Developing a Data-Driven Categorical Taxonomy of Emotional Expressions in Real World Human Robot Interactions

by

Ghazal Saheb Jam

B.Sc., University of Tehran, 2019

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:	**Ghazal Saheb Jam**
Degree:	**Master of Science**
Thesis title:	**Developing a Data-Driven Categorical Taxonomy of Emotional Expressions in Real World Human Robot Interactions**
Committee:	**Chair:** Mo Chen Assistant Professor, Computing Science
	Angelica Lim Supervisor Assistant Professor, Computing Science
	Parmit Chilana Committee Member Associate Professor, Computing Science
	Philippe Pasquier Examiner Associate 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.

Abstract

Emotions are reactions that can be expressed through a variety of social signals [14] [23]. For example, anger can be expressed through a scowl, narrowed eyes, a long stare, or many other expressions. This complexity is problematic when attempting to recognize a human's expression in a human-robot interaction: categorical emotion models used in HRI typically use only a few prototypical classes, and do not cover the wide array of expressions in the wild. We propose a data-driven method towards increasing the number of known emotion classes present in human-robot interactions, to 28 classes or more. The method includes the use of automatic segmentation of video streams into short (<10s) videos, and annotation using the large set of widely-understood emojis as categories. We then investigate the meaning behind these emojis by studying how humans perceive these emojis. We showcase our results in a taxonomy which includes each emoji and the different meanings people perceived from it. This framework for social signal analysis can be used in the future by researchers to capture what social signals are happening in the wild. Furthermore, researchers can use emojis as social signal representation labels for training machine learning models, towards more accurate human emotion recognition by robots.

Keywords: emotion-recognition, social signals, in-the-wild, emotion categories, emojis

Acknowledgements

During the writing of this thesis I have received a great deal of support and assistance. I would like to thank my supervisor, Dr. Angelica Lim, who has provided invaluable insight and guidance. I would like to express my sincere gratitude to Dr. Mo Chen for supporting me throughout my degree. I would like to express my appreciation to Dr. Parmit Chilana and Dr. Philippe Pasquier for accepting to being my committee members and for providing insightful suggestions.

I would like to thank Dr. Jimin Rhim for her friendship and advices that she have shared during the project. Also I would like to express my infinite and endless gratitude to my family for always supporting me in life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Robots must be able to recognize, interpret, and respond effectively to social cues displayed by a human interactant in order to participate in bi-directional social interactions with humans [58]. A robot capable of interpreting emotion will have a greater capability to make decisions and aid humans because of its ability to respond to human non-verbal cues [58]. Such robots would encourage more effective and interesting interactions with users, resulting in a higher level of acceptability of these robots among users [58]. For instance consider this scenario: a social robot is placed in a public space and people can interact with it. One person asks the robot a question and the robot responds. The person shows confused face as they did not understand the answer properly. If the robot is able to recognize this social signal, it can act accordingly. Humans show a wide variety of social signals in everyday conversations. However, previous work in human robot interaction (HRI), has mostly focused on 6 prototypical human affect states (happiness, surprise, fear, disgust, anger, and sadness) and do not cover this wide variety. As Barrett et al. suggest there is an urgent need for research into how people actually move their faces to express emotions and other social information in the various circumstances that make up everyday life, as well as a thorough examination of the mechanisms by which people perceive instances of emotion in one another [8]. This thesis investigates two main research questions: First, what social signals do humans make when interacting with a social robot in a public space? and second, what is the meaning behind all these social signals?

1.1 Contributions

We aim to address these two important issues by proposing a novel data-driven framework for social signal analysis in HRI. First, we contribute a data-driven method to segment, analyze and annotate streams of video to capture what dynamic social signals people make during interactions with a social robot in public. Here, we use challenging human-robot interaction data captured from a robot's head, but the technique can be applied to any front-facing videos. In addition, we contribute the first use of emojis as representations of social signals for videos. This is advantageous since emojis have been found to be mostly cross-lingual and are used every day in-the-wild [2]. We then provided an initial taxonomy of social signals in the wild in HRI. Our taxonomy includes a wide variety of social signals humans make when interacting with a social robot in public in HRI and the rich collection of meanings behind each of these social signals. We contribute a finding that confusion was a common social signal in our in-the-wild HRI dataset. The reason for this may be the fact that most of the participants were interacting with the robot for the first time. We also found that although most of the social signals we collected were rated as from one specific emotion category (from Love, Fear, Joy, Anger, Fear, Neutral, Sadness, Surprise categories), some of them were from two or more categories. Based on our experiments, we provide concrete distributions of different ways humans perceive these social signals (see Appendix).

1.2 Thesis Outline

This thesis is organized as follows. Chapter 2 gives a brief overview of related work. Chapter 3 examines the question of what dynamic social signals people make when interacting with a social robot in public by collecting and annotating a rich set of social signals in the wild. Chapter 4 investigates how humans perceive these collected social signals. Chapter 5 organizes the results of both experiments in a final taxonomy of social signals in the wild in HRI. Finally, our conclusions and implications of this work are drawn in the Chapter 6.

Chapter 2

Related Work

2.1 Emotion models

Researchers proposed affect representation models to describe how people perceive and classify affect. Generally, there are three main types of affect representation models: categorical, dimensional, and hybrid.

2.1.1 Categorical models

Categorical models include discrete states, each representing a particular affect category [68]. Initial work on categorical models of emotion started with Darwin. Darwin considered emotions as separate discrete entities, or modules, such as anger, fear, disgust, etc [31]. Ekman focused primarily on six basic affective states. He claimed that affective states were happiness, sadness, fear, disgust, anger, surprise (and possibly contempt). [29] [42] [30] [24]. Later, researchers added neutral to these categories [77] [36] and generated new categories by combining these basic categories [27].

Plutchik proposed eight primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust and joy, and arranged them in a color wheel. Half of these emotions are positive, while the other half are negative [40]. They are seen as opposite to each other such as joy being opposed to sadness, surprise being opposed to anticipation, trust being opposite to contempt, and rage being opposite to fear. In a wheel-shaped mechanism, he discussed each emotion in detail and separated it into subgroups, treating them as secondary and tertiary emotions. His work depicts an intriguing link between emotions, their intensities, and their polarities. He also noted that the strength of an emotion is greatest when it is at the centre of the wheel and decreases as one moves away from the centre [1].

Furthermore, Shaver proposed a hierarchical structure for the emotion domain to organize the interrelated set of emotion categories [71].

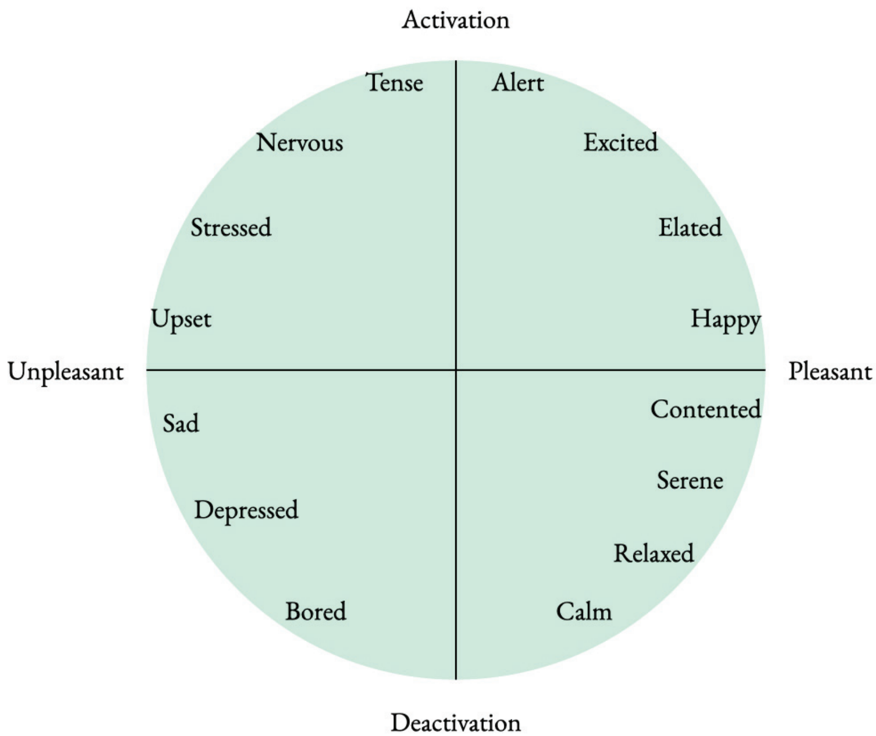


Figure 2.1: Russel Circumplex Model of Emotion (Reproduced) [66]

2.1.2 Dimensional models

Dimensional models define emotional expressions as a feature vector in a continuous space. [58]. One of the most recognized dimensional models is Russell's circumplex model of emotion [66]. He argued that there is a link or conflict between emotions. In this model, two orthogonal dimensions known as pleasure and arousal are used for affect categorization as shown in the Figure 2.1 [81]. As a result, the following eight emotional terms were collected from the preceding literature and placed in two dimensions to define emotions: happy and excited, amazement, disappointment, misery, depression, fatigue, and satisfaction.

Another study defined the Pleasure-Arousal-Dominance (PAD) model which consisted of three nearly orthogonal dimensions. Dominance described whether the person feels controlling or controlled [60]. Anger, for example, can be defined as having a negative valence, a high arousal, and a high dominance, whereas fear has a negative valence, a high arousal, and a low dominance, and sadness has a negative valence and a low arousal and dominance [49].

Dimensional models are often harder to understand than categories. For example, [21] states that "categorical labels such as amusement better capture reports of subjective experience than commonly measured affective dimensions."

2.1.3 Hybrid Models

An example of a hybrid model is the Geneva Emotional Wheel [69], which combines dimensional and categorical models. This model consists of 20 distinct emotion families. This model also has two dimensions corresponding to valence (horizontal axis) and control (vertical axis) [22]. Geneva Emotional Wheel arranges emotion terms in two dimensional space and the intensity of the associated subjective feeling is distance from origin.

Although all of these representations are beneficial for improvements and progression in the HRI field, a number of considerations remain to be addressed.

2.1.4 Emotion models used in HRI

An increasing number of HRI studies have used categorical models for classifying emotional expressions such as facial expressions [78] [54] [25] [52], body language [59], etc. [70]. However, it is now accepted that the current set of categories are too limited to adopt for real-world emotion recognition of humans [8]. As stated in a recent review, "humans can display a large variety of affective states during social interactions, and affect-detection systems that can identify only a small set of states (e.g., two to five) will not allow a robot to effectively participate in natural bi-directional affective communications with humans [58]."

2.2 Emotional expressions in videos

2.2.1 Emotional expressions representations

Emotional expression representations include techniques to show a facial configuration or non-verbal behavior that displays a human's emotion.

Facial action coding system

FACS (Ekman, Friesen, & Hager, 2002) is a systematic way of describing what a face appears like when facial muscle movements have happened [8][28]. The presence and intensity of facial movements are described by FACS [1] codes [8][28]. It divides face expressions into Action Units, which are discrete components of muscle activity (AUs) [8][28]. For example, AU2 corresponds to outer brow raise. Although AUs are scored and analysed as distinct pieces, the underlying architecture of many face muscles prevents them from moving independently of one another, resulting in AU interdependence [8]. Coders can manually code all potential face displays into action units (AUs) [8][28].

[1] https://www.paulekman.com/facial-action-coding-system/

Affectemes and allaffects

Axelrod et al. provided a method for coding emotional and expressive behaviours during computer interactions to HCI researchers [5]. They modeled their approach based on the way that spoken language is analysed: by using different levels of analysis [5]. They identified distinctive affect-related behavioural events, as ‘affectemes’ [5]. They then considered the variations or ’allaffects’ displayed by the participants, since there may be variations of exact representations of these affectemes [5]. In the first-level coding they asked participants to time-stamp any points where they perceived any communicative signals from the videos of interactions with computer [5]. These videos included transcripts [5]. In second-level, using a series of seven passes the participants looked for incidences of audible activity, whole body movements, head movements, upper face movements, lower face movements, gaze and blink patterns, and activities such as keyboard entry and screen activity, which were logged by spyware and pasted into the relevant transcript [5]. They used keywords and keyword families [5]. In the third-level of this top-down approach they coded variations on an affecteme and episodes could be further categorised within the keyword family [5].

Emojis

Emoji is a Japanese term that means "picture characters." Emojis, unlike emoticons, are images rather than characters. Shigetaka Kurita invented them in the late 1990s using Unicode characters. They were intended to be an evolution of eastern emoticons, representing not only faces but also concepts, objects, and so on [41]. Emojis have a number of different roles such as sentiment enhancement, sentiment expression and sentiment modification [41]. One of their primary applications has been to convey emotion, particularly through facial expression emojis [72]. Emojis are a visual simplified form of (affective) communication that increases the amount of information (e.g., cues and gestures) that can be shared between humans and virtual/embodied artificial entities [67]. Emojis have undoubtedly become a part of mainstream communication around the world since their beginnings, allowing people of different languages and cultural backgrounds to share and interpret ideas and emotions more accurately [67]. In this vein, it has been proposed that emoji will become a universal language due to their universal communication features and ever-expanding lexicon [67]. However, this notion is controversial, since emoji usage during communication is influenced by factors such as context, user interrelationships, users’ first language, repetitiveness, socio-demographics, and so on [67]. Previous research used emojis in several ways. Al-Halah et al. used emoji as low-dimensional embedding of images in the emoji space for visual sentiment analysis [2]. Marengo et al. explored whether emojis can replace text-based items in personality assessments [56]. They found that "the emoji were significantly related with the traits that have shown the most consistent links with emotions and affective processing" [56]. They concluded that emojis might be utilised to create a language-free assessment tool for personality [56]. To the best of our knowledge, until now using emojis as social signal representations has only been applied to images and not videos.

2.2.2 Emotional expressions in naturalistic settings

One important limitation that should be addressed is that laboratory conditions have been more widely used up to this time in emotion studies, and the majority of HRI studies use prototypical emotion classes [46]. Matsumoto et al. utilised FACS to code facial behaviors of medal winners of the judo competition at the 2004 Athens Olympic Games [57]. Winners' spontaneous expressions were captured as soon as they finished their medal matches, got their medal from a dignitary, and posed on the podium [57]. Benitez-Quiroz et al. tested the ability of computer vision algorithms to analyse a significant number of photos of facial expressions in the wild automatically [11]. First, they tested the system to detect 11 AUs. Then, they tested the system to recognize 16 basic and compound emotion categories [11]. The most frequent discrete affect states used by robots in HRI settings are disgust, sad, surprise, anger, disgust, fear, sad, happy, surprise, and neutral [68]. We therefore aim to develop a new taxonomy based on the data of humans interacting with a robot in the wild, to incorporate naturalistic settings. To this end, we use a dataset containing recordings of participants interacting with a social robot in a *real world public space*.

2.2.3 Dynamic vs. static emotional expressions

Another limitation is that although some HRI studies use dynamic data [12] [79], many HRI studies performing facial expression recognition use static and frozen in time images of prototypical emotional expressions [53] [18]. These images only show part of the behaviour. A key feature of emotional expressions is their dynamic nature [64]. It seems that using dynamic information is more helpful when expressions are low intensity. In addition, dynamics enable us to observe how expressions change over time. It also helps to differentiate between authentic and fake expressions [48]. Therefore in this study we focus on short video clips ($\leq$ 10s) instead of images to preserve the temporal sequence of emotional expressions.

2.2.4 Perceiving emotions from emotional expressions

For almost a century, academics have investigated whether humans can reliably and explicitly interpret emotional meaning from facial expressions [8]. The majority of tests on emotion perception ask participants to infer emotion from images of posed facial configurations [8]. In most cases, the configurations were posed by persons who were not at an emotional state at the time the images were taken [8]. Recent advances in emotion perception measurement make use of in-the-wild data from the internet [62], or computer-generated faces or heads rather than pictures of staged human features [8]. One technique, known as reverse correlation, assesses participants' internal models of emotional expressions (i.e., their mental representations of which facial configurations are likely to express instances of emotion) by observing how they label an avatar head that displays random combinations of animated facial action units [80][43][8].

Perceiving emotions from emojis

The emergence of emojis as a novel, more visual method of expression has been a major development [72]. The popularity of emojis appears to be mostly due to their capacity to convey emotion [72] [75]. This is demonstrated by the fact that the most commonly used emojis are smileys and other facial expression symbols with a direct link to emotional expression [72]. Many attempts have been made in order to investigate relationship between emojis and emotions. Smailović et al. proposed a emoji sentiment ranking and developed sentiment map of the 751 most frequently used emojis [73]. The sentiment of emojis was computed from sentiment of tweets by polarity(negative, neutral, or positive) [73]. Another study classified a sample of 15 emojis into four different emotion categories [65]. Barbieri et al. explored usage and meaning of emojis over different languages [7]. Their result indicated that the overall semantics of the subset of emojis investigated are intact across all languages studied. Some emojis, however, are understood differently from language to language, which may be due to socio-geographical factors [7]. Shoeb et al. conducted an experiment to obtain real-valued emotion intensity scores for 150 most popular emojis on Twitter [72]. The results were collected for eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust from the Plutchik wheel of emotions model [72]. Annamalai et al. attempted to investigate 210 undergraduates' interpretations of 75 emojis in WhatsApp Messenger [4]. However, not all of these words are emotional.

Chapter 3

Data-driven discovery of social signals

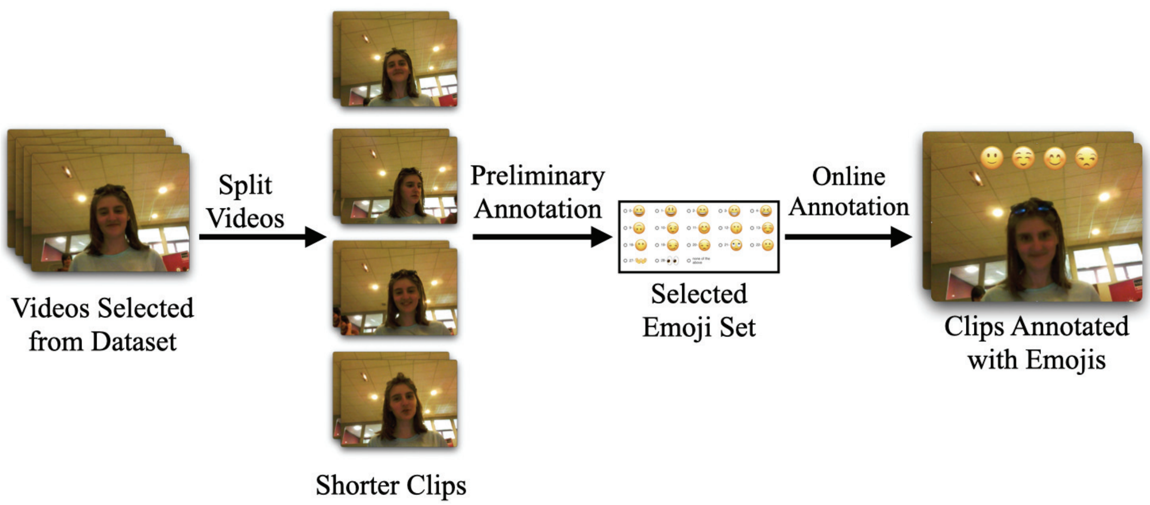


Figure 3.1: Proposed data-driven method to discover human expressions in HRI in the wild

In this chapter, given an in-the wild dataset that consists of recordings of humans interacting with a social robot, we aim to find a new categorical taxonomy of social signals in HRI utilizing emoji-based annotations. Humans use emojis to add non-verbal behavioral cues to text-based communications, and also emojis are widely available in many languages and they are cross-lingual [19].

This process involves three steps.

- First, we segment and pre-process the raw data streams into short video clips.
- Second, we conduct a pre-selection phase by asking researchers to label clips by assigning emojis to each clip.

Figure 3.2: Example of UE-HRI dataset Interactions

- Finally, we run a validation phase asking online annotators to annotate clips using the emojis selected in the previous step as a basis.

Several studies have introduced multimodal human-robot interaction datasets [45], [74], [15] [50]. One of the serious limitations of these studies is that these studies are conducted in the lab. We are most interested in in-the-wild human-robot interaction datasets, such as [13] which consists of 133 interactions which are not publicly available, and [35] [51], two publicly available in-the-wild datasets. While our ultimate goal is to analyze a broad range of datasets, as a first step, we select an in-the-wild, publicly available dataset which focuses on social signals, the User Engagement in HRI (UE-HRI) dataset [9].

This dataset contains 195 recordings (54 among them are publicly available) of participants interacting with the Pepper robot. The robot is placed in a public space and participants are free to start and end the interaction whenever they want. The recording starts automatically when the robot recognizes the presence of a person. If the participant confirms the agreement presented on the robot's tablet, the recorded data is sent to the local server. First, the Pepper robot presents itself and gives instructions. Then, the robot asks questions from participants and talks about their favorite book or movie. The robot records the interaction using all of Pepper's data streams, which are then packaged into the Robot Operating System (ROS[1]) framework's ROSbag[2] file format. The dataset consists of 54 interactions (36 males, 18 females), where 32 are mono-user and 22 are group interactions [9].

