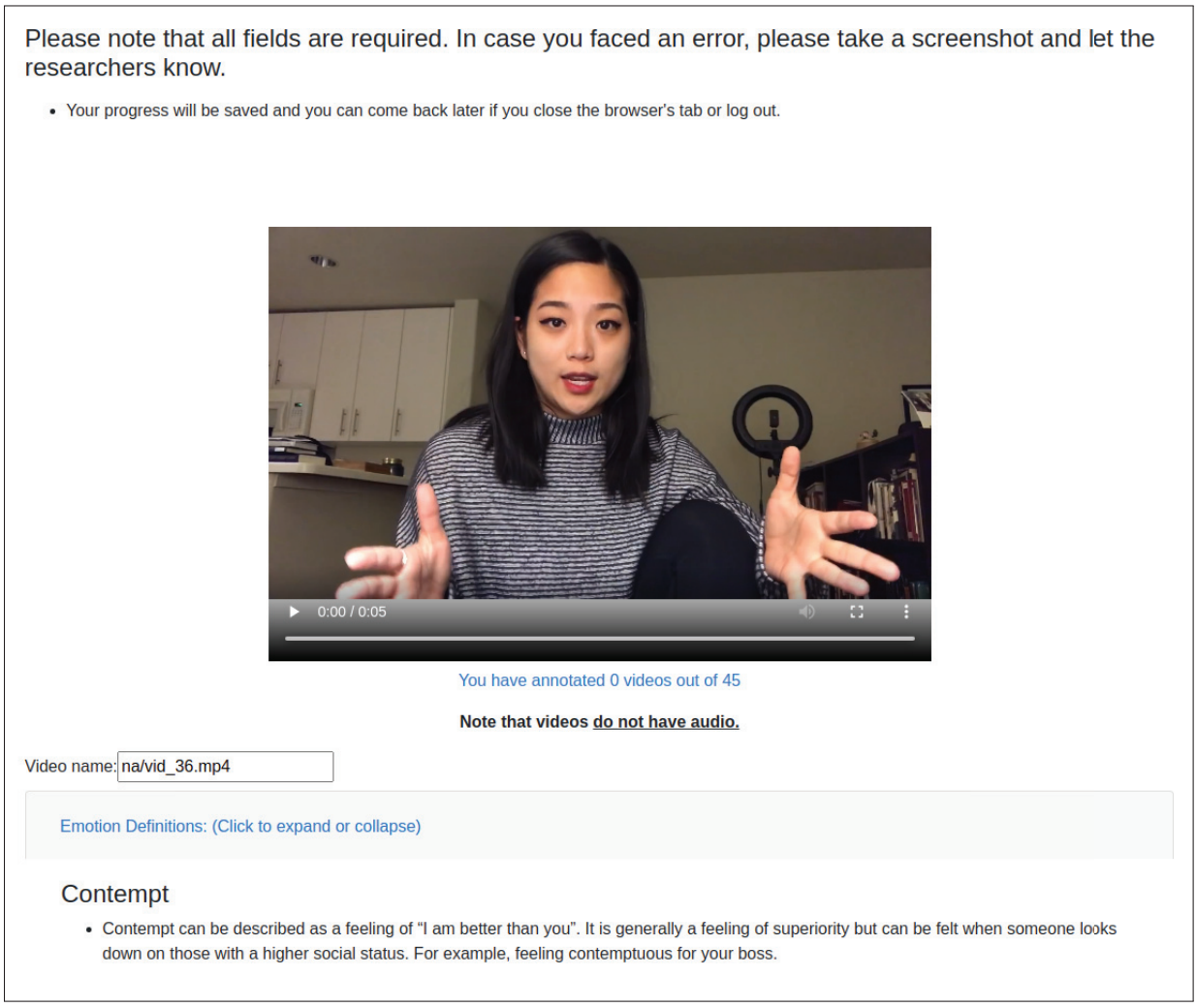


Figure 3.3: An example of a video in the multi-cultural interface

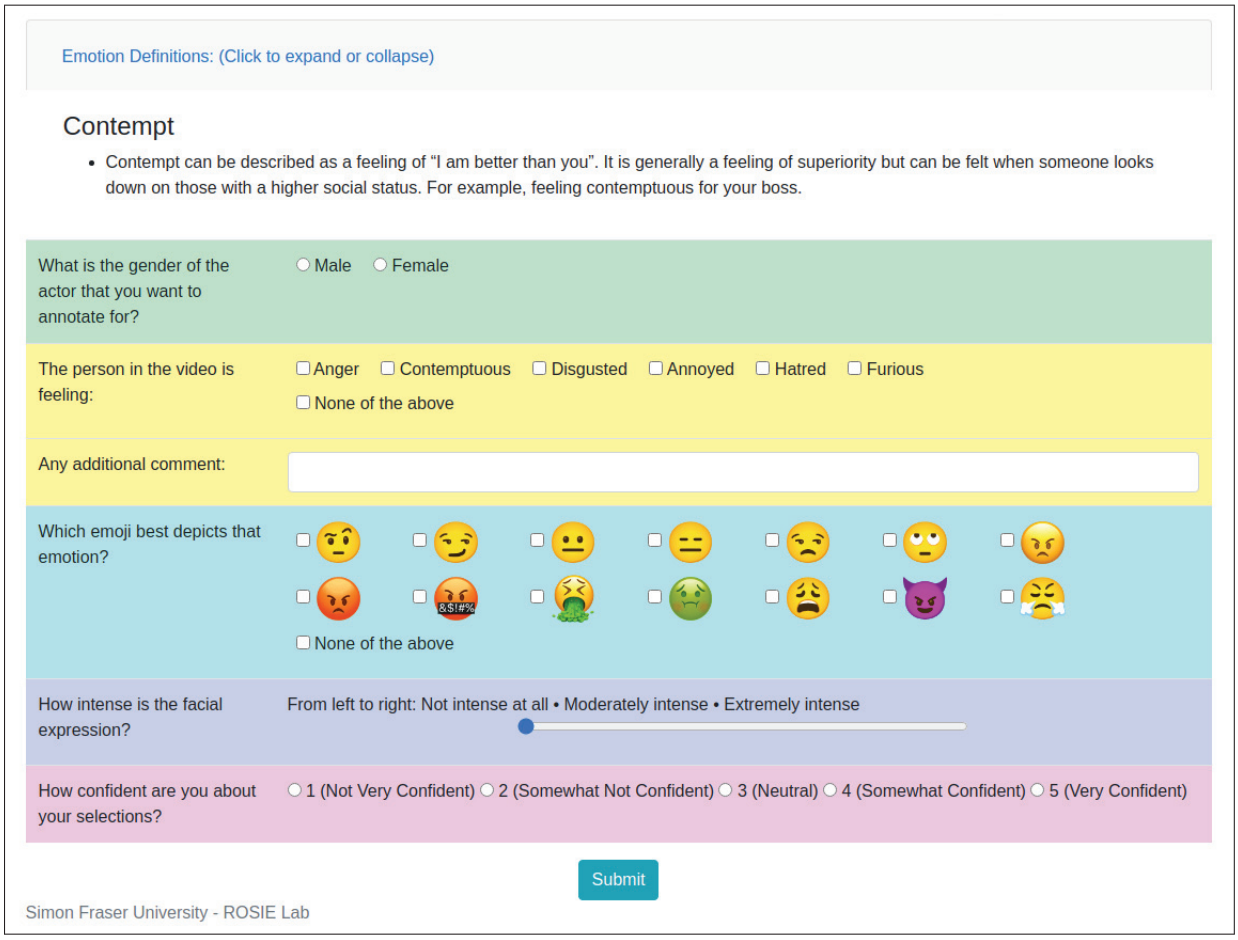


Figure 3.4: Questionnaire in the multi-cultural interface

Chapter 4

Data Analysis

In this chapter, we describe all the analyses that we performed on the dataset.

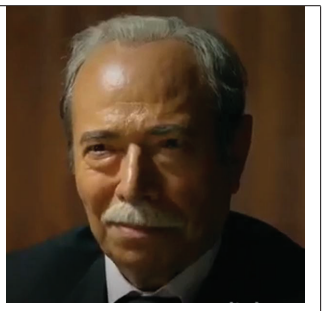
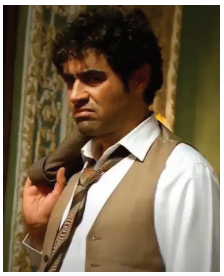
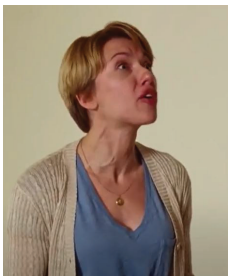
Initially, our dataset aimed to cover the emotions of Contempt, Anger, and Disgust (also known as the CAD triad [26]). The first steps of annotation revealed that annotators perceived some fine-grained emotions like annoyed, as well as complex emotions such as disgust-anger or contempt-disgust. Therefore, we decided to further refine the labels and allow people to choose several emotions. We emphasize that the purpose of this thesis is not to obtain perfect agreement, but to consider that, similar to [16], the distribution over annotators *is* the ground truth for this challenging in-the-wild data. Nevertheless, as a descriptive measure, we used Krippendorff's Alpha (used when multiple labels can be chosen) to calculate the agreement between annotations. The agreement scores for NA and Persian emotion dataset were 0.252 and 0.076, respectively. Since Jeni et al. [12] indicated that label imbalance can dramatically affect metrics such as Krippendorff's Alpha, and we can see that labels were indeed imbalanced when observing the number of videos for each emotion/emoji in Fig. 4.3 and Fig. 4.4.

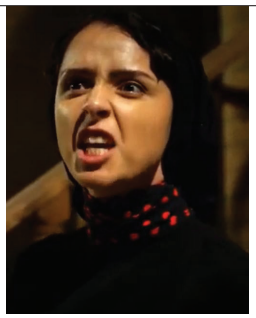
4.1 Labels Statistics

The number of videos for each emotion is illustrated in Fig. 4.3 and Fig. 4.4. Table 4.1 illustrates example frames for each emotion and their assigned emojis.