3.1 Data preparation method

Since the UE-HRI dataset contains raw data streams recorded by the robot packaged into ROSbags, we wrote a script that extracts synchronized front images and turns the image

[1]http://wiki.ros.org/

[2]http://wiki.ros.org/rosbag

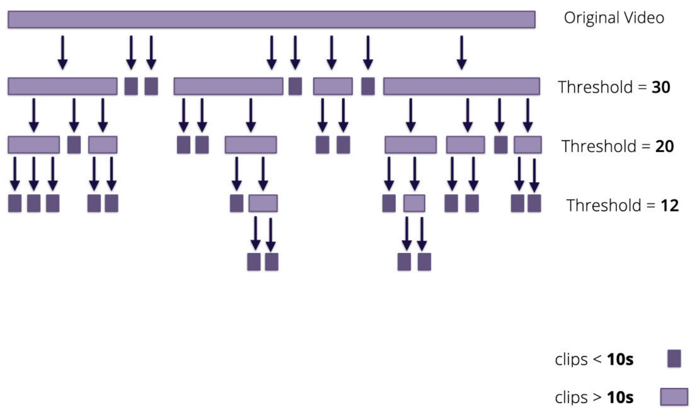


Figure 3.3: Video segmentation process

sequence into a video (excluding audio) using ffmpeg [3] [10]. In order to avoid erroneous data, e.g. participants working with the tablet instead of talking to the robot, participants taking video of the robot, etc., we then manually selected 27 of the 54 videos of interactions; in these videos, participants completed the scenarios, meaning they did not leave the robot unexpectedly. The combined length of these 27 videos was 278.7 minutes. In order to ease the final annotations, we performed a semi-automatic segmentation of the videos into clips, each containing, ideally, a single emotional expression. We aimed for clips of less than 10 seconds because upon manual inspection we found that clips longer than 10 seconds frequently contained several emotional expressions. The scene-detection method used for automatically splitting videos into clips was PyScene-Detect [4]. This library has a content-aware detection mode which finds moments where the difference between two subsequent frames exceeds a threshold value and cuts the video at these points. We first used this tool with threshold $T = 30$ which is the default threshold. In the next step, we removed clips with one frame and clips where the face was not visible, resulting in 627 clips. Since there still remained 156 clips longer than 10 seconds, and these long clips often showed more than one social signal, we repeated this process with threshold $T = 20$ resulting in 351 clips, with 69 videos longer than 10s. As a final step we applied threshold $T = 12$ on the remaining 69 long clips. Overall, we acquired 842 total clips of less than 10 seconds, with a total combined length of 77.55 minutes, corresponding to 27% of the 27 selected videos from dataset. The process of choosing optimal threshold values was done manually through trial and error.

[3] http://www.ffmpeg.org

[4] https://github.com/Breakthrough/PySceneDetect

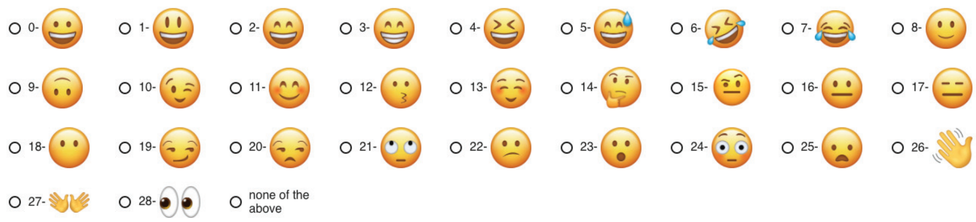


Figure 3.4: Selected emoji set

3.2 Preliminary study: Emoji selection

In this work, we target emotional expressions as a starting point, but are interested in all social signals used for communication between humans and robots. Instead of relying on traditional emotion models, we use a set of social signals in use every day by people around the world: emojis, otherwise known as smileys. Humans use emojis to add non-verbal behavioral cues to text-based communications, and also emojis are widely available in many languages and they are cross-lingual [19]. "emoji serve as a visual simplified form of (affective) communication that broadens the total amount of information (e.g., cues and gestures), which can be shared between humans and virtual/embodied artificial entities." [67]

The Unicode Emoji set[5] contains a rich palette of expressions of over 100 face and hand gestures that are frequently updated as new communicative needs are established. In [32], a machine learning system was proposed to convert images of faces into 16 different emojis. Building on this idea, our aim is to discover which of the many other emojis can be used as a representation and annotation tool for dynamic, in-person interactions with robots. The Unicode emoji dataset contains 150 emojis of body and face. In order to decrease cognitive load on the annotators, we performed pre-selection of emojis. As mentioned in Section 2.2, in the first step we manually selected 27 of 54 interactions. In this step, we randomly selected up to 3 clips (after the pre-processing phase, some of the videos had only one remaining clip) from each of these 27 videos in order to cover a broad selection of interactions. The result was 61 clips.

Each clip was annotated independently by two researchers given the full Unicode Emoji set. They selected up to 5 emojis from the given set and specified the confidence of their choice based on a Likert-scale from 1 (not confident) to 5 (very confident). We then aggregated the data into a set of emojis containing all of the emojis both researchers used to annotate videos. The obtained set comprised 29 emojis. We computed the inter-rater agreement between two researchers using Cohen's kappa [20] measure. The agreement was 0.39 which can be interpreted as fair agreement according to [20] [76].

[5]http://unicode.org/emoji/charts/full-emoji-list.html

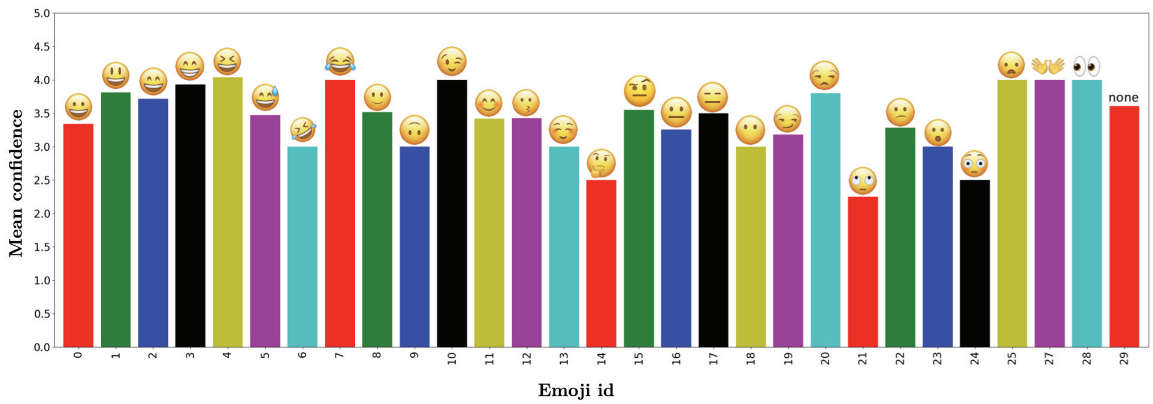


Figure 3.5: Mean confidence of each emoji obtained by annotations in Mturk

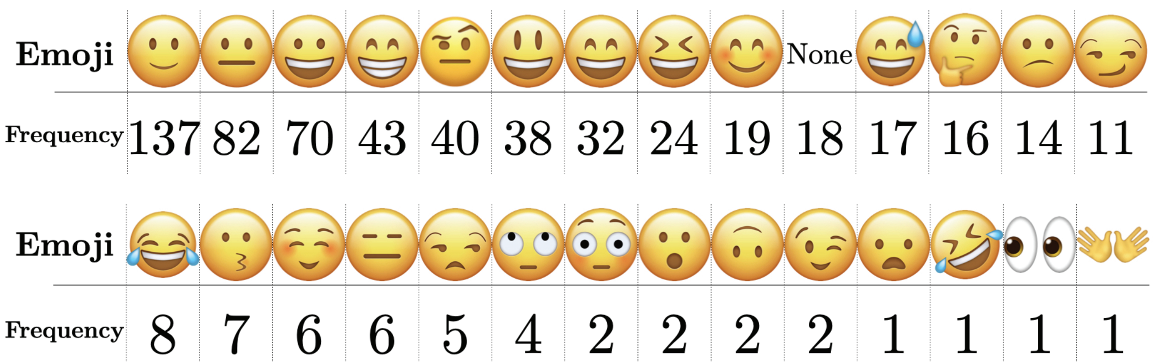


Figure 3.6: Frequency of emojis obtained by annotations in MTurk

3.3 Experiment 1: Social signal annotation

3.3.1 Method

Annotators were recruited from Amazon Mechanical Turk (MTurk). MTurk is a crowdsourcing website which allows researchers to pay a fee to the MTurk workers to complete Human Intelligence Tasks (HITs). HITs may include surveys or any other work that requires human intelligence. Previous studies have shown that although MTurk samples are not representations of the general population, these samples tend to be more diverse with respect to education and age than college samples [47]. In this experiment, we posted a HIT (Human Intelligence Task) on MTurk containing the 61 clips that researchers had previously annotated. To increase the likelihood of high quality annotations, we set our requirements to a minimum HIT approval rate of $\geq 95\%$ (percentage of completed work that has been approved by other requestors) and having approved HITs ≥ 5000. We required all of our annotators to be located in Canada for ethical board approval. The assignment duration for each task was set to 1 minute and each video clip was annotated by 10 unique annotators. Annotators first provided informed consent about study details and procedure. Then they viewed each of these clips along with the set of 29 emojis obtained from Section 2.4. They were asked to choose one emoji that best matches what the person is expressing to the camera and also specify confidence of their choice based on the Likert-scale from 1 (not confident) to 5 (very confident). They could also write the ID of more than one emoji if they observed multiple expressions, or they could describe the expression if it was not available. Each annotator was paid per HIT such that their per hour rate was equivalent of the minimum wage in their country. The final number of annotations was 610.

3.3.2 Results

The annotators used all of the 29 emojis in the provided emoji set except the waving hand (👋).
As shown in Figure 3.6 the results demonstrate that most common emoji is (🙂) which is a slightly smiling face and the second most common emoji is (😐) which is the neutral face.
Figure 3.5 shows the mean of confidence of each emoji. It can be seen in this figure that the 6 emojis that annotators were most confident about were (😆, 😂, 😉, 😦, 👐, 👀) and 6 emojis annotators were least confident about were (🙄, 😳, 🤔, 😮, 😶, ☺️).
We used the Fleiss' kappa [34] as the statistical measure for inter-rater agreement between more than two raters. This measure is used when data consists of categorical data. The inter-rater agreement for the experiment was 0.123 which is a slight agreement. However, if we use the higher categories such as smile, neutral, etc. agreement equals 0.26 which is fair agreement.

3.3.3 Discussion and limitations

Nearly all of the emojis were used by annotators, this means that each of the emojis was perceived in at least one clip by one annotator. The results showed that (🙂) and (😐) were the first and second most common emojis this shows that in most of the dataset's videos, the participants interacting with the robot were slightly smiling or were neutral.

It is possible that participants matched literally what they saw with the emojis, meaning that they matched the exact facial configuration to emoji showing that, however, we found that participants sometimes chose emojis where the configuration did not appear in videos e.g. (no participants placed their hands under their head for thinking: 🤔) to annotate videos. It remains an open challenge to know exactly to what extent annotators selected emojis based on the physical attributes versus their meanings.

This inter-rater agreement number was slightly lower than the value we expected. There are several possible explanations for low level of inter-rater agreement. Our results demonstrated that if we use the higher level categories specified in Figure 5.1 such as smile, neutral, worried, etc., and combine corresponding emojis into same class, the agreement changes to 0.26 which is fair agreement. This suggests that the lower agreement value in the main

Figure 3.7: Some frames include more than one participant

experiment was due to a higher number of categories. Another reasonable explanation for this may be the semi-automatic segmentation of videos. Although our goal was to include only one emotional expression per clip, we can not assure that. Future work can use audio features to segment the videos as well. Another cause for this is having multiple people in one frame. Although participants were instructed to participate alone in the interaction with the Pepper robot [9], the dataset contains several interactions that include more than one participant. It is very likely that annotators may have erroneously identified social signals for different people in one frame. Inter-rater agreement increased from 0.123 to 0.134 when we removed the videos containing more than one participant in the frame.

Another explanation for this low agreement can be emotion residue remaining from prior expressions in the face. Albohn et al. suggests that "faces explicitly intended to show no expressive cues nonetheless still convey the emotional tone of a prior expression" [3]. This may have occurred because we segmented the videos into shorter clips and the clips are only small part of the full video of primary interaction. Also, emoji perception confusion may have happened. As we had several participants, each participant may have interpret the emojis in a different way. Especially for the emojis that are used less in the wild. Clearly, there may be other possible explanations such as limitations in emoji designs, as some of the participants wrote comments such as "emoji 14 (🤔) but with a slight smile" and "some emoji in between emoji 19(😏) and emoji 14 (🤔)"

We aware of several limitations in this work. The first is that the people interacting with the robot in the UE-HRI dataset were primarily students located in France and speaking in French. Another limitation is the annotators were from Canada which is a different but related culture to French. An additional limitation that should be discussed is emotional expressions are multi-modal (verbal, body posture, etc.), however we only incorporated the face in this experiment. Also given that we only focused on facial and limited body gestures, there is a possibility that dissimilar evaluations would have arisen if the focus had been on audio content or other modalities. Another possible limitation is the influence of mood of each individual at the time of filling out the survey. We did not ask our annotators their mood they had when filling out the survey. Additional limitation is perceiving emotions can be related to individual differences as well.

Future work may take into account culture and language when considering expressions and annotation, and applying our method to other available in-the-wild datasets. The second is the matching of textual descriptions of emotions to the emojis, which we will discuss in Chapter 4.

Chapter 4

Social signals to emotions

In the previous chapter, we collected a set of emojis as social signals proxies in the wild in HRI. In this chapter, we aim to find the meaning behind these social signals. To the best of our knowledge previous studies investigated perception of emojis used within text or did not include all the emojis that we used [61]. We present a novel approach to analyze these social signals. The aim of this experiment is to study how each of the annotators would interpret each emoji and also find appropriate textual labels for social signal categories.

4.1 Experiment 2: Emoji meanings

4.1.1 Emotion word selection

The first step in interpreting the emojis, is to provide a vocabulary to the users to select. For emotion words we used the Junto emotion wheel [1] [16]. It is a "large wheel shape, with three inner circles as shown in Figure 4.1 . Broader emotional groups are in the center, and emotions become more specific as they move toward the edge". [2] We used only two inner circles and also we added neutral to the centre of the wheel. In total there are 7 words on the inner circle and 34 words on the outer circle. Junto emotion wheel is used in the wild for counselling and group sessions. Furthermore, it includes Love as a primary category which sub-categorizes the positive emotions as well as negative emotions. For mental states we used Baron-Cohen et al. emotion taxonomy which encompasses a wide range of emotional and cognitive thought states. There are 412 mental state concepts included in this taxonomy, each of which is assigned to one (and only one) of 24 mental state classes as shown in Table 4.1. The 24 classes were chosen to preserve the semantic distinction of the emotion categories within each class [33] [39].

We selected five classes out of the 24 that are particularly relevant to human-robot interaction context: kind, disbelieving, interested, thinking, unsure to add to the wheel.

[1] www.thejuntoinstitute.com/emotion-wheels/

[2] https://greatist.com/grow/wheel-of-emotions

Mental States	
Surprised	Touched
Wanting	Liked
Hurt	Unfriendly
Disgusted	Unsure
Kind	Afraid
Happy	Excited
Sad	Angry
Disbelieving	Fond
Sneaky	Bothered
Romantic	Thinking
Interested	Bored
Sorry	Sure

Table 4.1: The 24 groups of emotion and cognitive labels

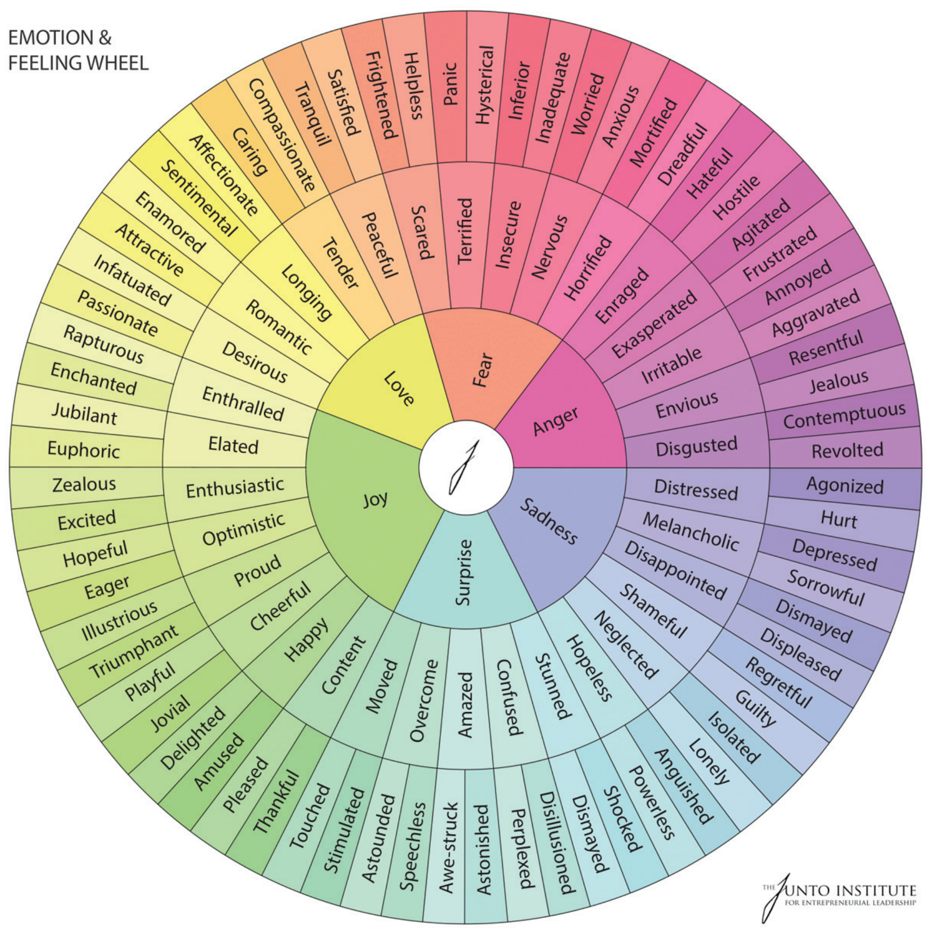


Figure 4.1: Junto emotion wheel

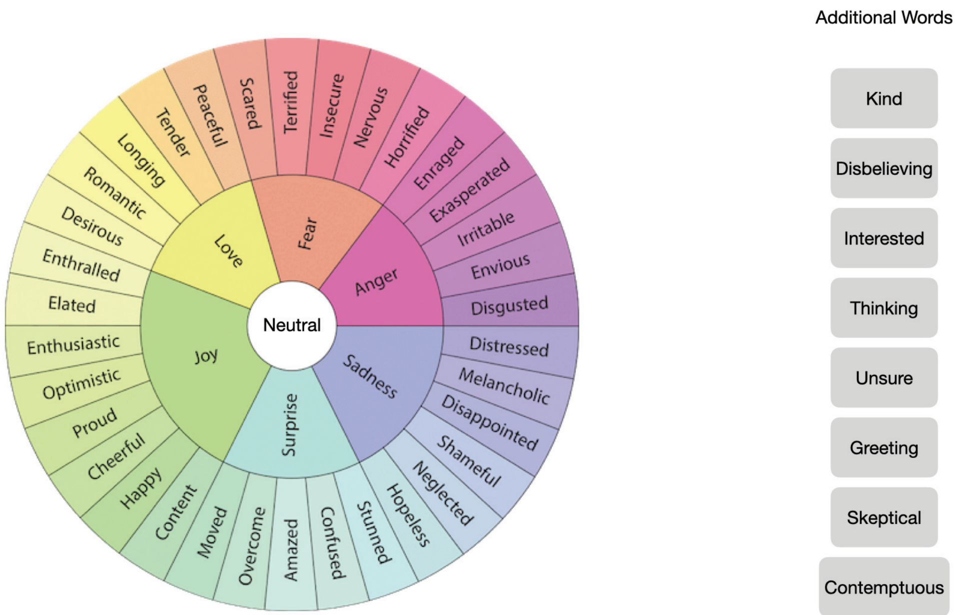


Figure 4.2: Emotion wheel and additional words shown to the annotators

We then added greeting, skeptical and contemptuous as they were related to human-robot interaction in the wild.

For emojis, we used the exact emoji set from experiment 1 as shown in Figure 3.4.