Table 4.1: Example frames for Many Faces of Anger

Emotion	NA	Persian

Contempt		
Emojis	😑, 😏	😏
Annoyed		
Emojis	😑, 😠, 🙄, 😒	😒
Anger		
Emojis	😠	😤
Hatred		
Emojis	😤, 😡	😤

Furious		
Emojis	🤬	🤬

4.1.1 Gender Statistics

The plots in Fig. 4.5 and Fig. 4.6 show the emotion labels statistics for each gender. Persian dataset has more gender imbalance than NA.

4.2 Annotators' Confidence

We calculated the mean of annotator's confidence for each label, which you can see in Fig. 4.7. The noticeable difference between the two cultures is the confidence of disgust, which is higher in NA. Upon analysis, we realized that this is because most disgust videos in NA dataset were reactions to foods, whereas in Persian 'disgust' videos were social disgust, hence more challenging to label.

4.3 Co-Occurrence of Emoji-Emotions

We calculated which emojis co-occur with which emotions. Fig. 4.10 show the emojis that appeared in more than 15% of the videos for a given label. Interestingly, annotators associated red emojis for higher arousal forms of anger, which supports the research that red faces map to higher levels of anger [11]. Also, annotators chose other emojis like 😒 over 😏 for contempt. In NA dataset, we see an abundance of 🤢 emoji for disgust while in Persian annotations there is none; the reason is that there are several videos labeled 'disgust' in NA in which a person reacts to foods, whereas Persian had social disgust. Another finding from this analysis is that contempt is not fine-grained enough; users selected 😒 label when it is combined with social disgust, while they preferred 😏 for contempt alone.

4.4 Correlation between Emotion Labels

We think the correlation between emotions may shed light on how emotions are related to each other, so we computed the Pearson correlation between the labels given by annotators

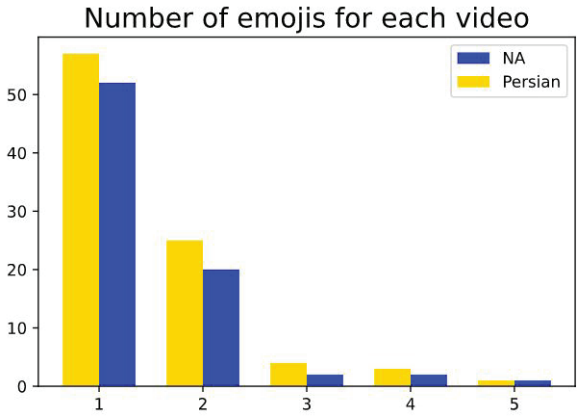


Figure 4.1: Distribution of number of emoji labels

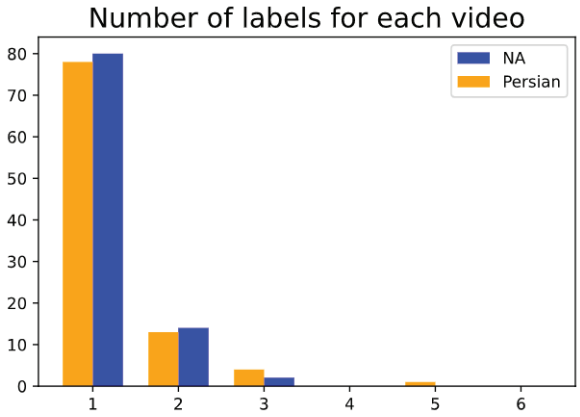


Figure 4.2: Distribution of number of emotion labels

(i.e. before voting the labels). The results are in Fig. 4.8 and Fig. 4.9, which provides us another view of the relationship between negative emotions. It is not surprising that anger, furious and hatred have a relatively high positive correlation. It is worth noting that disgust and hatred have a positive correlation in Persian dataset while their correlation is negative in NA dataset. Disgust and contempt show different patterns as well. The correlation between them is positive in Persian and negative in NA. Such differences may suggest that we should distinguish between 'social disgust' and 'physical disgust'. Social disgust might be a mix of hatred, annoyance and contempt according to the result, whereas 'physical disgust' can be considered a basic emotion. These tables also show the importance of having multiple labels for samples in order to capture compound negative feelings such as 'angry and disgusted' or 'contemptuous and disgusted', which are seen in Persian dataset labels. This heterogeneity in the "disgust" class may also explain the very low recognition scores for "aversion", previously noted as the worst recognized emotion in [18].

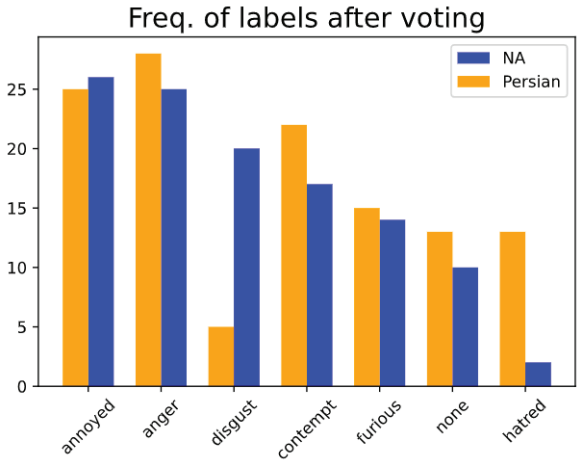


Figure 4.3: Frequency of emotion labels

	😠	😑	😒	🤬	😐	😡	😏	🙄	🤢	😩	😤	😐	🤮	😈	None
NA	16	19	19	16	12	9	5	12	11	6	2	3	6	0	7
Persian	32	2	29	10	8	15	18	8	2	12	17	3	0	1	2

Figure 4.4: Frequency of emoji labels

4.5 Action Unit Analysis

For all videos under each emotion category, we calculated the mean of peaks of each AU (one value per AU, per video) and visualized it using a radar chart in Table 4.3. This table can give a better view of differences between cultures. In our dataset, the AU values for Persian are generally smaller than NA. We see noticeable differences in the activated AUs in contempt, annoyance, hatred, and anger. The AU descriptions is in Table 4.2. We summed all values in each radar chart to quantify the difference between NA and Persian emotion expressiveness in Table 4.4. The values for NA are greater than Persian, except for None. The prototypical, high intensity expression of furious is similar across the two cultural datasets (31.52 and 31.47), but that there is less expression in the Persian dataset for milder forms of anger such as annoyance (30.42 vs. 26.86).

4.6 Social Signal Analysis

Another purpose of this thesis is to identify the key social signals for each emotion. We extracted co-occurrences of facial expressions (and body movements, such as arms crossed) with final labels, presented in Fig. 4.11 for NA and Fig. 4.12 for Persian. Each cell in the heatmap is normalized by the number of samples under each emotion class. The social signals of “mocking” in NA and “smiling” in Persian had 0 occurrences so their corresponding row is empty. Persian videos exhibited a wider variety of social signals except in disgust, which

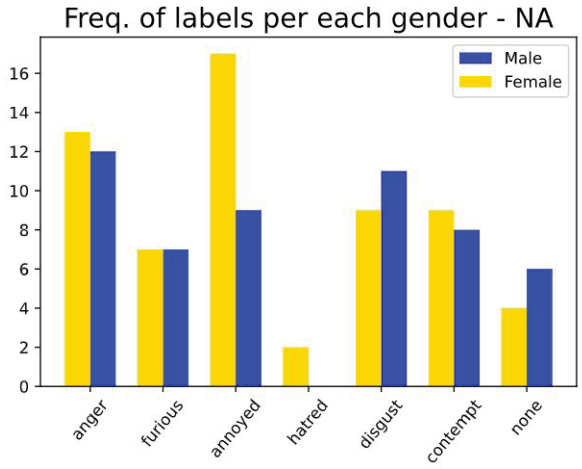


Figure 4.5: Frequency of emotion labels per each gender - NA

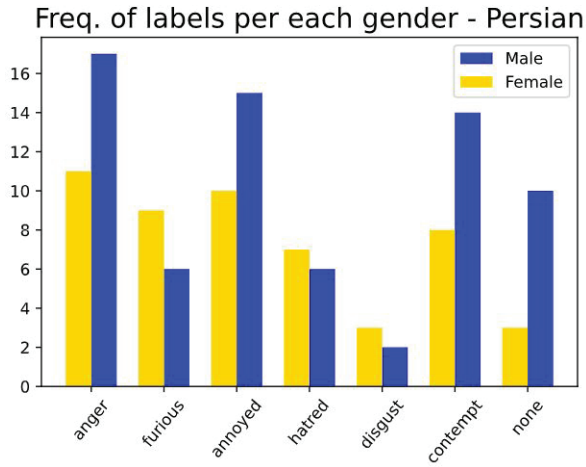


Figure 4.6: Frequency of emotion labels per each gender - Persian

may be attributed to the low number of videos in this class. “Eyebrows pushed together" was the most prevalent of all social signals in all emotions, though it was more common in Persian. In the Persian results, we notice some similarities between disgust and hatred social signals. “Raised eyebrows” is also common in NA anger and contempt but not in Persian.