4.1.2 Method

A total of 28 people were recruited for this study from Amazon Mechanical Turk. All of the participants were aged 19 to 40 in order to include digital natives generation. To increase the likelihood of high quality annotations, we set our requirements to a minimum HIT approval rate of $\geq 95\%$ (percentage of completed work that has been approved by other requester's) and having approved HITs ≥ 5000. We required all of our annotators to be located in Canada for ethical board approval. The assignment duration for each task was set to 20 minutes based on the estimates of the SurveyMonkey website and the survey was completed by 28 unique annotators.

In the first step, we posted a HIT(Human Intelligence Task) containing 29 emojis. Annotators first provided informed consent about study details and procedure. Then they viewed each of these emojis along with the Junto emotion wheel [3] of emotions and the additional words shown in Figure 4.2. They were asked to choose one word that best matches with the provided emoji. They could also write any other words that were not included in the end of

[3] www.thejuntoinstitute.com/emotion-wheels/

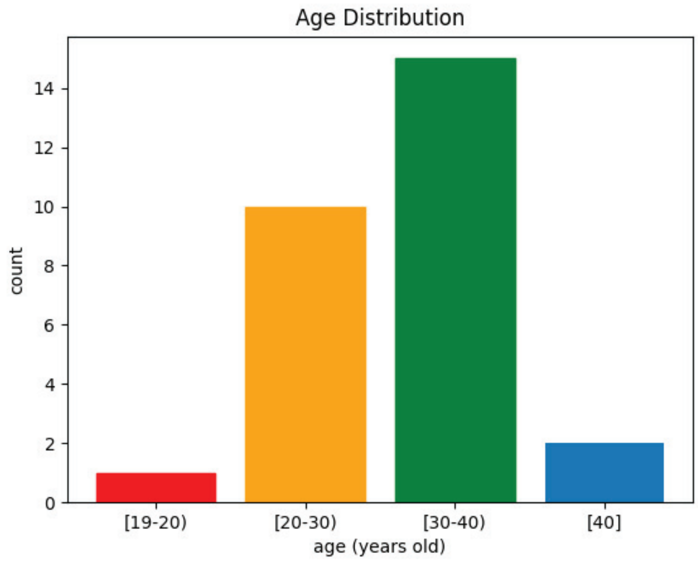


Figure 4.3: Age distribution of participants in the experiment

the survey. Each annotator was paid per HIT such that their per hour rate was equivalent of the minimum wage in their country.

4.1.3 Results

Of the study population, all of the participants completed the survey. However, since the age range was between 19 to 40 years old, we rejected two participants with ages > 40. So the total number of participants was 28. Over half (60%) of participants were aged between 30 to 40. Almost 36% of participants were aged between 20 to 30 and the rest 10% were aged 40 or < 20 years old. We visualized the age distribution of participants in Figure 4.3.

Familiarity

When the participants were asked about their familiarity with the emojis, half of the participants responded 'Very familiar'. Approximately 30% answered 'Extremely familiar' and almost 20% replied 'Somewhat familiar'. Very few participants(3.57%) reported 'Not so familiar' and no one reported 'Not at all familiar'. We visualized emoji familiarity among raters in Figure 4.4.

4.1.4 Emoji label analysis

After initial analysis we decided to categorize emojis and their labels into 3 categories: Unimodal, bimodal and multimodal. We used the inner wheel from Junto's emotion wheel in order to decide which category each emoji falls into. Summary of this analysis is shown in Table 4.2 and Table 4.3.

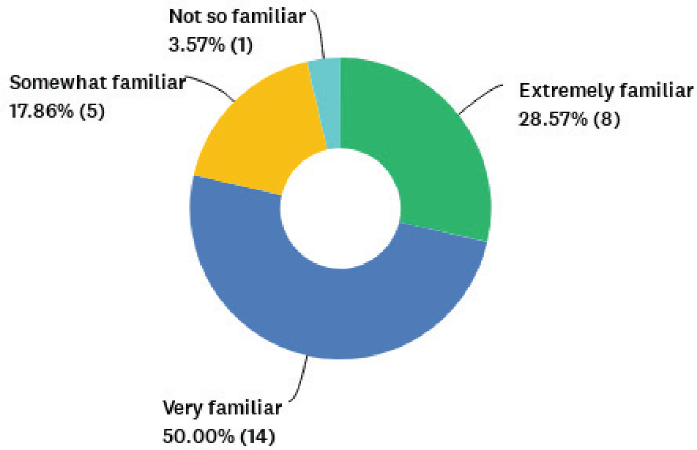


Figure 4.4: Emoji familiarity of participants in the experiment

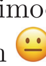
Emojis with unimodal label distributions

We defined the emoji having a unimodal label distribution if more than half of the raters (>50%) rated it as one specific category from the inner-wheel (Love, Joy, Surprise, Sadness, Neutral, Anger). As a result 17 of 29 emojis were rated as unimodal emojis. One example of this kind is . The majority of voters (88.89%) rated it as 'Joy' as shown in Figure 4.5. There were 17 emojis with unimodal label distribution in total as shown in Table 4.2.

Emojis with bimodal label distributions

We defined the emoji having a bimodal label distribution if from the inner-wheel (Love, Joy, Surprise, Sadness, Neutral, Anger) two categories had between 33% and 50% ratings. For instance, in  emoji, 'Anger' (33.33%) and 'Additional Words' (40.74%) (Skeptical and Disbelieving) labels were chosen by respondents as shown in Figure 4.6. There were 7 emojis with bimodal label distribution in total as shown in Table 4.3.

Emojis with multimodal label distributions

We defined the rest of emojis not belonging to unimodal or bimodal emojis as multimodal. These emojis encompass 3 or more categories. For instance, in  emoji, Neutral (17.86%), Sadness (28.57%), Surprise (10.71%), Anger (10.71%) and Additional Words (32.14%) were selected as shown in Figure 4.7. There were 5 emojis with multimodal label distribution in total as shown in Table 4.3.

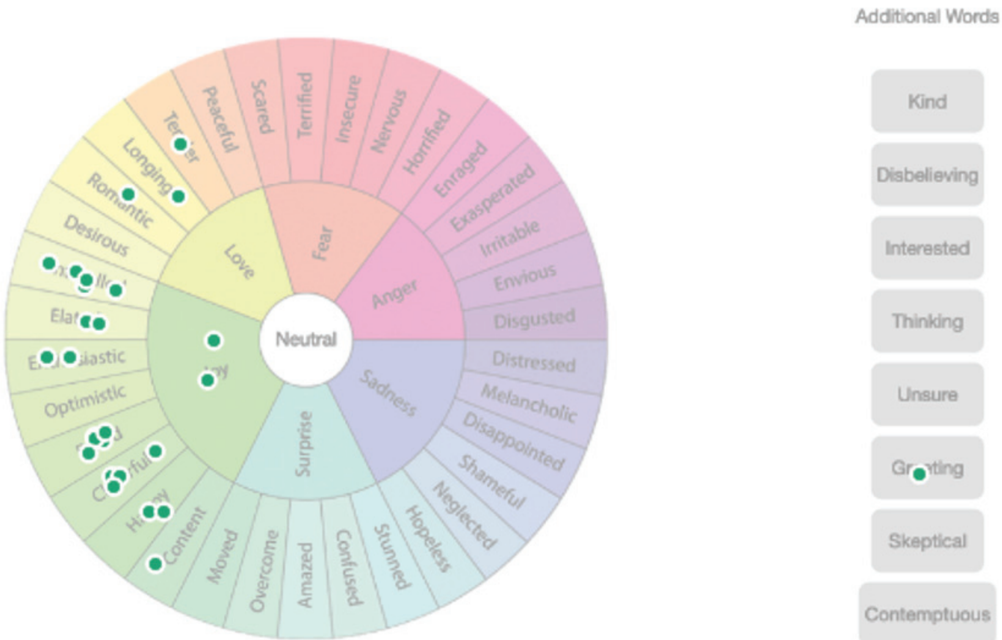


Figure 4.5: Unimodal distribution for (88.89% rated as Joy)

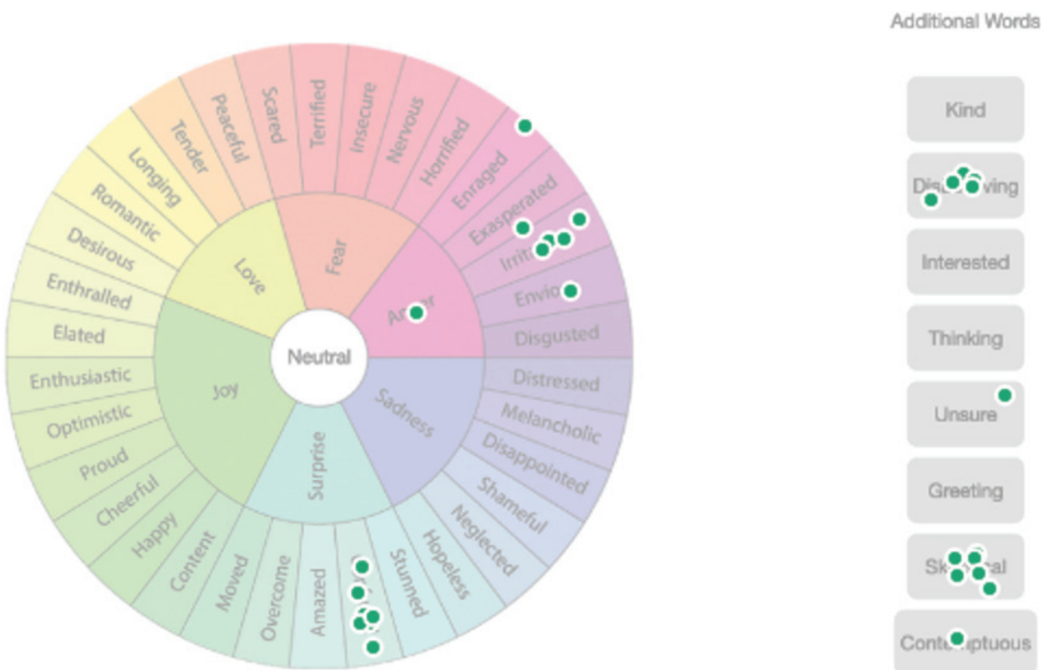


Figure 4.6: Bimodal distribution for (33.33% rated as 'Anger' and 40.74% rated as 'Additional Words')

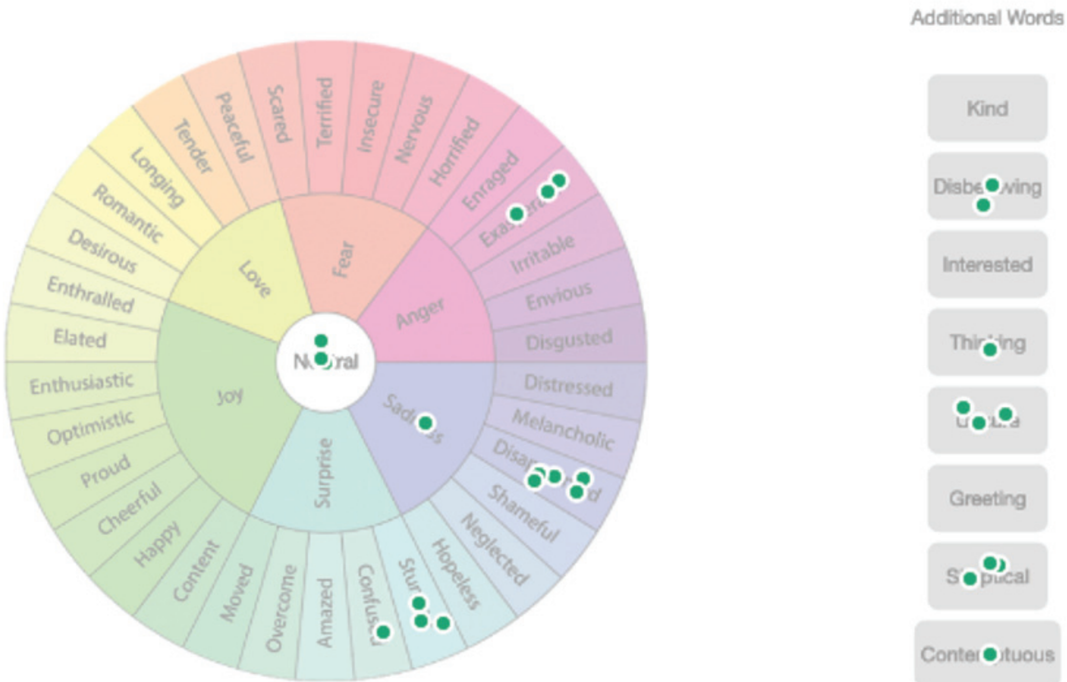


Figure 4.7: Multimodal distribution for (17.86% rated as Neutral, 28.57% rated as Sadness, 10.71% rated as Surprise, 10.71% rated as Anger and 32.14% rated as Additional Words)

Category	Emoji	High-Level Labels	Low-level Labels
Unimodal	🙂	Joy	Peaceful, Content, Happy, Cheerful, Kind, Greeting, Neutral
	🙃	Joy	Disappointed, Confused, Amazed, Overcome, Content, Happy, Cheerful, Enthusiastic, Elated, Disbelieving, Skeptical
	😊	Joy	Romantic, Longing, Tender, Content, Happy, Cheerful, Proud, Enthusiastic, Elated, Enthralled
	😃	Joy	Amazed, Content, Happy, Cheerful, Optimistic, Enthusiastic, Elated, Enthralled
	😀	Joy	Stunned, Content, Happy, Cheerful, Proud, Optimistic, Enthusiastic, Elated
	😂	Joy	Stunned, Overcome, Content, Happy, Cheerful, Enthusiastic, Elated, Enthralled
	🤣	Joy	Stunned, Amazed, Content, Happy, Cheerful, Enthusiastic, Elated, Enthralled, Unsure
	😆	Joy	Confused, Amazed, Overcome, Content, Happy, Cheerful, Optimistic, Enthusiastic, Elated, Enthralled
	😁	Joy	Content, Happy, Cheerful, Proud, Optimistic, Enthusiastic, Elated

Category	Emoji	High-Level Labels	Low-level Labels
Unimodal	😄	Joy	Overcome, Content, Happy, Cheerful, Proud, Optimistic, Enthusiastic
	😶	Surprise	Insecure, Melancholic, Disappointed, Neglected, Hopeless, Stunned, Confused, Amazed, Disbelieving, Thinking, Unsure, Neutral
	😕	Sadness	Distressed, Melancholic, Disappointed, Shameful, Hopeless, Confused, Unsure, Skeptical
	🤔	Additional Words	Stunned, Confused, Disbelieving, Thinking, Skeptical
	😮	Surprise	Scared, Exasperated, Stunned, Amazed, Overcome, Disbelieving
	😗	Love	Desirous, Romantic, Longing, Tender, Content, Happy, Cheerful, Enthusiastic, Elated
	👋	Additional Words	Longing, Peaceful, Irritable, Content, Cheerful, Unsure, Greeting, Neutral
	😉	Love	Desirous, Romantic, Tender, Content, Happy, Cheerful, Proud, Optimistic, Enthusiastic, Kind, Interested, Greeting

Table 4.2: Summary of Experiment 2: Unimodal Emojis

Category	Emoji	High-Level Labels	Low-level Labels
Bimodal	☺️	Love, Joy	Romantic, Longing, Tender, Nervous, Overcome, Moved, Content, Cheerful, Proud, Enthusiastic
	😏	Additional Words, Love	Desirous, Romantic, Longing, Insecure, Envious, Disgusted, Shameful, Overcome, Optimistic, Enthralled, Interested, Skeptical, Contemptuous
	🤨	Additional Words, Anger	Enraged, Exasperated, Irritable, Envious, Confused, Disbelieving, Skeptical
	😒	Anger, Sadness	Exasperated, Irritable, Disgusted, Distressed, Melancholic, Shameful, Neglected, Confused, Moved, Disbelieving, Skeptical, Contemptuous
	😑	Anger, Sadness	Peaceful, Horrified, Exasperated, Irritable, Disgusted, Distressed, Melancholic, Disappointed, Shameful, Neglected, Hopeless, Moved, Disbelieving, Neutral
	😳	Fear, Surprise	Terrified, Insecure, Nervous, Horrified, Shameful, Stunned, Confused, Amazed, Disbelieving
	😦	Fear, Sadness	Scared, Terrified, Insecure, Horrified, Disgusted, Distressed Disappointed, Shameful, Hopeless, Stunned, Overcome, Unsure

Category	Emoji	High-Level Labels	Low-level Labels
Multimodal	😅	Joy, Fear, Surprise, Love	Desirous, Insecure, Nervous, Hopeless, Stunned, Confused, Overcome, Moved, Happy, Cheerful, Proud, Optimistic, Enthusiastic, Elated
	😐	Anger, Sadness, Surprise, Additional Words, Neutral	Exasperated, Disappointed, Shameful, Stunned, Confused, Disbelieving, Thinking, Unsure, Skeptical, Contemptuous
	🙄	Anger, Fear, Sadness, Surprise, Additional Words	Scared, Nervous, Exasperated, Irritable, Disgusted, Disappointed, Shameful, Hopeless, Stunned, Confused, Overcome, Disbelieving, Skeptical, Contemptuous
	👐	Love, Fear, Anger, Joy, Additional Words, Neutral	Longing, Peaceful, Nervous, Exasperated, Cheerful, Optimistic, Enthusiastic, Elated, Enthralled, Kind, Unsure, Greeting, Neutral
	👀	Love, Fear, Anger, Surprise, Additional Words	Desirous, Scared, Terrified, Nervous, Envious, Stunned, Confused, Amazed, Interested, Unsure, Skeptical

Table 4.3: Summary of Experiment 2: Bimodal and Multimodal Emojis

4.1.5 Discussion and limitations

As the results suggest, 17 out of 29 emojis were rated as prototypical emotional expressions since the majority of raters chose one category for them. Most of these emojis (10) were rated as Joy. This suggests that the annotators had higher agreement on smile/joy emojis.

Another interesting finding was 2 prominent categories for 7 emojis. This suggests that in these cases the participants were unsure between two categories. For example, 😏 emoji had both a positive and negative interpretation such as Contemptuous and Desirous. This points towards the idea that with only facial expression, an underlying emotional category is not clear; using multiple modalities, verbal communication components or context can help towards choosing one of the two categories. It is plausible that a number of limitations might influenced the results obtained. To begin with, we did not include context of HRI when asking annotators to fill out the form. Also, we only asked annotators from Mechanical Turk in Canada. It is possible that participants were more familiar with some emojis more than other emojis; this could have also influenced the results obtained. Another limitation is the cultural difference. Gao et al. suggests that there are cultural disparities in the use of mouth vs eye clues to identify emotions and that these differences extend to paralinguistic cues like emojis [37]. Another probable limitation is the choice of better additional words. Additional words were chosen by discussion until agreement between researchers. However, there might be additional words suitable for this experiment.

Chapter 5

Social Signal Taxonomy

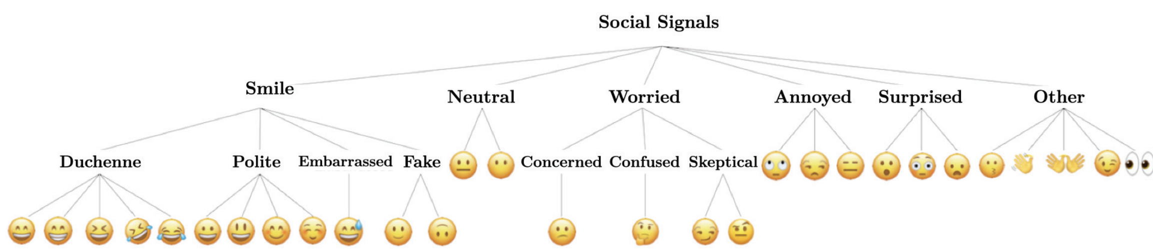


Figure 5.1: Initial taxonomy of social signals in the UE-HRI dataset

5.1 Initial taxonomy from experiment 1

We devised a hierarchical taxonomy from the initial results obtained, as shown in Figure 5.1. Our emoji annotation results illustrated that expressions in HRI are not just emotional, therefore, our taxonomy comprises social signals in a broader sense. Researchers discussed until agreement on an initial categorization. As detailed in Figure 5.1, in the first step, we categorized social signals in the dataset into 6 categories: Smile, Neutral, Worried, Surprised and Other. Inspiration was taken from the taxonomy of the Unicode Emoji dataset. [1] Once the initial categorization had been done, we then sub-categorized social signals into fine-grained sub-categories. From Figure 5.1 we can see that sub-categories include Duchenne or genuine smile, Polite, Fake and Embarrassed smiles and Confused, Concerned and Skeptical emotional expressions. The Other category includes the social signals that have no clear emotional meaning.