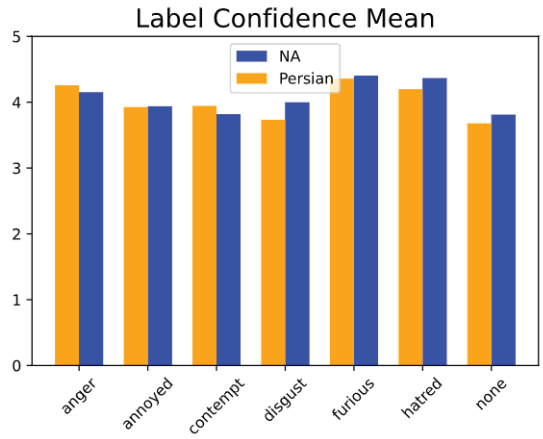


Figure 4.7: Annotators' confidence for labels

Table 4.2: Short description of AUs

AU	Description	AU	Description	AU	Description
AU01	Inner Brow Raiser	AU02	Outer Brow Raiser	AU04	Brow Lowerer
AU05	Upper Lid Raiser	AU06	Cheek Raiser	AU07	Lid Tightener
AU09	Nose Wrinkler	AU10	Upper Lip Raiser	AU12	Lip Corner Puller
AU14	Dimpler	AU15	Lip Corner Depressor	AU17	Chin Raiser
AU20	Lip stretcher	AU23	Lip Tightener	AU25	Lips part
AU26	Jaw Drop	AU45	Blink		

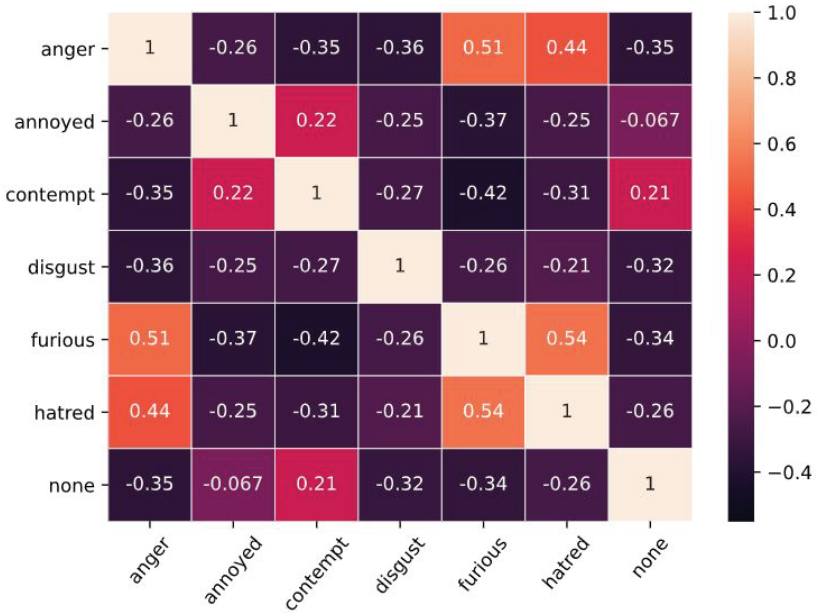


Figure 4.8: Correlation between annotators' assigned labels - NA

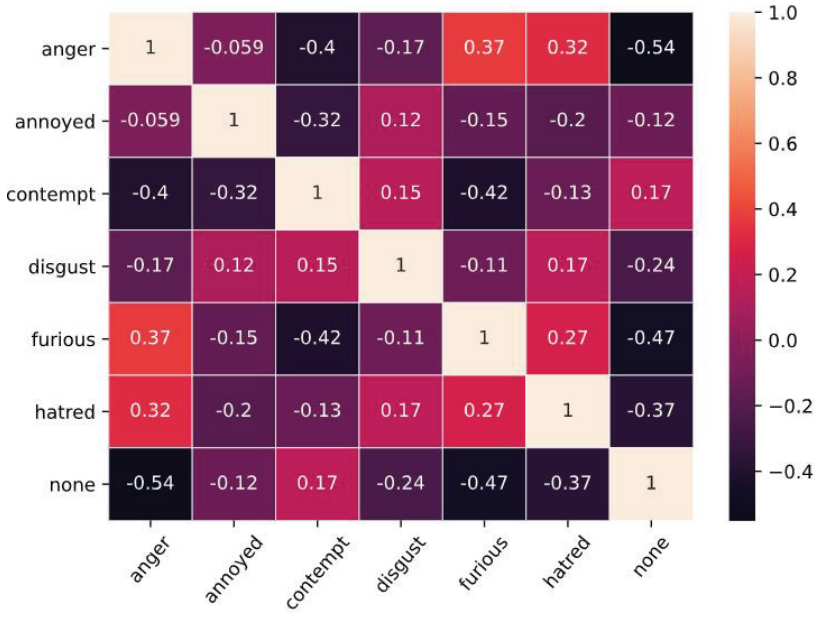


Figure 4.9: Correlation between annotators' assigned labels - Persian

	NA	Persian
annoyed	😒, 😑, 🙄, 😠	😒, 😠, 😩
anger	😠, 🤬, 😡	😠, 😤, 😡
contempt	😑, 😒, 😏	😒, 😏
disgust	🤢, 🤮, 😩	😒
hatred	😡, 😒	😠, 😡, 😤
furious	🤬, 😡	🤬, 😡, 😠
none	😐, none	😒, 😐, 😏

Figure 4.10: Most common emojis for each emotion

Table 4.3: Radar charts of the mean of the peak of AUs

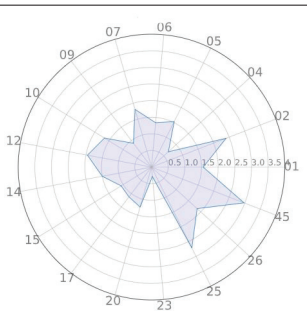
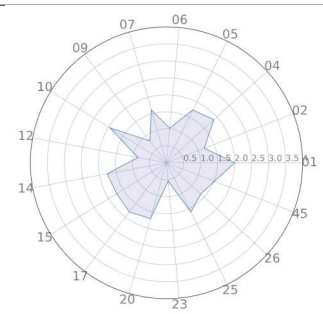
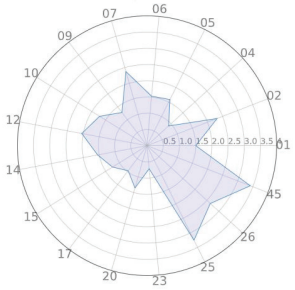
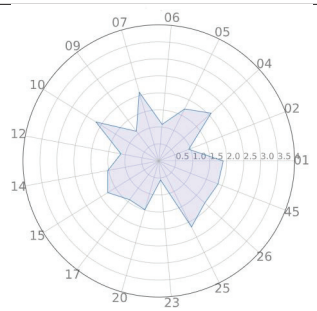
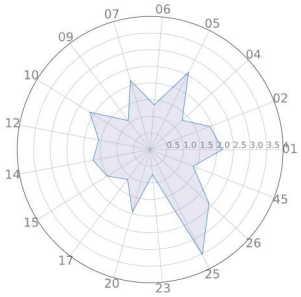
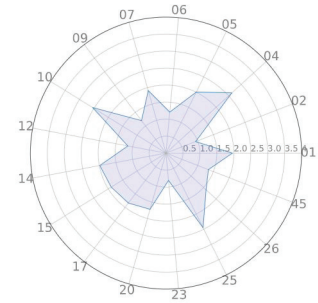
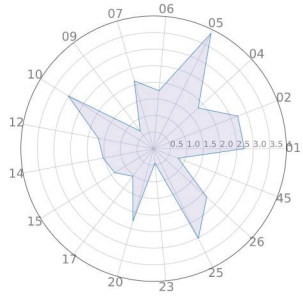
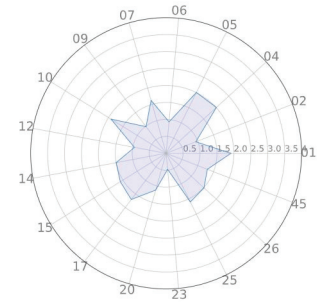
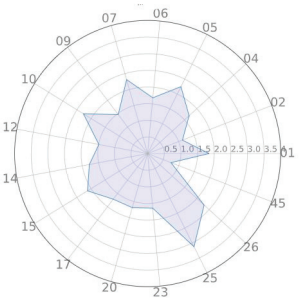
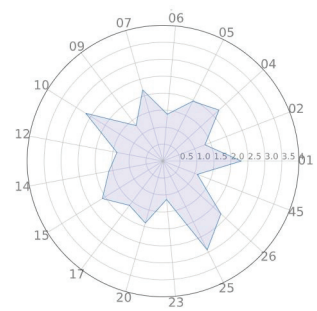
Emotion	NA	Persian
Contempt		
Annoyed		
Anger		
Hatred		
Furious		

Table 4.4: Sum of AU values in

	Annoyed	Contempt	Anger	Hatred	Furious	None
NA	30.42	26.61	30.71	33.06	31.52	20.28
Persian	26.86	25.10	29.17	24.09	31.47	26.13