[1] https://unicode.org/emoji/charts/full-emoji-list.html

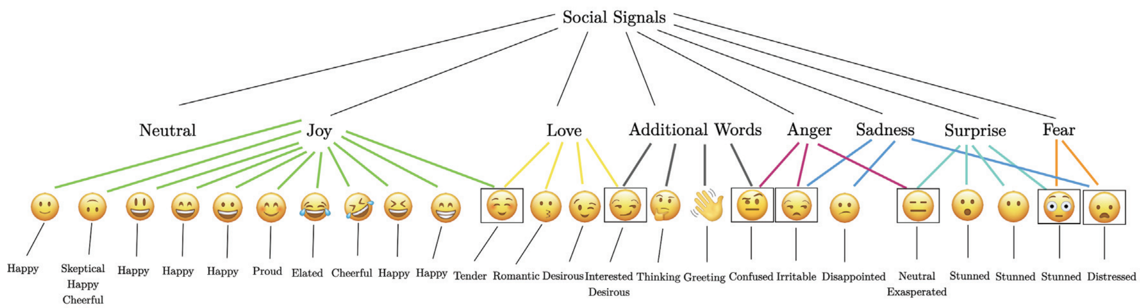


Figure 5.2: Improved taxonomy of social signals (square shows the bimodal emojis)

5.2 Improved taxonomy from experiment 2

We further improved the taxonomy from experiment 1 by incorporating the emotion labels for each emoji from experiment 2. In this step, instead of researchers choosing the categories for each emoji, we used the emotion labels for unimodal and bimodal emojis as shown in Figure 5.2. Bimodal emojis are specified with square. We excluded the multimodal emojis in this taxonomy as they would fall under 3 or more categories. As illustrated in Figure 5.2, 7 emojis fall into two categories.

5.3 Alternative representation

As stated in previous section, we only included the unimodal and bimodal emojis in the refined taxonomy. The refined taxonomy included the high-level categories such as neutral, love, joy, etc. However, our main aim was to capture the diverse ways that humans perceive these social signals. We decided that the best way of showing our result was to use a circular tree as shown in Figure 5.3. We used the circular form for facilitating the visualization of categories and sub-categories. We used the original categories from Junto's emotion wheel for the higher level categories and we included the subcategories that included more than one rating. The number of ratings is showed as a color gradient as well as on each branch. This kind of visualization also helps to illustrate the unimodal, bimodal, multimodal emojis. As shown in Figure 5.3, this emojis falls under the unimodal category. This visualization helps to preserve the rich vocabulary of emotion words that participants used for describing their perception of each emoji. By using this visualization we can comprehend the different meanings humans perceive from each emoji. Please refer to Appendix for circular tree distribution plots of all emojis.

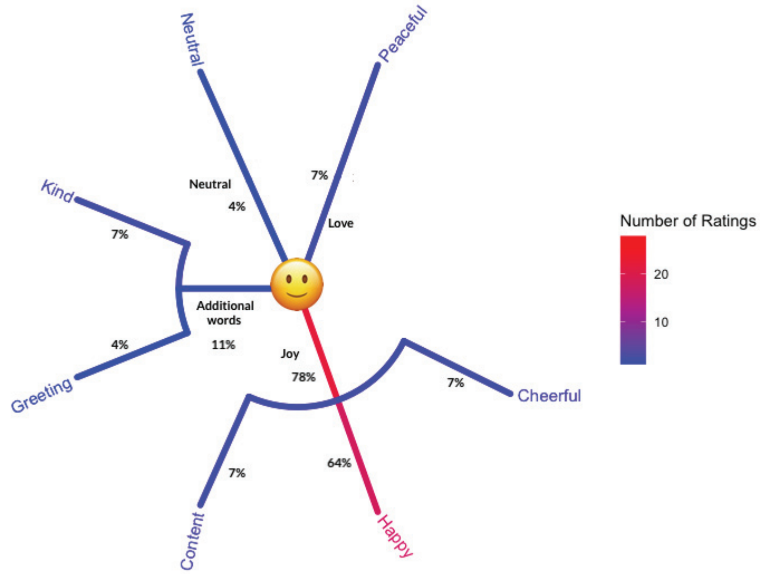


Figure 5.3: Unimodal distribution for 🙂

Chapter 6

Conclusion and Future Work

In this thesis we aimed to discover the social signals occurring in the wild in HRI and investigate the meaning behind them. Our study provides the framework for a novel way to analyze and categorize social signals in the wild in the HRI. We conducted two different experiments supporting our study. We used a data-driven approach of starting from expressions seen in the wild, converting them to low-level social signal representation and then finding the emotion words associated with them. We used a data-driven method to study social signals. We collected a rich collection of social signals in the wild by segmenting and labeling videos from a HRI dataset collected in the wild. We used emojis as social signal representations. We devised a hierarchical initial taxonomy of social signals and emotional expressions in real-world human-robot interactions. We showed that there are additional emotion categories present in real-world social settings other than the existing prototypical categories. We also found that skeptical expressions were relatively common in HRI in real world settings. We studied how humans perceive each emoji. We used a comprehensive emotion wheel to find associated affective states with each emoji. Finally, we proposed an alternative model for social signals analysis in the wild.

In chapter 2, we used a data-driven method to devise a hierarchical taxonomy of social signals in the wild. We segmented the videos from a HRI dataset in the wild. We then recruit annotators to label these videos with emojis as social signal proxies. The hierarchical taxonomy involved two levels of hierarchies. The higher level includes six coarse-grained categories. The lower level includes fine-grained sub-categories. We also found a relatively high number of skeptical expressions.

In chapter 3, we refined the initial taxonomy based on the new results obtained. We studied how humans perceive each emoji. Our results suggest that there is a many to many relation between emotion words and emojis meaning each emoji was correlated with several emotion words and each emotion word was used for several emojis. We categorized emojis into Unimodal, Bimodal and Multimodal categories based on the annotations.

6.1 Limitations

The present study has only investigated a social interaction between a human and a robot placed in a social space in France. Furthermore, the majority of people interacting with the robot were students speaking in French. Another limitation is that the full scene was not shown in the videos. As a result, annotators could only see the face of person interacting with the robot and not robot surroundings which could affect the emotional expressions produced. We randomly selected up to 3 clips from each of the pre-selected videos in order to cover a broad selection of interactions. However, given the sample size caution must be taken when analyzing the results of experiment 1. Finally, given low inter-rater agreement in Experiment 1, we believe the results can be taken as a first step, and future work to discover social signals in HRI should involve more raters, similar to Experiment 2. This would allow the categorization of expressions into emojis as unimodal, bimodal or multimodal. A multilabel annotation scheme could also be considered.

In experiment 2, we used the Junto Emotion Wheel for the sake of comprehensiveness. However, there could exist several more emotion words. Furthermore, influence of culture is also another limitation to be considered. As different cultures include emotion words specific to that culture. Another limitation to be considered is since the unicode emoji set is expanding each year, the emoji set is thus still incomplete. Overall, the extra step of converting videos to emojis, then emojis to text labels may add more potential for noise. However, we still believe that emojis were a good choice of representation compared to the alternatives of text labels and action units. On one hand, compared to text labels, emojis are a curated set of communicative social signals and used in everyday communication, and this step was helpful in finding which labels would be relevant for annotation. On the other hand, compared to raw action units, emojis were created to have communicative meaning and experts are not required for coding. Our work clearly had some limitations. Despite this we believe our work could be a starting point for social signal and emotion label analysis.

6.2 Future work

Further experimental investigations are needed to study varied contexts. One example of these contexts can be group vs. solo scenarios. More broadly, we suggest further research should be done on cultural influences and language. Our results should be validated by another HRI dataset in the wild as well. Further experimental tests are needed to establish whether the relation from videos to emojis and from emojis to emotion words are transitive. Future studies should also use newer set of unicode emojis.

6.3 Application and implications

Our approach has three main implications which we will discuss below.

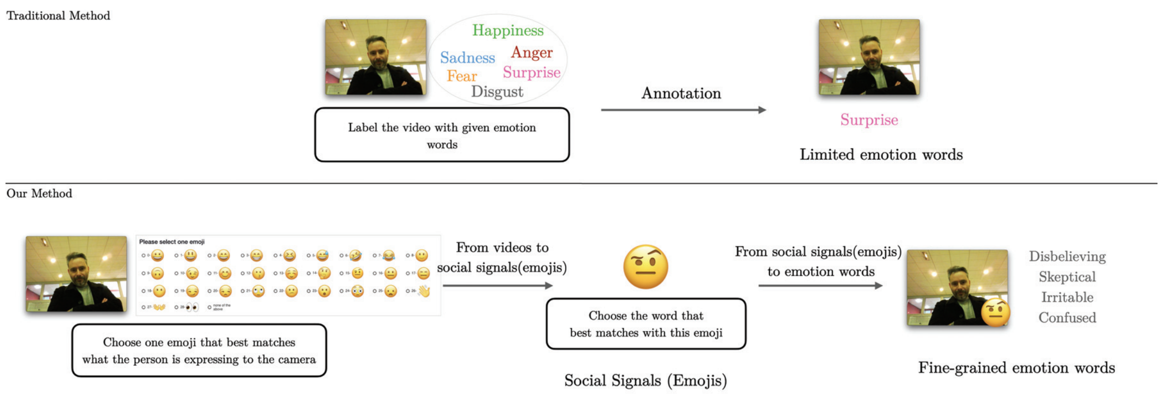


Figure 6.1: Traditional approach and our method

6.3.1 Dataset creation based on social signals

We hope that our research will be beneficial for creating datasets based on social signals rather than emotion words. For example, a confused face in the dataset can be labeled as 😕. Our research suggests that other researchers can use emojis as social signal representations for their future work. "Several linguists, psychologists, and anthropologists have demonstrated that emotion talk, concepts, and scripts may differ across cultures and that, as a result, some emotion words may have no translation equivalents" [26] However emojis are the same independently of where one lives [7]. As a result, we suggest that researchers label the collected datasets with emojis. Furthermore, many machine learning problems require large amount of labeled data, by using emojis as labels datasets can be used cross-lingual. As a result, this study can support machine learning problems such as creating classifiers and emotion recognition by using emojis as low-dimensional social signal representation for problems.

6.3.2 Narrow-down social signal analysis problem

The present findings can improve human robot interaction by developing new modules in robot architecture for analysis of social signals. As we mentioned in chapter 3, we defined several emojis as bimodal emojis. This could eventually help to narrow-down the scope of social signal analysis. For instance, if the identified social signal maps to ☺️ emoji, we know the decision space is between Love and Joy categories. This can be a useful aid to reduce the decision making space.

6.3.3 Proposed data-driven framework for social signal analysis in HRI

Top-down vs. bottom-up approach

The traditional approach to study videos that contain social signals is to presume a fixed set of emotion categories and ask annotators to select videos that fall into each category. Several facial emotion recognition datasets use this traditional approach. In this top-down approach annotators are asked to annotate videos with a set of pre-configured categories such as six basic emotions (anger, disgust, fear, happiness, sadness and surprise) [55] or the basic 6 combined with neutral and some complex emotion states [63] .

As shown in Figure 6.1 instead of starting from emotion words such as happiness, sadness, etc. and then finding the relative video for these emotions, we started from videos, used emojis as social signal representations and then asked participants to label the emojis in order to get appropriate label for each emoji. This data-driven bottom-up approach helps us to capture and categorize rich collection of social signals humans perceive. This bottom-up approach consists of two main stages which we will discuss below.

Stage 1: Videos to emojis

In order to investigate social signals in the wild, we used emojis as a proxy of emotional expressions and social signals since emojis are a visual simplified form of affective communication [67]. People frequently rely on nonverbal behaviours to properly express emotion in face-to-face conversations. Non-verbal expression is not as easily transmitted in text-based communication. As a result, individuals utilise emojis as a quick and simple way to express nonverbal behaviors through text.

The reason for this stage is to have a low dimensional embedding of social signals in the videos in emoji space since emojis encode the social signal in the videos and they are low-dimensional. This can also make machine learning models more efficient and easier to work with. One example of real world usage of this kind of embedding is the application that takes a stream of video or photos from a person as input and generates an emoticon based on the image face. The technology recognizes the facial expression of the person who is sending the message. The device generates a message with the appropriate emoticon once that facial expression is detected [17][38][2].

Stage 2: Emojis to emotion words

Emojis are non-verbal cues with rich emotional meanings [6]. Jaeger et al. found that most emoji can express one or more emotions [44]. In order to study the corresponding emotion labels for each emoji, we setup an experiment that investigated which emoji relates to which emotion labels. This step could also help us in creating a taxonomy with appropriate labels for emojis.

Although our research may have possible limitations, we believe that this framework can be used in the future for HRI research using in-the-wild data. Taken together, we hope that researchers use this 2-stage, bottom-up approach to discover the social signals occurring in the wild and the meaning behind them.

Bibliography

[1] Mohsin Manshad Abbasi and AP Beltiukov. Summarizing emotions from text using plutchik's wheel of emotions. In *Scientific Conference on Information Technologies for Intelligent Decision Making Support*, 2019.

[2] Z. Al-Halah, A. Aitken, W. Shi, and J. Caballero. Smile, be happy :) emoji embedding for visual sentiment analysis. In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4491–4500, 2019.

[3] Daniel N Albohn and Reginald B Adams Jr. Emotion residue in neutral faces: implications for impression formation. *Social Psychological and Personality Science*, 12(4):479–486, 2021.

[4] Subashini Annamalai and Sobihatun Nur Abdul Salam. Undergraduates' interpretation on whatsapp smiley emoji. *Jurnal Komunikasi: Malaysian Journal of Communication*, 33(4), 2017.

[5] Lesley Axelrod and Kate S Hone. Affectemes and allaffects: a novel approach to coding user emotional expression during interactive experiences. *Behaviour & Information Technology*, 25(2):159–173, 2006.

[6] Qiyu Bai, Qi Dan, Zhe Mu, and Maokun Yang. A systematic review of emoji: Current research and future perspectives. *Frontiers in psychology*, 10:2221, 2019.

[7] Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. How cosmopolitan are emojis? exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 531–535, 2016.

[8] 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.

[9] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. Ue-hri: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the ACM international conference on multimodal interaction*, pages 464–472, 2017.

[10] Atef Ben-Youssef, Giovanna Varni, Slim Essid, and Chloé Clavel. On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks. *International Journal of Social Robotics*, 11(5):815–828, 2019.

[11] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*, 2017.

[12] Anne C Bloem, Emilia Barakova, Inge M Hootsmans, Lena M Opheij, Romain HA Toebosch, Matthias Kerzel, and Pablo Barros. Improving emotional expression recognition of robots using regions of interest from human data. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 142–144, 2020.

[13] Dan Bohus and Eric Horvitz. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*, pages 2–9, 2014.

[14] Anna Bortolan. A phenomenological discussion of antonio damasio's theory of emotions. *Phenomenology and mind*, (1):142–150, 2011.

[15] O. Celiktutan, E. Skordos, and H. Gunes. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 10(4):484–497, 2019.

[16] Raman Chadha. The junto emotion wheel: Why and how we use it. *The Junto Institute*, 2020.

[17] Ajay B Chembula. Generating emoticons based on an image of a face, February 21 2017. US Patent 9,576,175.

[18] Luefeng Chen, Mengtian Zhou, Wanjuan Su, Min Wu, Jinhua She, and Kaoru Hirota. Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. *Information Sciences*, 428:49–61, 2018.

[19] Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. Emoji-powered representation learning for cross-lingual sentiment classification. In *The World Wide Web Conference*, pages 251–262, 2019.

[20] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[21] Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. In *Proceedings of the National Academy of Sciences*, volume 114, pages E7900–E7909, 2017.

[22] Adam K Coyne, Andrew Murtagh, and Conor McGinn. Using the geneva emotion wheel to measure perceived affect in human-robot interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 491–498, 2020.

[23] Antonio Damasio. Fundamental feelings. *Nature*, 413(6858):781–781, 2001.

[24] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[25] Jia Deng, Gaoyang Pang, Zhiyu Zhang, Zhibo Pang, Huayong Yang, and Geng Yang. cgan based facial expression recognition for human-robot interaction. *IEEE Access*, 7:9848–9859, 2019.

[26] Jean-Marc Dewaele and Aneta Pavlenko. Emotion vocabulary in interlanguage. *Language learning*, 52(2):263–322, 2002.

[27] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. In *Proceedings of the National Academy of Sciences*, volume 111, pages E1454–E1462, 2014.

[28] Paul Ekman. Facial action coding system. 1977.

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

[30] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[31] Paul Ekman. Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3449–3451, 2009.

[32] Rana El Kaliouby, May Amr Fouad, Abdelrahman N Mahmoud, Seyedmohammad Mavadati, and Daniel Mcduff. Emoji manipulation using machine learning, 2020.

[33] Rana El Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005.

[34] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[35] Mary Ellen Foster, Bart Craenen, Amol Deshmukh, Oliver Lemon, Emanuele Bastianelli, Christian Dondrup, Ioannis Papaioannou, Andrea Vanzo, Jean-Marc Odobez, Olivier Canévet, et al. Mummer: Socially intelligent human-robot interaction in public spaces. *CoRR*, abs/1909.06749, 2019.

[36] NH Frijda and L Berkowitz. Advances in experimental and social psychology, 1969.

[37] Boting Gao and Doug P VanderLaan. Cultural influences on perceptions of emotions depicted in emojis. *Cyberpsychology, Behavior, and Social Networking*, 23(8):567–570, 2020.

[38] Amanda N Gesselman, Vivian P Ta, and Justin R Garcia. Worth a thousand interpersonal words: Emoji as affective signals for relationship-oriented digital communication. *PloS one*, 14(8):e0221297, 2019.

[39] Ofer Golan, Simon Baron-Cohen, and Jacqueline Hill. The cambridge mindreading (cam) face-voice battery: Testing complex emotion recognition in adults with and without asperger syndrome. *Journal of autism and developmental disorders*, 36(2):169–183, 2006.

[40] Simeng Gu, Fushun Wang, Nitesh P Patel, James A Bourgeois, and Jason H Huang. A model for basic emotions using observations of behavior in drosophila. *Frontiers in psychology*, 10:781, 2019.

[41] Gaël Guibon, Magalie Ochs, and Patrice Bellot. From emojis to sentiment analysis. In *WACAI 2016*, 2016.

[42] Ursula Hess and Pascal Thibault. Darwin and emotion expression. *American Psychologist*, 64(2):120, 2009.

[43] Rachael E Jack, Carlos Crivelli, and Thalia Wheatley. Data-driven methods to diversify knowledge of human psychology. *Trends in cognitive sciences*, 22(1):1–5, 2018.

[44] Sara R Jaeger, Leticia Vidal, Karrie Kam, and Gastón Ares. Can emoji be used as a direct method to measure emotional associations to food names? preliminary investigations with consumers in usa and china. *Food Quality and Preference*, 56:38–48, 2017.

[45] Dinesh Babu Jayagopi, Samira Sheiki, David Klotz, Johannes Wienke, Jean-Marc Odobez, Sebastien Wrede, Vasil Khalidov, Laurent Nyugen, Britta Wrede, and Daniel Gatica-Perez. The vernissage corpus: A conversational human-robot-interaction dataset. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 149–150, 2013.

[46] Malte Jung and Pamela Hinds. Robots in the wild: A time for more robust theories of human-robot interaction, 2018.

[47] Melissa G Keith, Louis Tay, and Peter D Harms. Systems perspective of amazon mechanical turk for organizational research: Review and recommendations. *Frontiers in Psychology*, 8:1359, 2017.

[48] Eva G Krumhuber, Arvid Kappas, and Antony SR Manstead. Effects of dynamic aspects of facial expressions: A review. *Emotion Review*, 5(1):41–46, 2013.

[49] Brent Lance and Stacy Marsella. Glances, glares, and glowering: how should a virtual human express emotion through gaze? *Autonomous Agents and Multi-Agent Systems*, 20(1):50–69, 2010.

[50] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scassellati. Comparing models of disengagement in individual and group interactions. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 99–105, 2015.

[51] Séverin Lemaignan, Charlotte ER Edmunds, Emmanuel Senft, and Tony Belpaeme. The pinsoro dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PloS one*, 13(10):e0205999, 2018.

[52] Tzuu-Hseng S Li, Ping-Huan Kuo, Ting-Nan Tsai, and Po-Chien Luan. Cnn and lstm based facial expression analysis model for a humanoid robot. *IEEE Access*, 7:93998–94011, 2019.

[53] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, and J. Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica*, 4(4):668–676, 2017.

[54] Ren C Luo, Pei H Lin, Yen C Wu, and CY Huang. Dynamic face recognition system in recognizing facial expressions for service robotics. In *IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 879–884, 2012.

[55] 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.

[56] Davide Marengo, Fabrizia Giannotta, and Michele Settanni. Assessing personality using emoji: An exploratory study. *Personality and Individual Differences*, 112:74–78, 2017.

[57] David Matsumoto and Bob Willingham. The thrill of victory and the agony of defeat: spontaneous expressions of medal winners of the 2004 athens olympic games. *Journal of personality and social psychology*, 91(3):568, 2006.