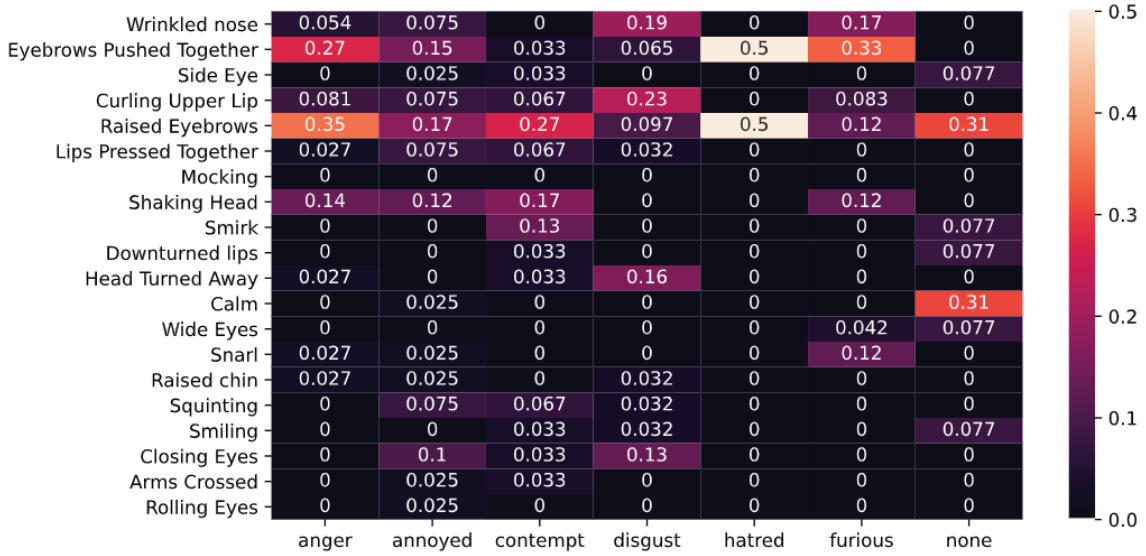


Figure 4.11: Co-occurrences of social signals for each emotion in NA videos

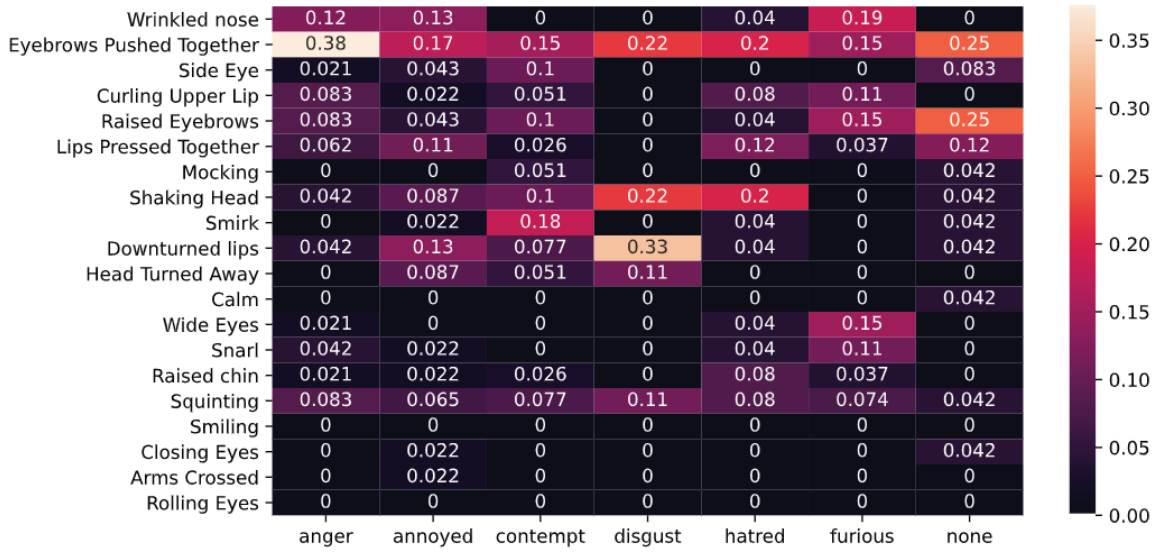


Figure 4.12: Co-occurrences of social signals for each emotion in Persian videos

Chapter 5

Experiments

Our dataset analysis showed multiple labels for many samples - especially emoji labels. Work by Du et al. [7] also showed that emotions can be mixtures, therefore, we used multi-label, multi-class classifiers for our baseline. Traditional classifiers output a single label for each sample. In order to use them for multi-label classification, we can use either of the following approaches: 1. Adapt existing algorithms to output multiple labels, 2. Transform the problem into another form, e.g. performing binary classification for each label. In the latter case, we use several classifiers that may or may not be independent of each other. We used Multi-Label K-Nearest Neighbors (MLKNN) [33] for the adaptive approach and Classifier Chains (CC) [25] for the latter.

5.1 Classifier Chains

For a given set of labels L the CC model learns $|L|$ classifiers in which all classifiers are linked in a chain through 22 features. The dataset is transformed in $|L|$ data sets where instances of j-th data set has the form $((x_i, l_1, ..., l_{j-1}), l_j), l_j \in \{0, 1\}$.

The advantage of classifier chains method is that it is capable of taking correlations between labels into account while maintaining acceptable computational complexity since the output of previous classifiers is fed into the next ones as additional features. Fig. 5.1 is an illustrated example with a classification problem of three categories $\{C1, C2, C3\}$ chained in that order.

5.2 Baseline Classification

Before doing the classification, we separated 25% of the videos for the test phase. We performed 5-fold cross-validation on the rest to obtain the best order of classifiers in chain and parameter k in MLKNN. In the classifier chains, we used the XGBoost [4] classifier, since it showed promising results on similar datasets [20]. We input all permutations of the order of labels in the classifier chain and selected the order that yielded the best result on the validation set. The inputs to the model were AUs, head rotation, and gaze angle for

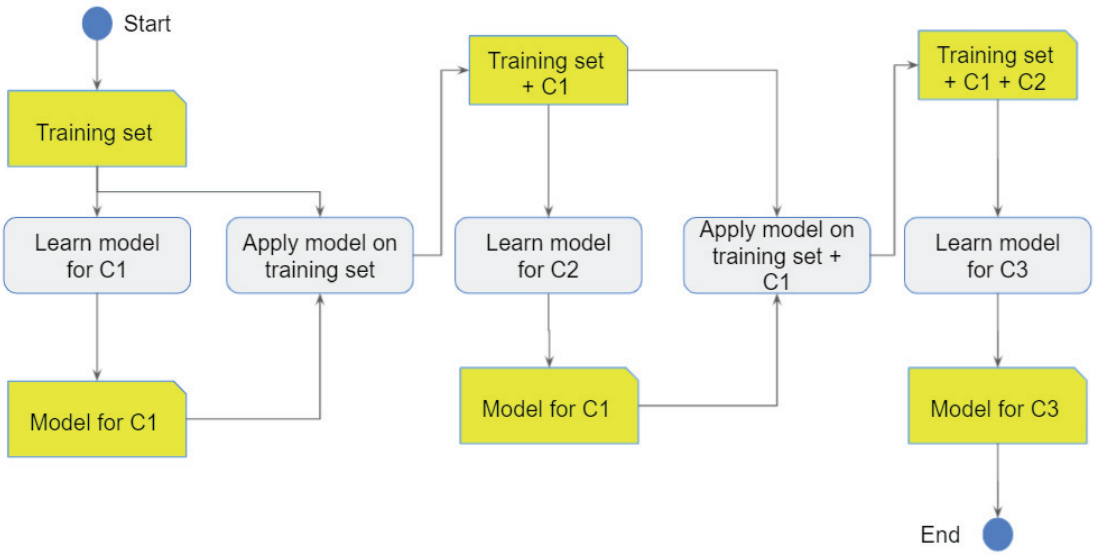


Figure 5.1: Classifier chains

each frame, and then the model outputs a set of labels for that frame. We report the sample average of F1-score for all of our models since it is widely used in multi-label classification models. We computed these metrics in two ways: 1. for each frame, and 2. assigning the label to the whole video by taking the majority of predicted labels on the frames. We also combined NA and Persian datasets for all classification models. The results are reported under Combined column in tables. We did three types of classification as follows:

Multi-Label Classification Using Emotions Words

First, we used word labels of emotions (annoyed, anger, hatred, furious, contempt, none) to perform classification. Table 5.5 shows examples of model confusion in emotion classification. The results are in Table 5.1.

Multi-Label Classification Using Emojis

In this experiment, we used 10 emoji labels to perform a multi-label classification. The results are in Table 5.2.

Hierarchical Classification Using Emojis

To find out the role of non-verbal labels, we conducted a hierarchical classification using emoji labels. First, we fed in our training and test set into the classifiers similar to what we did with emotion labels, and then mapped each predicted emoji label to its emotion class. Training and test set were the same as emotion classification. We used Fig. 4.10 to create a mapping of each emoji to its corresponding emotion. We mapped 😒 or 😏 to contempt, 🙄 or 😑 or 😩 to annoyed, 😠 to anger, 😤 or 😡 to hatred, 🤬 to furious and 😐 to none.