[58] Derek McColl, Alexander Hong, Naoaki Hatakeyama, Goldie Nejat, and Beno Benhabib. A survey of autonomous human affect detection methods for social robots engaged in natural hri. *Journal of Intelligent & Robotic Systems*, 82(1):101–133, 2016.

[59] Derek McColl and Goldie Nejat. Recognizing emotional body language displayed by a human-like social robot. *International Journal of Social Robotics*, 6(2):261–280, 2014.

[60] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[61] Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. “blissfully happy” or “ready tofight”: Varying interpretations of emoji. In *Tenth international AAAI conference on Web and social media*, 2016.

[62] 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, 2017.

[63] 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*, pages 137–144, 2016.

[64] Tommaso Querci, Yaocong Duan, Robin AA Ince, Chaona Chen, Lotta K Peussa, Oliver GB Garrod, Philippe G Schyns, and Rachael E Jack. Distinct face movements boost recognition of facial expressions of emotions. *Journal of Vision*, 20(11):1039–1039, 2020.

[65] Aisulu Rakhmetullina, Dietrich Trautmann, and Georg Groh. Distant supervision for emotion classification task using emoji 2 emotion. In *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji2018). Stanford, CA, USA. http://ceurws. org*, volume 2130, 2018.

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

[67] Guillermo Santamaría-Bonfil and Orlando Grabiel Toledano López. Emoji as a proxy of emotional communication. In *Becoming Human with Humanoid-From Physical Interaction to Social Intelligence*. IntechOpen, 2019.

[68] Shane Saunderson and Goldie Nejat. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics*, 11(4):575–608, 2019.

[69] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[70] Garima Sharma and Abhinav Dhall. A survey on automatic multimodal emotion recognition in the wild. In *Advances in Data Science: Methodologies and Applications*, pages 35–64. Springer, 2020.

[71] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061, 1987.

[72] Abu Shoeb and Gerard de Melo. Are emojis emotional? a study to understand the association between emojis and emotions. *arXiv preprint arXiv:2005.00693*, 2020.

[73] J Smailović, B Sluban, I Mozetič, et al. Sentiment of emojis. *Plos one*, 10(12):e0144296–e0144296, 2015.

[74] Dominique Vaufreydaz, Wafa Johal, and Claudine Combe. Starting engagement detection towards a companion robot using multimodal features. *Robotics and Autonomous Systems*, 75:4–16, 2016.

[75] Leticia Vidal, Gastón Ares, and Sara R Jaeger. Use of emoticon and emoji in tweets for food-related emotional expression. *Food Quality and Preference*, 49:119–128, 2016.

[76] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363, 2005.

[77] Sharon G Watson. Judgment of emotion from facial and contextual cue combinations. *Journal of Personality and Social Psychology*, 24(3):334, 1972.

[78] Matthias Wimmer, Bruce A MacDonald, Dinuka Jayamuni, and Arpit Yadav. Facial expression recognition for human-robot interaction–a prototype. In *International Workshop on Robot Vision*, pages 139–152, 2008.

[79] Chuang Yu and Adriana Tapus. Interactive robot learning for multimodal emotion recognition. In *International Conference on Social Robotics*, pages 633–642, 2019.

[80] Hui Yu, Oliver GB Garrod, and Philippe G Schyns. Perception-driven facial expression synthesis. *Computers & Graphics*, 36(3):152–162, 2012.

[81] Ke Zhong, Tianwei Qiao, and Liqun Zhang. A study of emotional communication of emoticon based on russell's circumplex model of affect. In *International Conference on Human-Computer Interaction*, pages 577–596. Springer, 2019.

Appendix A

Dataset

The data from this thesis can be downloaded from `www.rosielab.ca/datasets`.

Appendix B

Label Distribution Plots

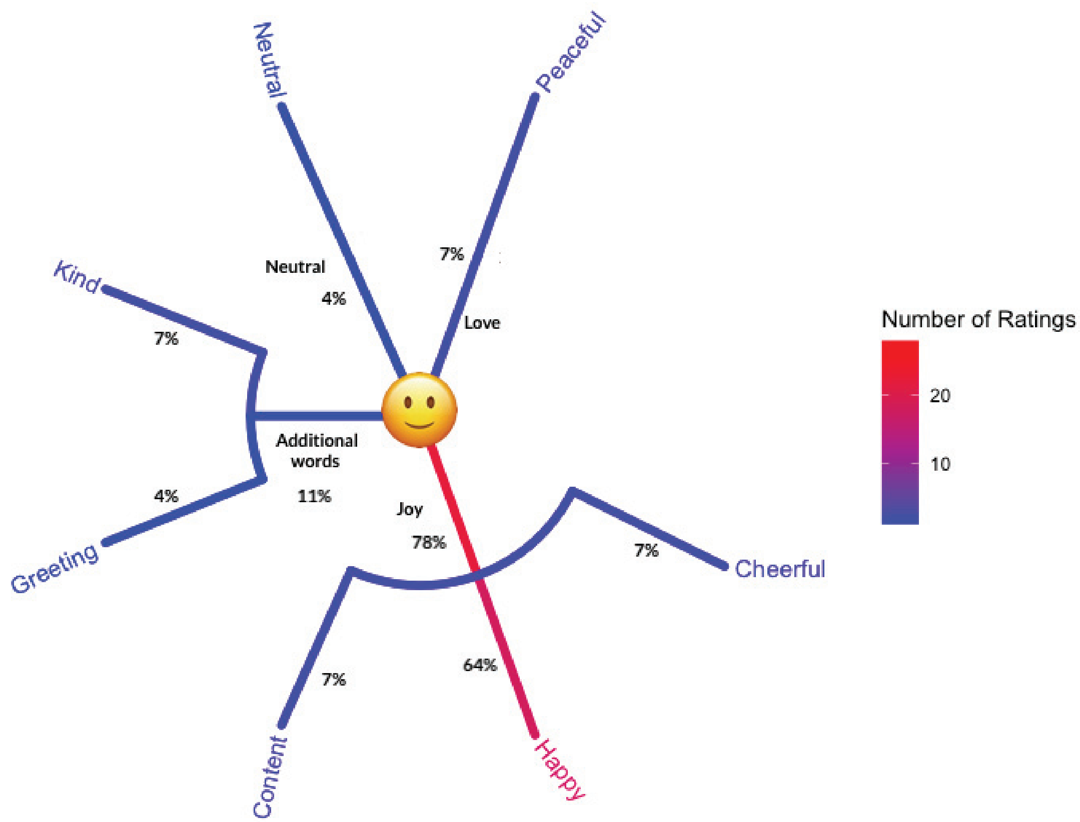


Figure B.1: Unimodal distribution for 🙂

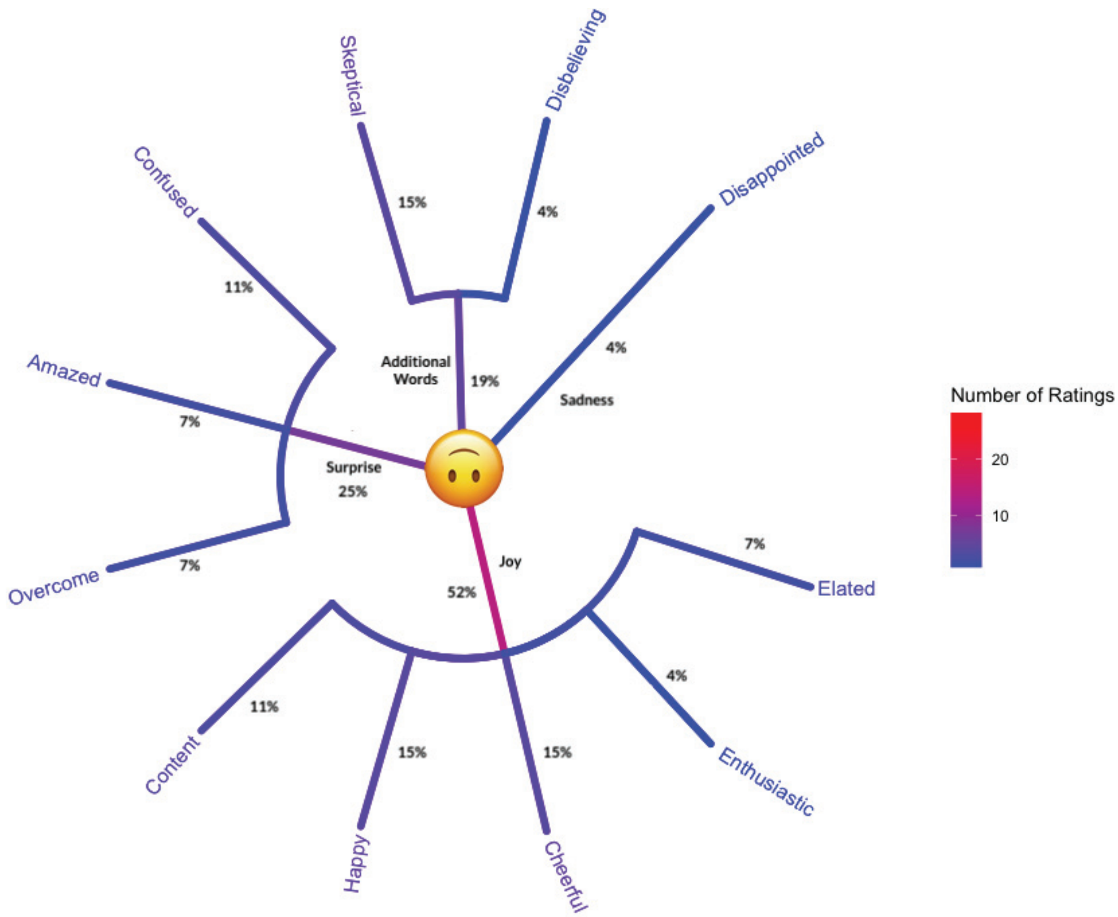


Figure B.2: Unimodal distribution for 🙃

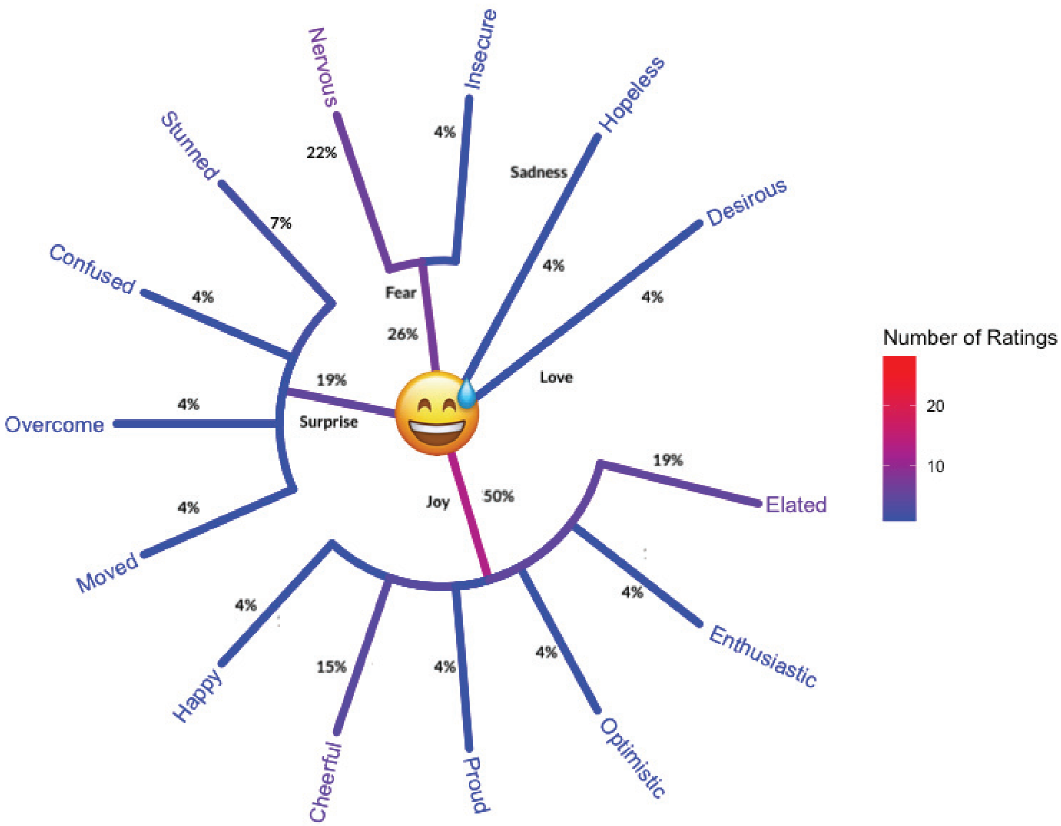


Figure B.3: Multimodal distribution for 😅

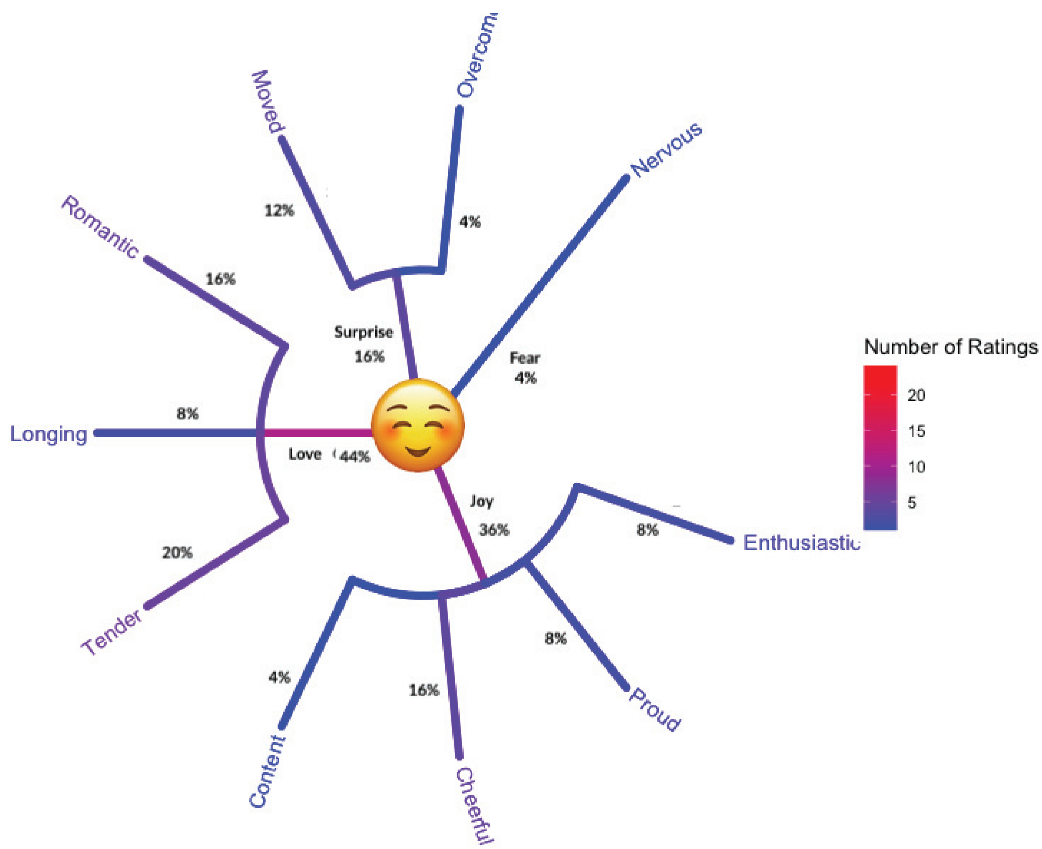


Figure B.4: Bimodal distribution for ☺️

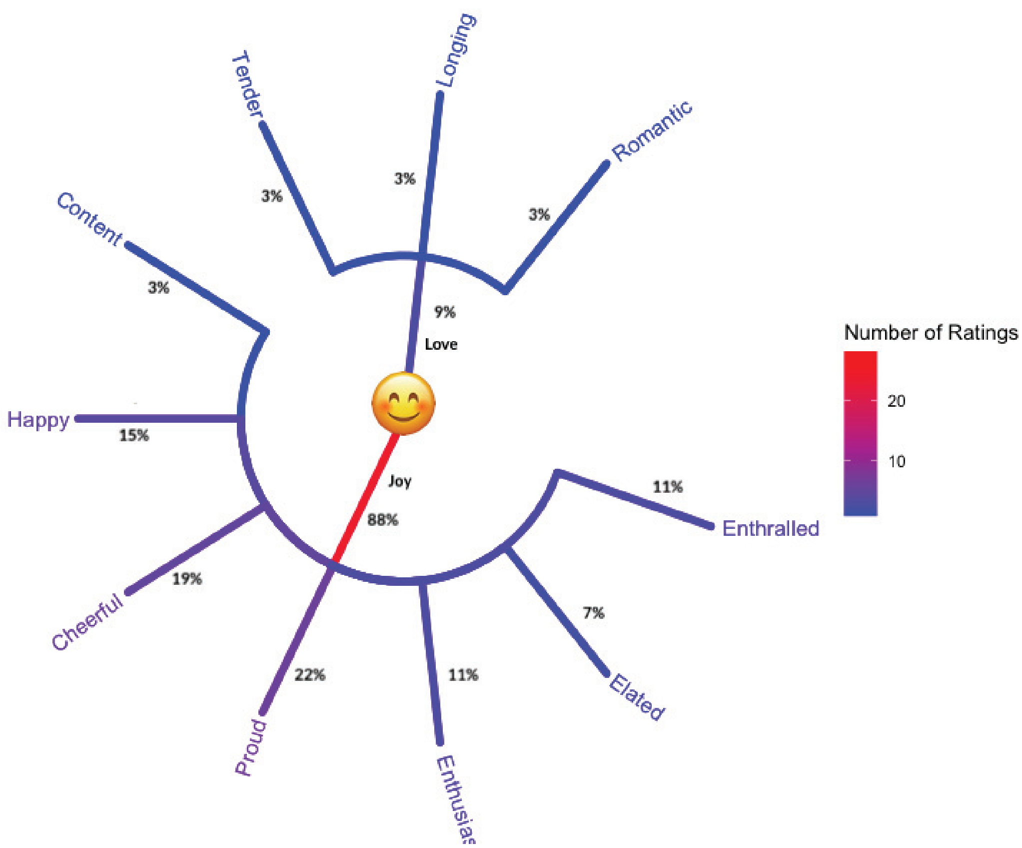


Figure B.5: Unimodal distribution for 😊

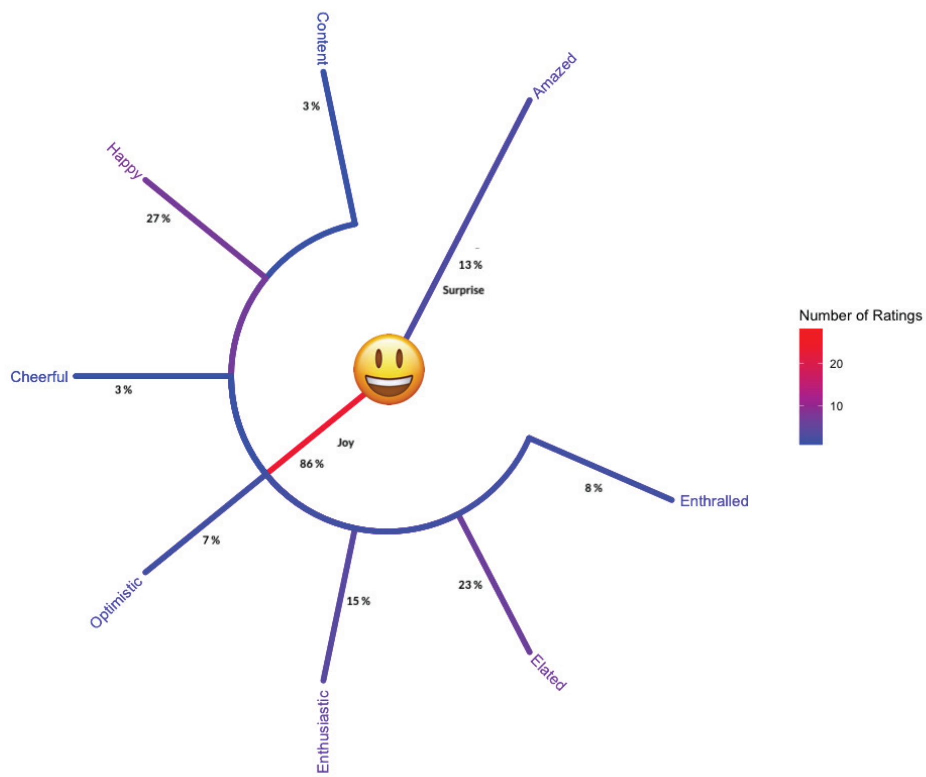


Figure B.6: Unimodal distribution for 😃

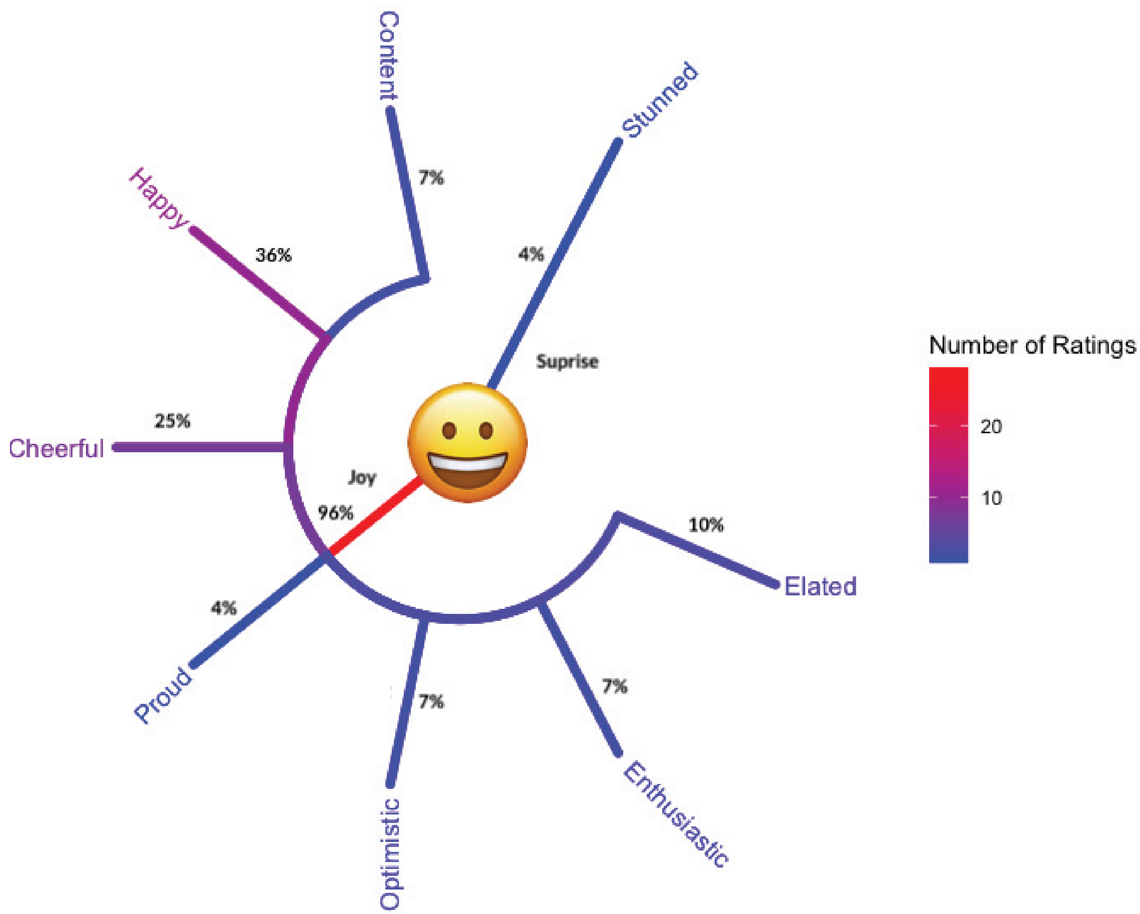


Figure B.7: Unimodal distribution for 😀

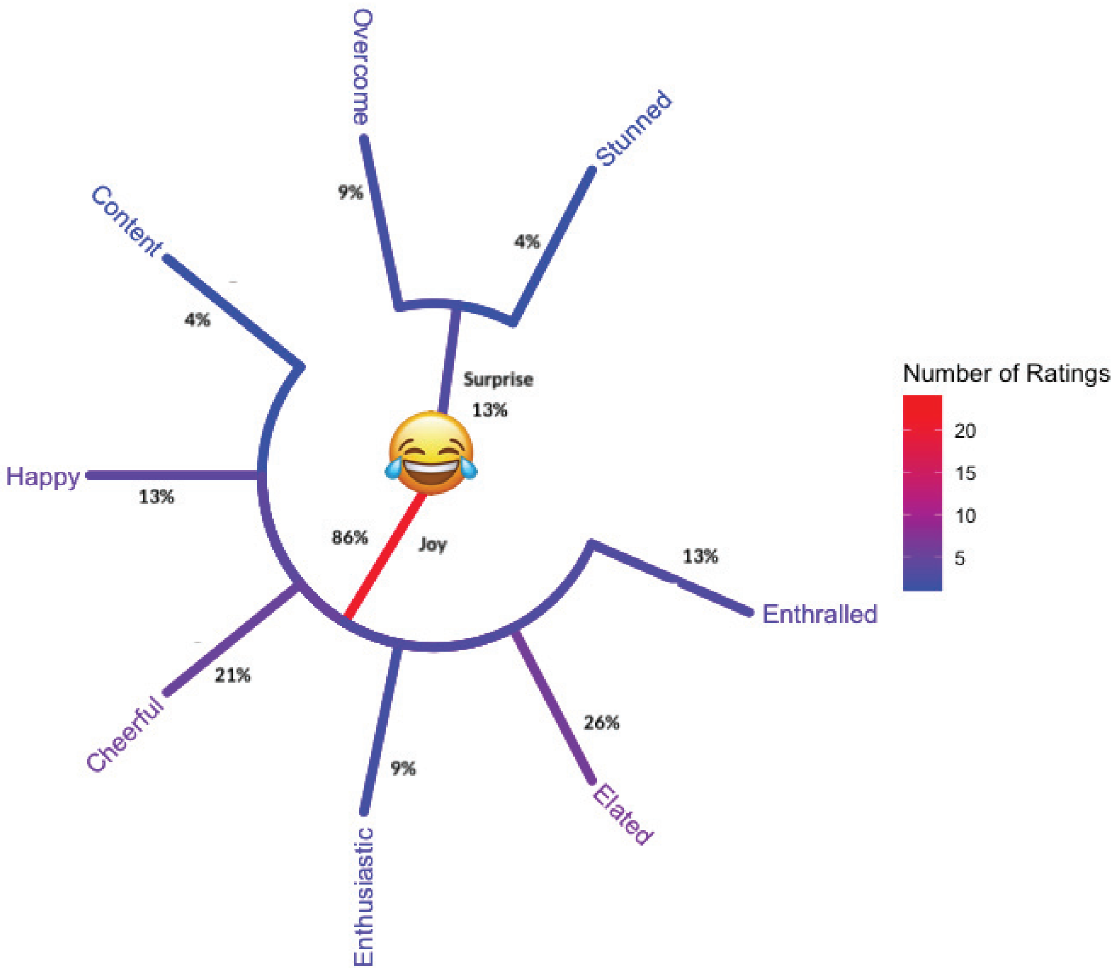


Figure B.8: Unimodal distribution for 😂

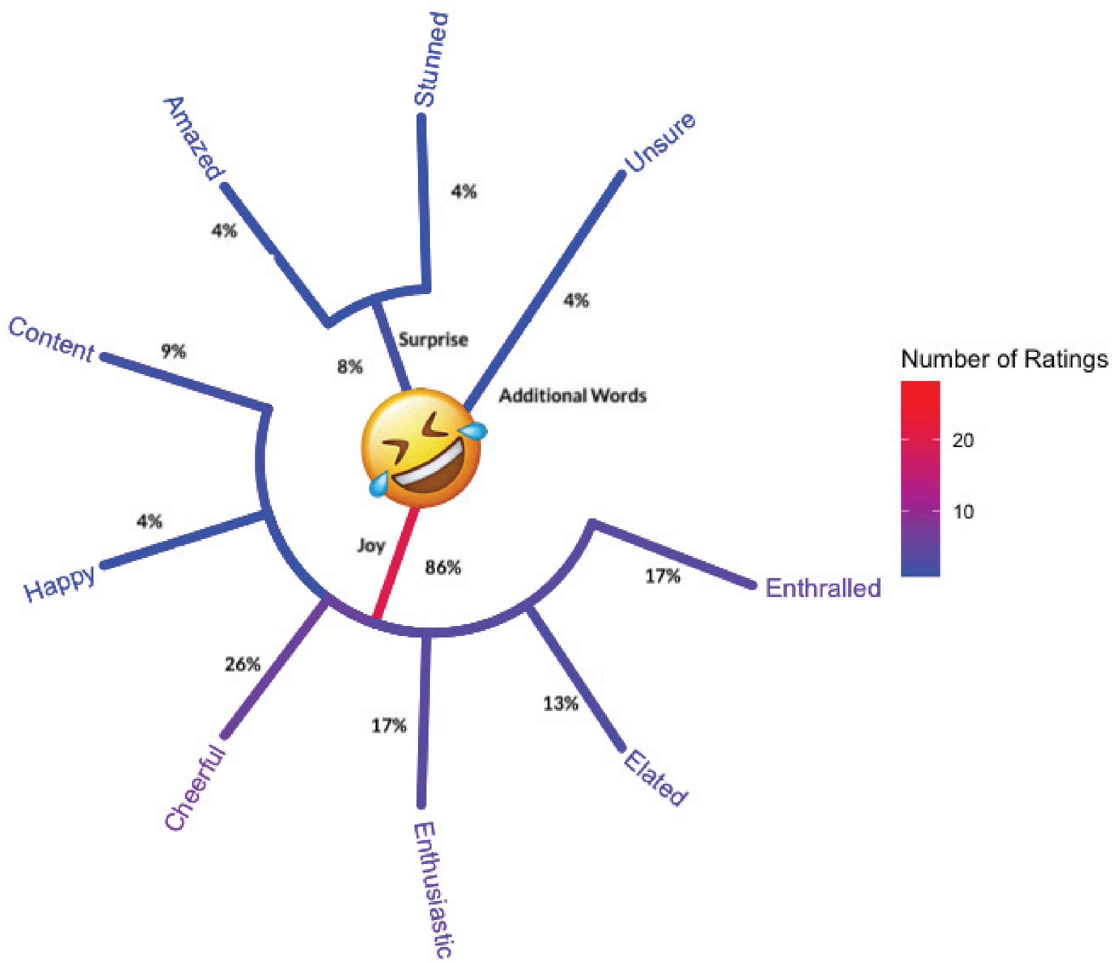


Figure B.9: Unimodal distribution for 🤣

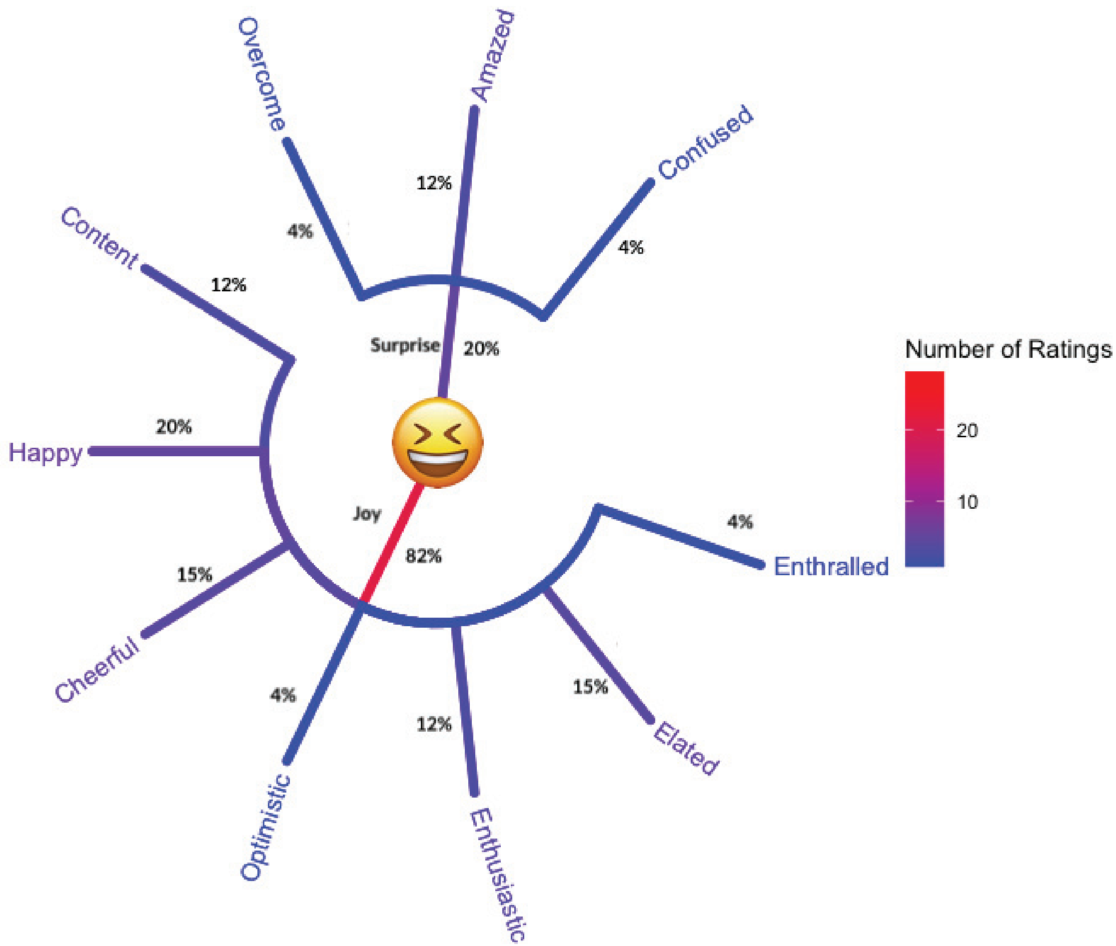