Table 5.1: Multi-label Classification results using 6 emotion categories

	F-F1 score			V-F1 score		
Model	NA	Persian	Combined	NA	Persian	Combined
CC - XGBoost	**0.42**	0.28	0.33	**0.42**	0.33	0.36
MLKNN	**0.42**	0.31	0.34	**0.42**	0.40	0.39

Table 5.2: Multi-label classification using 10 emoji categories

	F-F1 score			V-F1 score		
Model	NA	Persian	Combined	NA	Persian	Combined
CC - XGBoost	**0.24**	0.22	0.20	0.28	**0.31**	0.29
MLKNN	**0.27**	0.22	0.25	0.28	0.27	**0.30**

5.3 Cross-Dataset Generalization

In order to check the robustness of our proposed model, we trained a CC with XGBoost estimators and MLKNN with k=5 on ElderReact, a multi-label video dataset. Similar to our data processing, we used OpenFace to extract AUs, head pose, and gaze features. The F1-score for each label is represented in Table 5.4. The multi-label KNN model performed noticeably better for all labels except Happy and Sad. ElderReact's single label model was only marginally better in Sad category. This could suggest that using multiple labels provides a better representation of the data than single label, especially for emotions which are traditionally more poorly recognized. We believe the multilabel approach provides a new, more challenging baseline that better represents how humans perceive the videos and can be used even for samples that would otherwise be discarded for single-label classification.

5.4 Results

Overall, NA scores were better than Persian. Video-level scores did not follow a pattern; in some cases, they were more than frame-level scores and in some cases less than frame-level. This change in results from frame-level to video-level is highly dependent on test videos. If test videos have a short length (e.g. 50 frames) and the model misclassify a greater portion of its frames, it will affect video-level scores more than frame-level scores since the number of frames is far more than the number of videos (2000 frames vs 15 videos in the test set). In emotion classification where word labels are used, differences between NA and Persian were more than emoji and hierarchical classification. In hierarchical and emoji label classification, Persian's video-level F1-score is the highest of all. Combining the two datasets did not yield better results. Results ranged between the scores of Persian and NA, or less than both, except in emoji classification using MLKNN model.

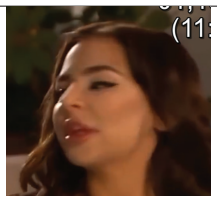
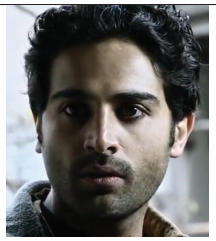
Table 5.3: Baseline of hierarchical classification using 10 emoji labels grouped into 6 labels

	F-F1 score			V-F1 score		
Model	NA	Persian	Combined	NA	Persian	Combined
CC - XGBoost	**0.34**	0.28	0.21	0.32	**0.36**	0.28
MLKNN	**0.38**	0.30	0.28	**0.39**	0.32	0.34

Table 5.4: Video-level F1-score for each emotion in ElderReact

	Happy	Surprised	Fear	Disgust	Anger	Sad
CC - XGBoost (multi-label)	0.62	0.56	0.20	0.25	0.31	0.33
MLKNN (multi-label)	0.58	**0.60**	**0.40**	**0.57**	**0.47**	0.34
ElderReact (single label)	**0.71**	0.54	0.25	0.36	0.43	**0.35**

Table 5.5: Example of misclassified frames

Frame	Actual Labels	Predicted Labels
	Anger, Furious	Annoyed
	Anger	Annoyed
	Annoyed, Contempt	None
	Anger, Annoyed	None
	Anger, Hatred	Hatred

Chapter 6

Discussion

As shown by the result of data collection, there is always some ambiguity between different emotions. We should also note that the object to which people are reacting may affect the way it is expressed. The result of hierarchical classification suggests that sometimes non-verbal labels may be a better alternative for word labels since they produced the similar results but the translation step can be skipped, especially in cross-cultural research where language might be a barrier. This is very useful in crowd-sourcing platforms like AMT, where labels usually are in English, but annotators' first language may not be. In this thesis, we cannot completely eliminate the effect of the questionnaire's language. Although people who participated in the Persian dataset were competent in English, they likely lacked a deep emotional connection with it and their affective processing may have been weaker [3].

6.1 Limitation

This thesis is not free of limitations. Firstly, it was more difficult to collect high-quality, emotionally-rich videos in Persian due to a lack of resources. For example, reaction videos akin to YouTube React channel are almost non-existent for Persian, which makes it an obstacle for this type of research on low-resource cultures. Another limitation is that North American culture is a mix of other cultures such as Asian American or European American due to immigration and these sub-cultures show differences in emotion expression as well [30]. Moreover, there are variances of expressions within cultures that we may not be able to capture in the dataset because they are different from one person to another and depend on the personality. Another challenge with collecting in-the-wild datasets is the change between the scenes or lighting conditions. We had to discard some of the videos because the camera moved rapidly from one person to another in a hot quarrel. OpenFace would also fail to detect faces in dark scenes. Additionally, we had label-imbalance in our dataset, both within labels and genders, that can negatively affect the result. The solution to this problem may be iterative data collection, that is, collecting more data and annotation to balance the labels

in second round. Another technical limitation is the features that we extracted through OpenFace because it only extracts 17 AUs out of 30 main AUs. Nevertheless, the extraction of AUs made it straightforward to have an analysis of differences in expressiveness. The limited number of videos in hatred category for NA and disgust for Persian would also make the analysis slightly inaccurate. For example, the AU values in NA's hatred radar chart were greater than furious, which is the high arousal form of anger, and it needs more data for a more precise analysis. It also adds to the skewness of the predictive models. Finally, emotional expressiveness may affect the result of classifications. Research suggests that more conservative cultures do not express negative emotions such as anger blatantly. We found supporting evidence in the radar charts of Table 4.3 that activated AUs in Persian have smaller values. This adds to the complexity of predictive models training on more conservative cultures like Persian.

6.2 Implications for Future Affective Computing Research

This research clarified some problems in the affective computing field. We found out that we should distinguish between social disgust and physical disgust in data collection. Also, in-the-wild acted videos (e.g. movies) should be separated from spontaneous in-the-wild videos (e.g. vlogs). Researchers should consider discrepancies in labels to be important multi-labels rather than the noise that they throw out.