Figure B.10: Unimodal distribution for 😆

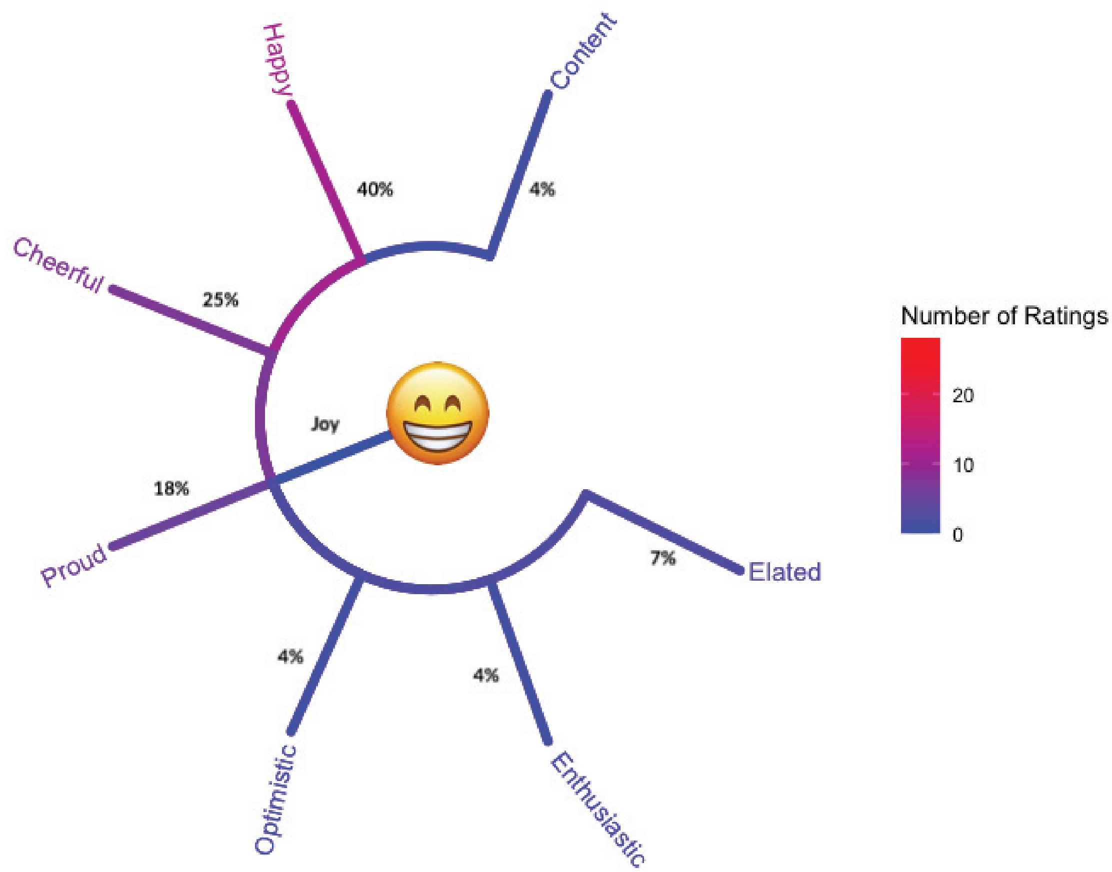


Figure B.11: Unimodal distribution for 😁

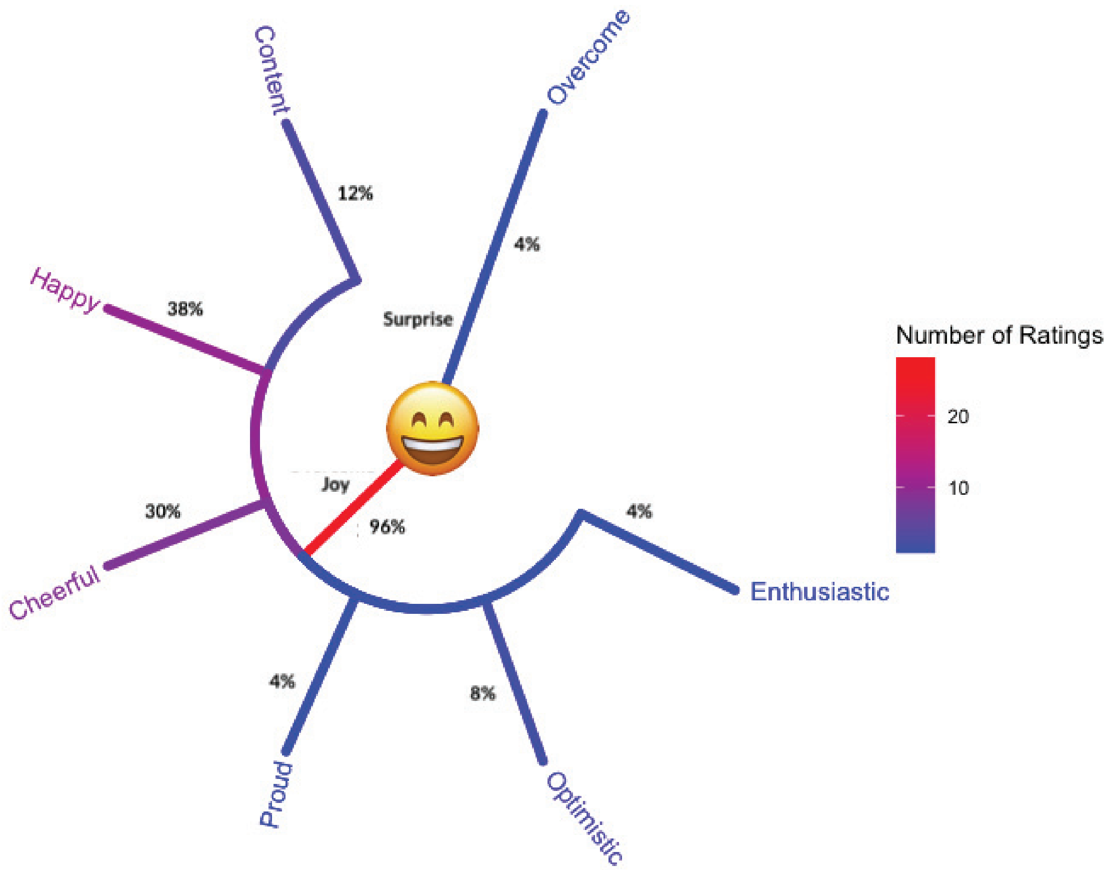


Figure B.12: Unimodal distribution for 😄

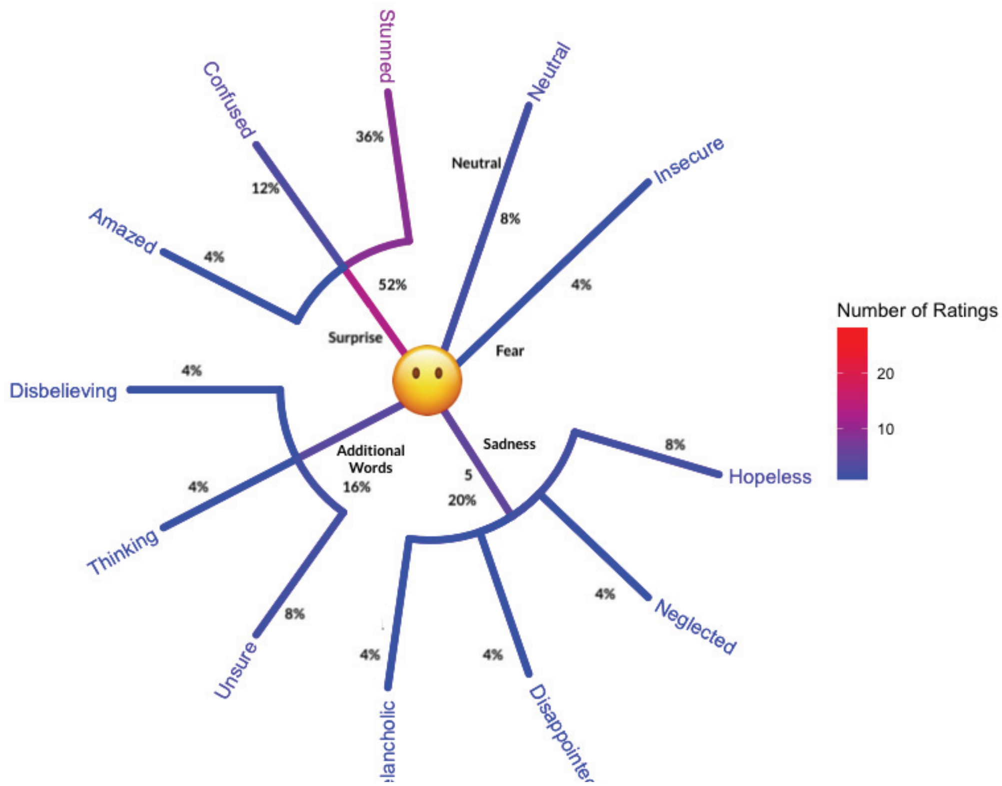


Figure B.13: Unimodal distribution for 😶

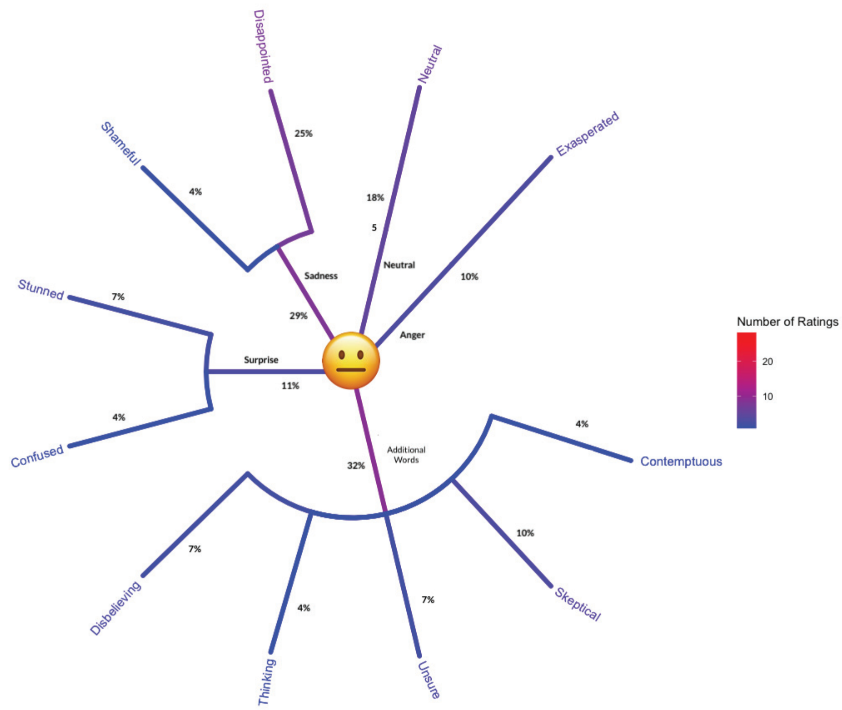


Figure B.14: Multimodal distribution for 😐

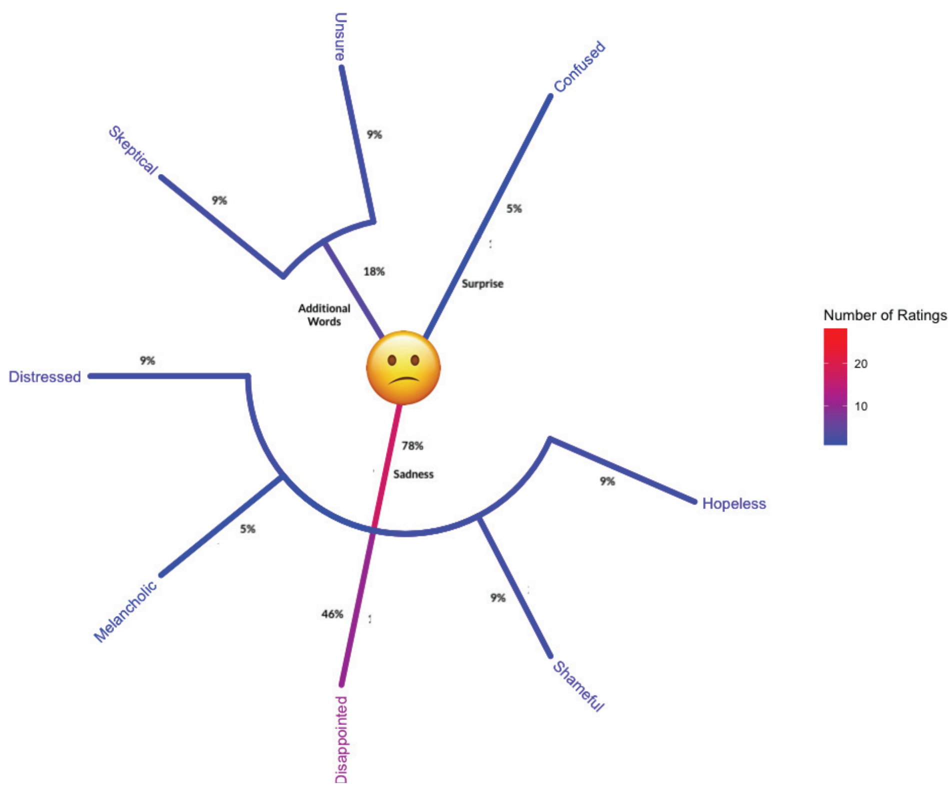


Figure B.15: Unimodal distribution for 🙁

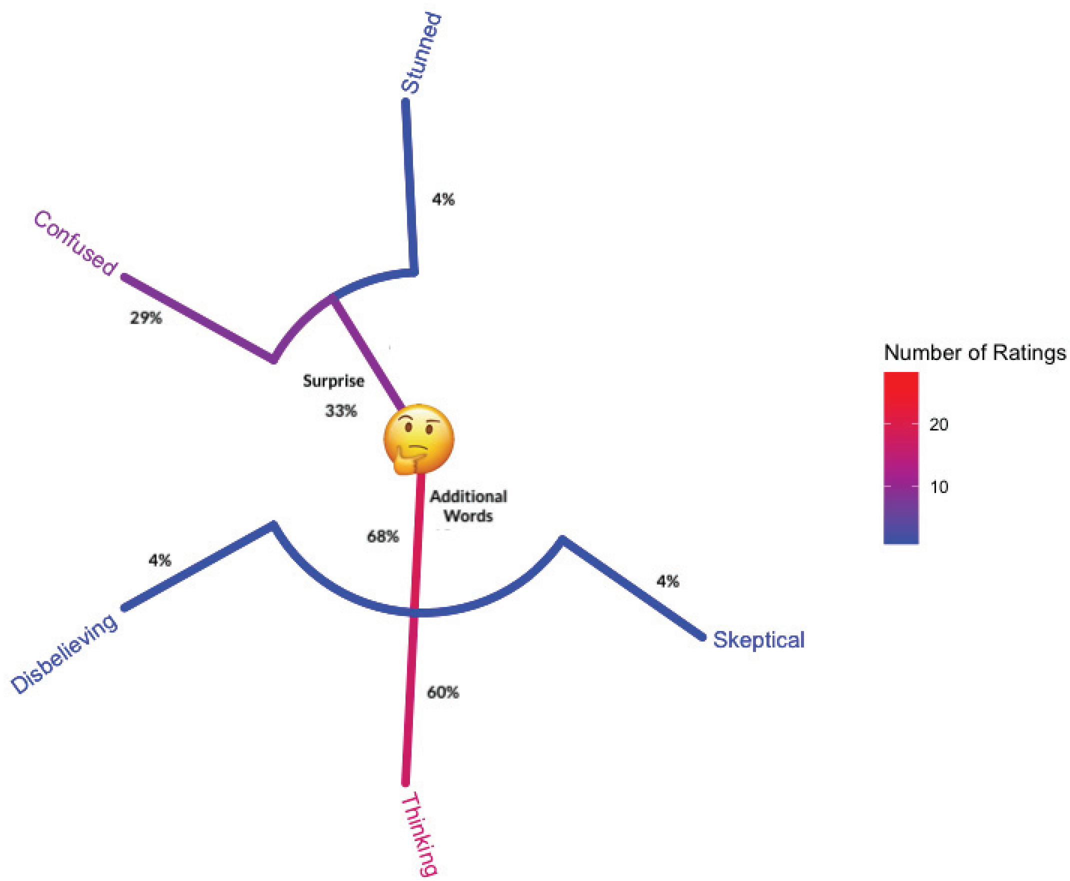


Figure B.16: Unimodal distribution for 🤔

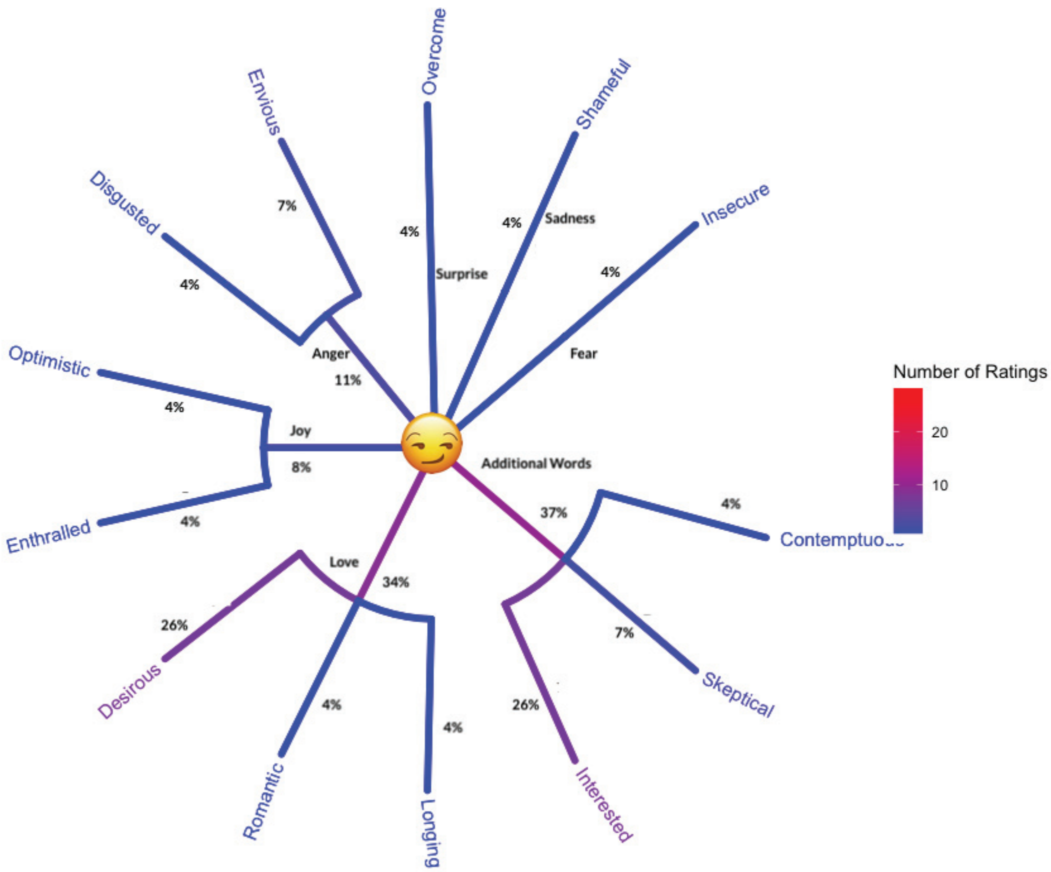


Figure B.17: Bimodal distribution for 😏

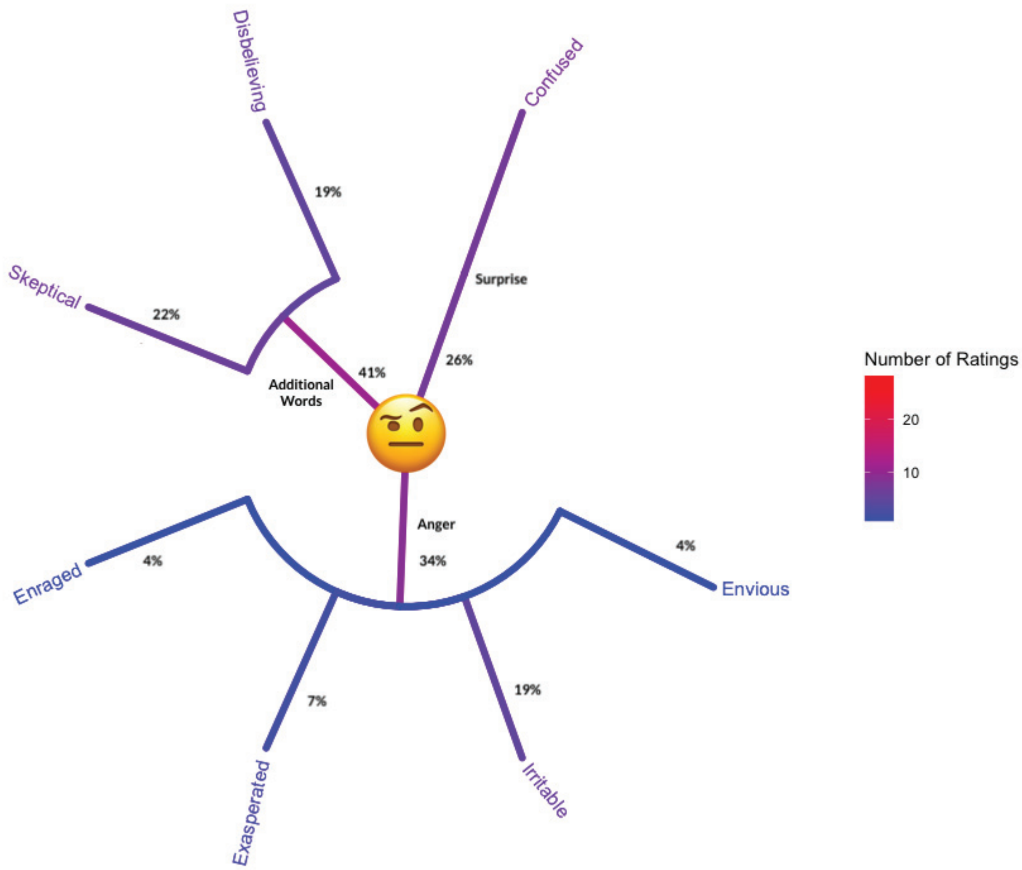


Figure B.18: Bimodal distribution for 🤨

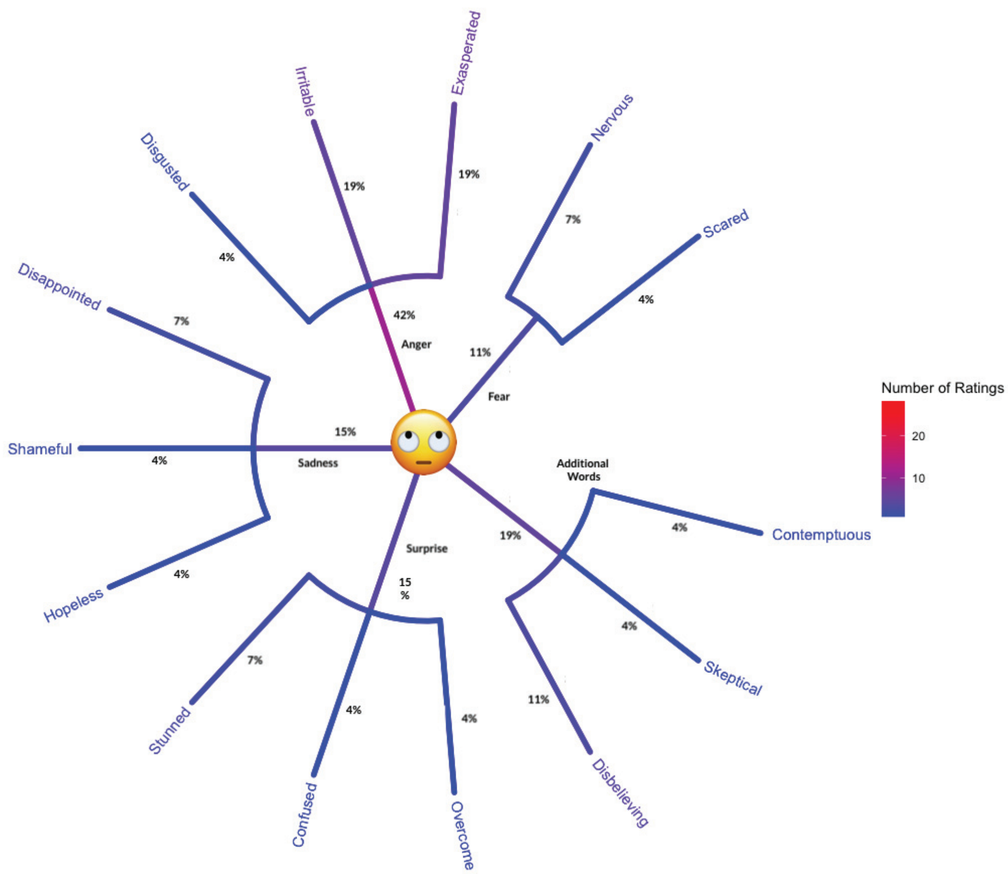


Figure B.19: Multimodal distribution for 🙄

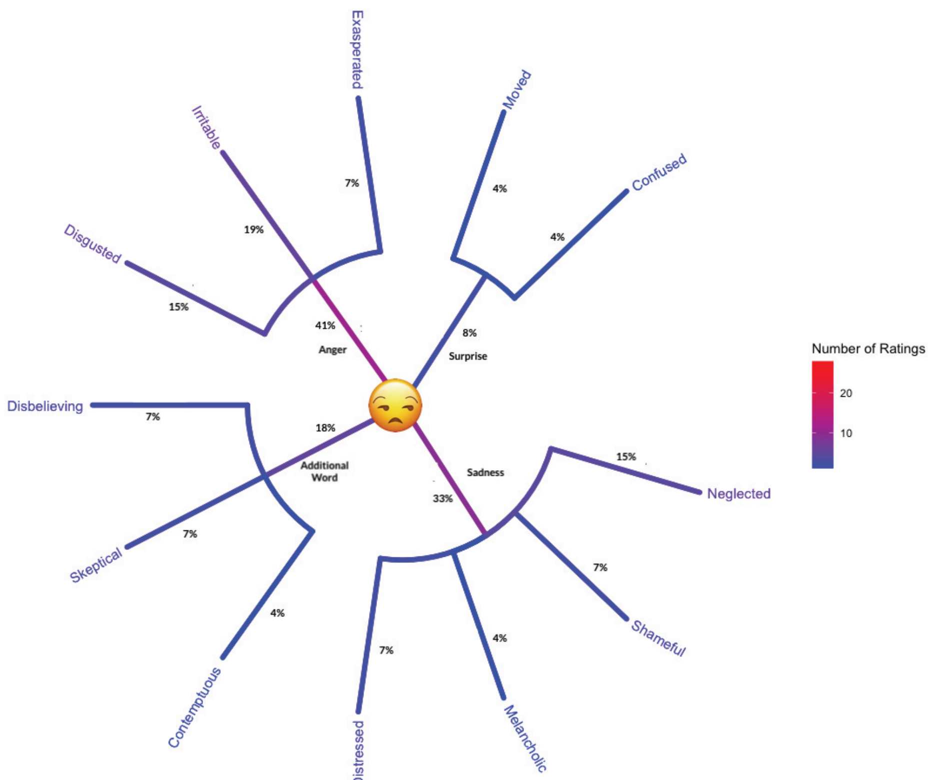


Figure B.20: Bimodal distribution for 😒

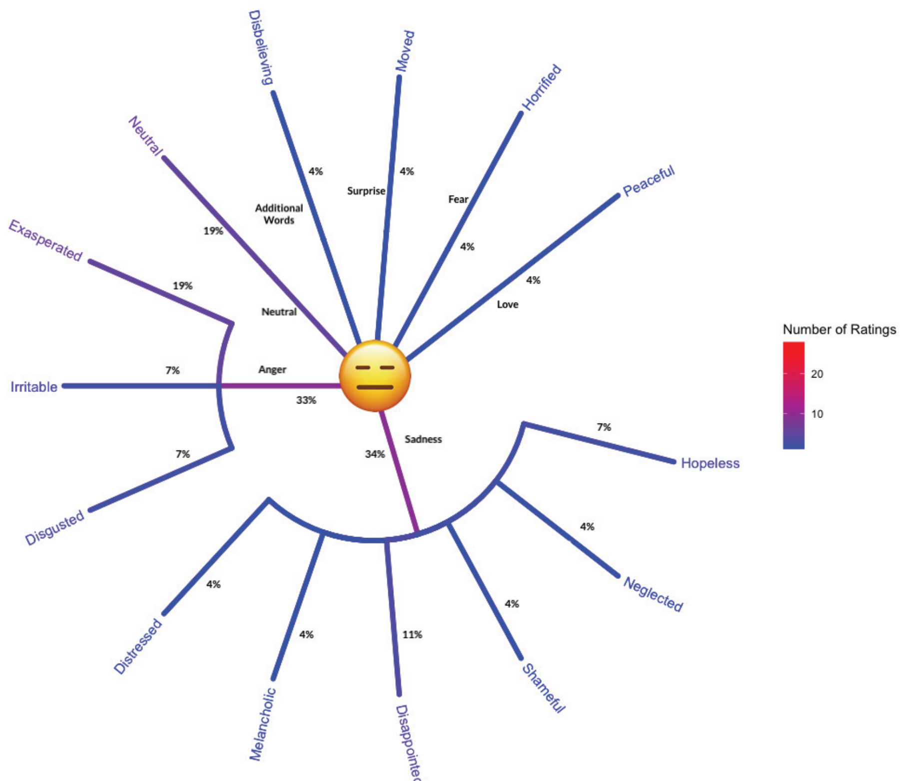


Figure B.21: Bimodal distribution for 😑

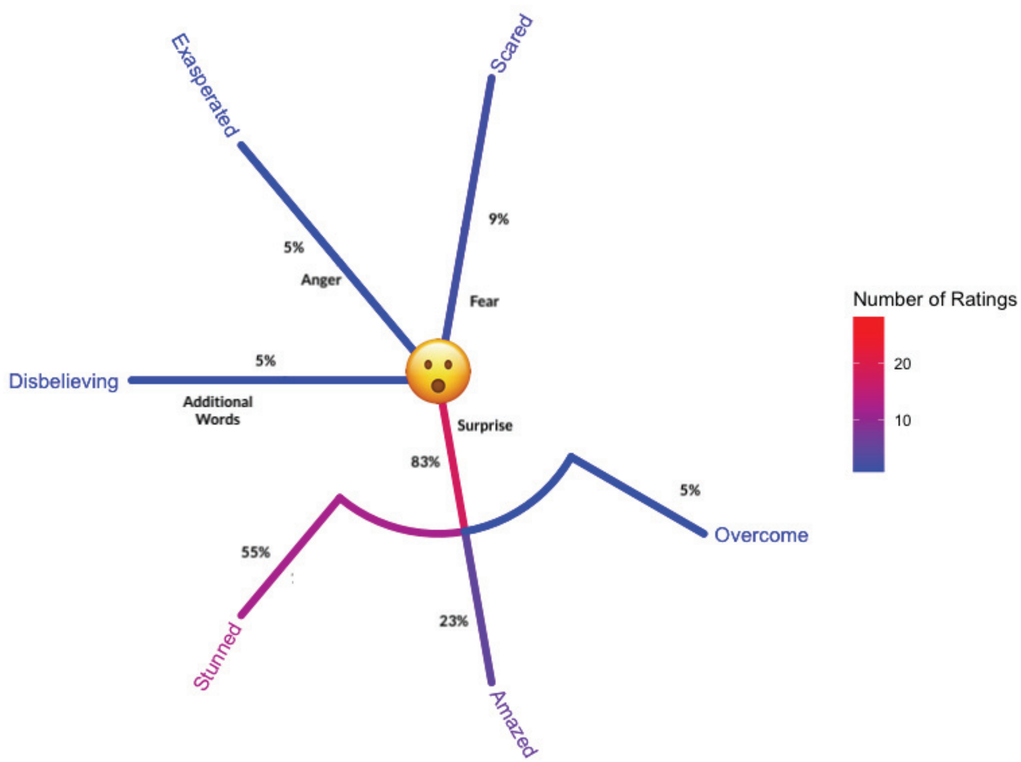


Figure B.22: Unimodal distribution for 😮

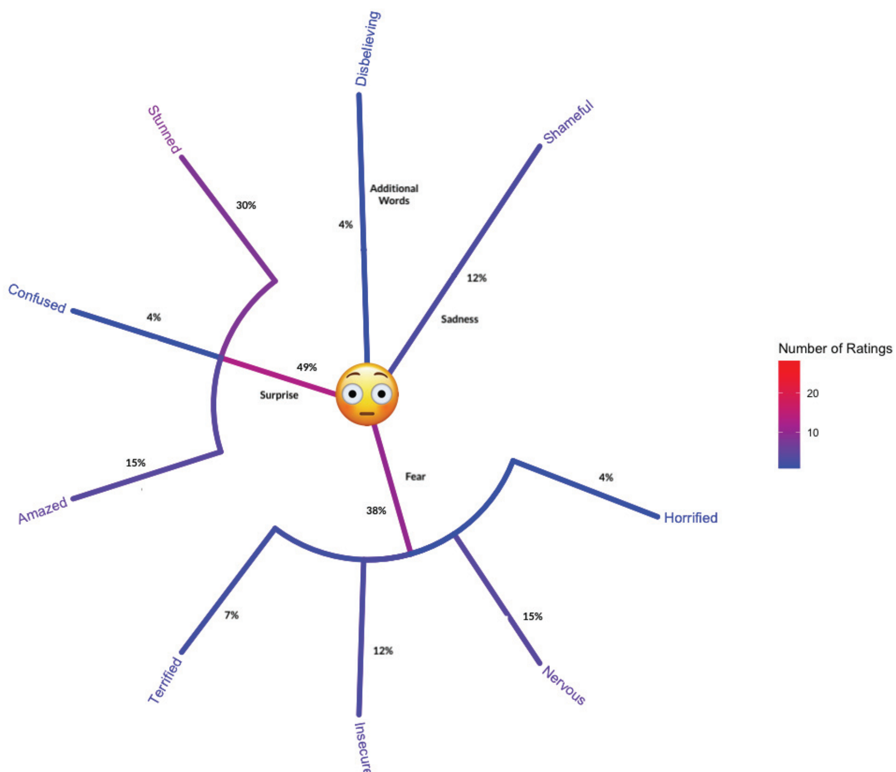


Figure B.23: Bimodal distribution for 😳

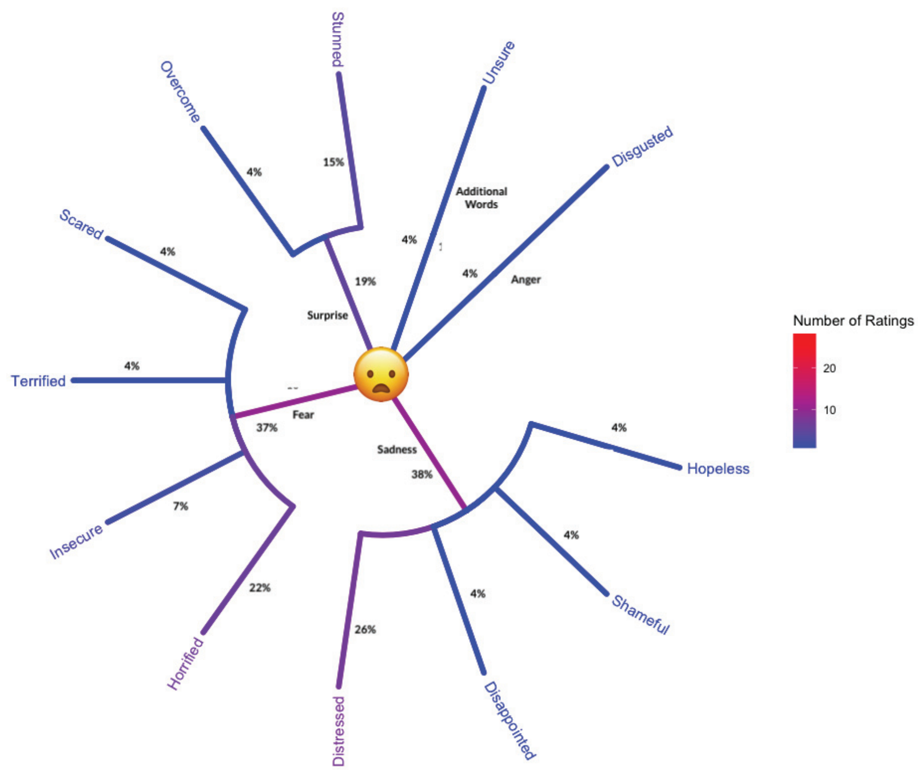


Figure B.24: Bimodal distribution for 😦

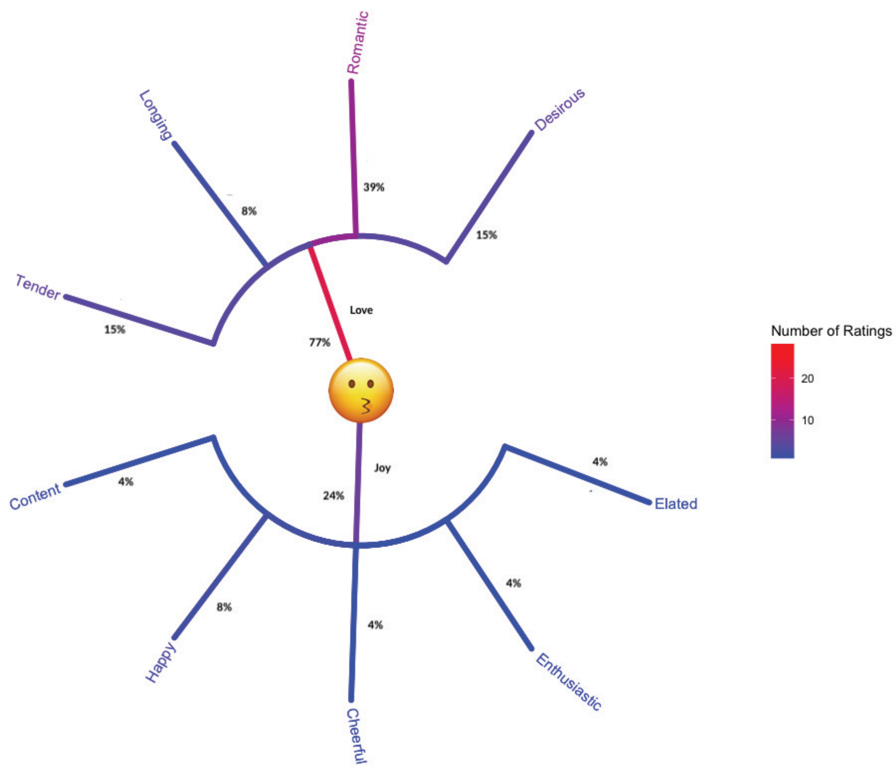
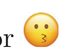


Figure B.25: Unimodal distribution for 

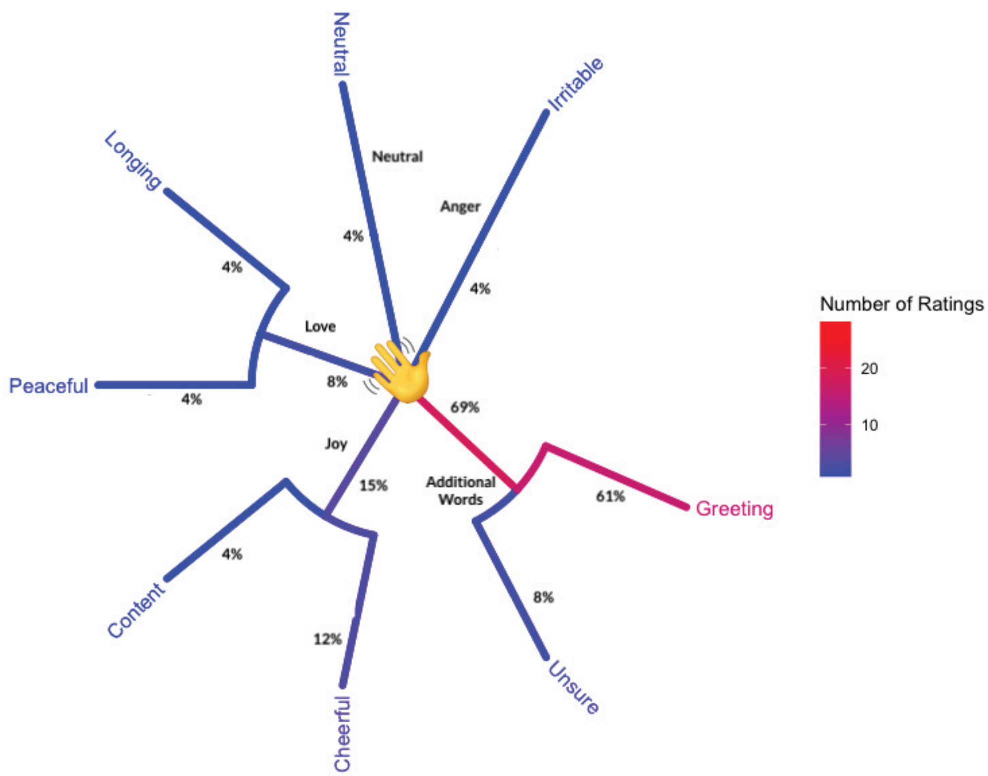


Figure B.26: Unimodal distribution for 👋

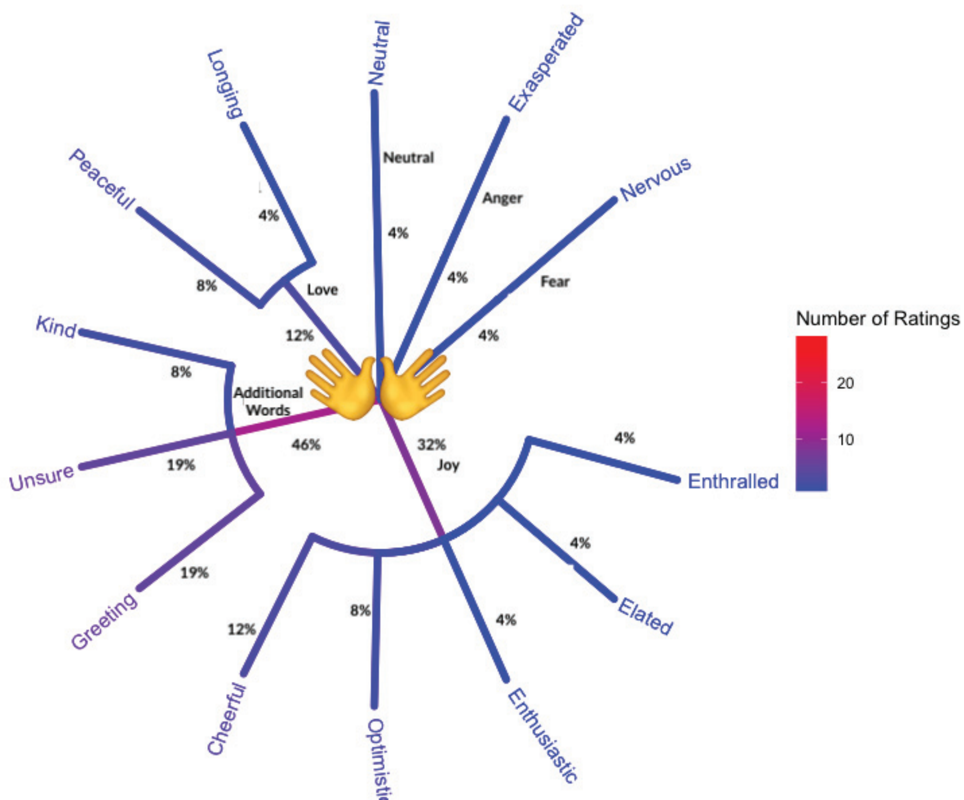


Figure B.27: Multimodal distribution for 👐

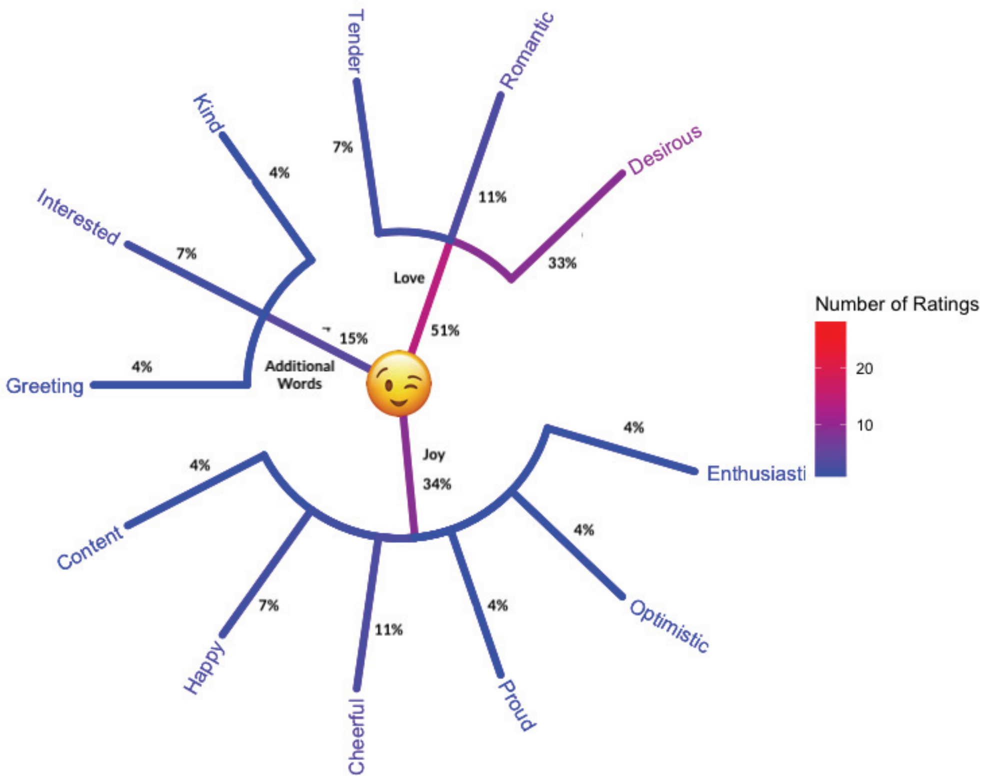


Figure B.28: Unimodal distribution for 😉

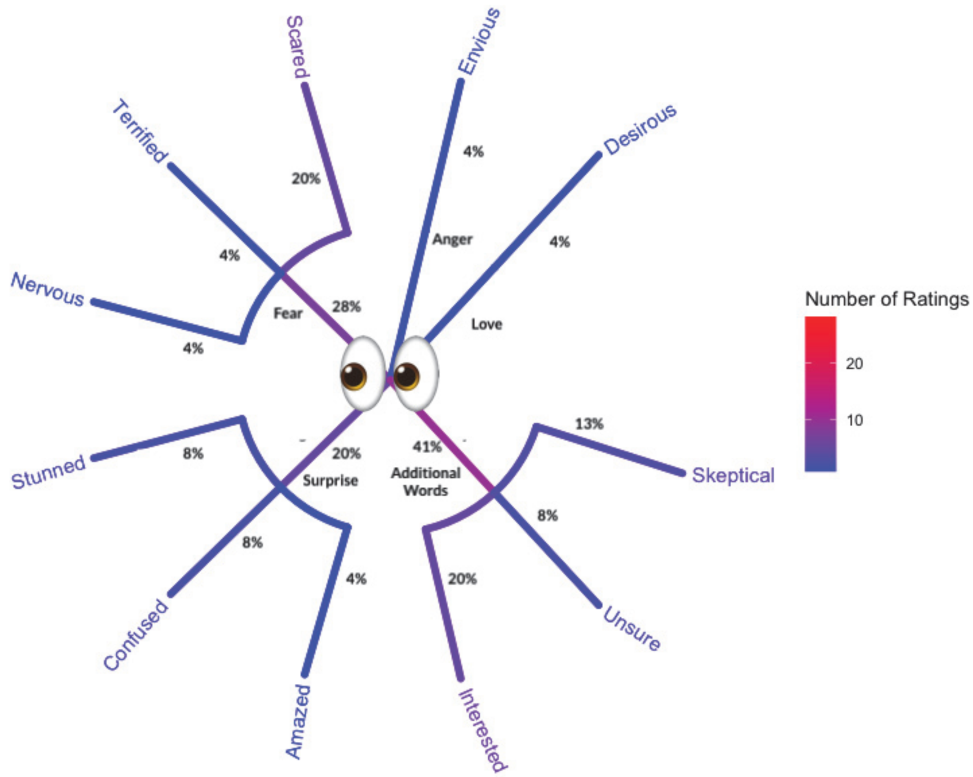


Figure B.29: Multimodal distribution for 👀