Chapter 7

Conclusion

7.1 Conclusion

The main contribution of this thesis is the collection of a multi-cultural dataset of videos annotated with (a) affect labels under anger category, (b) their associated emojis, (c) social signals for building more robust emotion recognition models for underrepresented cultures. We conducted statistical analyses to find the underlying expressions of each emotion and compare NA and Persian cultures in terms of emotion expression. Moreover, we provided multi-label classification baseline models that demonstrated how emojis can be used instead of or in addition to word labels. In order to examine the effectiveness of non-verbal labels, we built a similar model, this time with emoji labels. This opens the opportunity to language-agnostic labels, especially in cross-cultural emotion studies.

7.2 Future Work

We collected the presented dataset with an intention to classify them as a set of features varying over time (i.e. a multi-feature time series). However, there is no machine learning model that is adaptable to multi-label classification of multi-feature time series. Hence, a great improvement on the computational aspect of this field would be designing multi-label classifiers that are capable of handling 3-dimensional data (time, features, and samples). We can also design algorithms that are trained on a specific culture, but are able to adapt to personal variances, using a smaller dataset such as the data collected during interaction with the person. We are also interested to extend the approach presented here to positive affects such as joy and related emotions such as happiness, surprise, and cheerfulness. Using emojis in affective computing is also novel and researchers may investigate the perceived meaning of them in different groups, based on age, gender, or culture. Future work could compare the result using English labels fully translated to Persian (and back-translated to ensure accuracy). Augmenting the dataset with more videos (synthetically or generated) will allow us to use transfer learning and deep neural network algorithms and investigate

their effectiveness. However, we should note that simply transferring AUs to a generated or synthesized video will not be enough, and the resulted videos should be photo realistic. We also collected videos from Filipino culture, but due to the low number of annotators, we omitted them. We look forward to applying the method to Filipino and other cultural datasets. Training the model on one culture and testing it on another culture can also be a matter of investigation.

Bibliography

[1] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[2] Gibran Benitez-Garcia, Tomoaki NAKAMURA, and Masahide KANEKO. Multicultural facial expression recognition based on differences of western-caucasian and east-asian facial expressions of emotions. *IEICE Transactions on Information and Systems*, E101.D:1317–1324, 05 2018.

[3] Catherine L. Caldwell-Harris. Emotionality differences between a native and foreign language: Implications for everyday life. *Current Directions in Psychological Science*, 24(3):214–219, 2015.

[4] Tianqi Chen and Carlos Guestrin. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016.

[5] Weixuan Chen, Ognjen Oggi Rudovic, and Rosalind W. Picard. Gifgif+: Collecting emotional animated gifs with clustered multi-task learning. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 510–517, 2017.

[6] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep — a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964, 2014.

[7] Shichuan Du, Yong Tao, and Aleix M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[8] Itziar Fernández, Pilar Carrera, Flor Sánchez, Darío Páez, and L. Candia. Differences between cultures in emotional verbal and nonverbal reactions. *Psicothema*, 12:83–92, 2000.

[9] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. In Minho Lee, Akira Hirose, Zeng-Guang Hou, and

Rhee Man Kil, editors, *Neural Information Processing*, pages 117–124, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[10] Geert Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture, Unit 2*, 2(1), 2007.

[11] Shinnosuke Ikeda. Influence of color on emotion recognition is not bidirectional: An investigation of the association between color and emotion using a stroop-like task. *Psychological reports*, 123(4):1226–1239, 2020.

[12] László Jeni, Jeffrey Cohn, and Fernando De la Torre. Facing imbalanced data - recommendations for the use of performance metrics. volume 2013, 09 2013.

[13] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017.

[14] Trinh Khanh, S.H. Kim, Gueesang Lee, Hyung-Jeong Yang, and Eu-Tteum Baek. Korean video dataset for emotion recognition in the wild. *Multimedia Tools and Applications*, 80:1–14, 03 2021.

[15] Pooya Khorrami, Tom Le Paine, Kevin Brady, Charlie Dagli, and Thomas S. Huang. How deep neural networks can improve emotion recognition on video data. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 619–623, 2016.

[16] Yelin Kim and Jeesun Kim. Human-like emotion recognition: Multi-label learning from noisy labeled audio-visual expressive speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5104–5108, 2018.

[17] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition, 2018.

[18] Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2019.

[19] Nangyeon Lim. Cultural differences in emotion: differences in emotional arousal level between the east and the west. *Integrative Medicine Research*, 5(2):105–109, 2016.

[20] Kaixin Ma, Xinyu Wang, Xinru Yang, Mingtong Zhang, Jeffrey M Girard, and Louis-Philippe Morency. Elderreact: a multimodal dataset for recognizing emotional response in aging adults. In *2019 International Conference on Multimodal Interaction*, pages 349–357, 2019.

[21] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected "in-the-wild". pages 881–888, 09 2013.

[22] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, Jan 2019.

[23] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[24] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E. Hughes, and Louis-Philippe Morency. Emoreact: A multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 137–144, New York, NY, USA, 2016. Association for Computing Machinery.

[25] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. In Wray Buntine, Marko Grobelnik, Dunja Mladenić, and John Shawe-Taylor, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 254–269, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[26] Paul Rozin, Laura Lowery, Sumio Imada, and Jonathan Haidt. The cad triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, 76(4):574–586, January 1999.

[27] Ghazal Saheb Jam, Jimin Rhim, and Angelica Lim. Developing a data-driven categorical taxonomy of emotional expressions in real world human robot interactions. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, Mar 2021.

[28] Anna Schouten, Michael Boiger, Alexander Kirchner-Häusler, Yukiko Uchida, and Batja Mesquita. Cultural differences in emotion suppression in belgian and japanese couples: A social functional model. *Frontiers in Psychology*, 11:1048, 2020.

[29] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks, 2020.

[30] Jeanne L Tsai, Heather Mortensen, Ying Wong, and Dan Hess. What does" being american" mean?: A comparison of asian american and european american young adults. *Cultural Diversity and Ethnic Minority Psychology*, 8(3):257, 2002.

[31] Job Van Der Schalk, Skyler T Hawk, Agneta H Fischer, and Bertjan Doosje. Moving faces, looking places: validation of the amsterdam dynamic facial expression set (adfes). *Emotion*, 11(4):907, 2011.

[32] Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[33] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

Appendix A

Code

The dataset and source codes are available at this link.