NAIST-IS-MT1551128

Master's Thesis

Study of Social-Affective Communication: Emotion Recognition, Emotional Triggers Prediction, and Dialogue Response Selection

Nurul Fithria Lubis

March 16, 2017

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Master's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
MASTER of ENGINEERING

Nurul Fithria Lubis

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Assistant Professor Sakriani Sakti	(Co-supervisor)
Assistant Professor Koichiro Yoshino	(Co-supervisor)

Study of Social-Affective Communication: Emotion Recognition, Emotional Triggers Prediction, and Dialogue Response Selection*

Nurul Fithria Lubis

Abstract

In a social setting, emotion works in two ways: a person is expressing their emotion, and also affected by their conversational counterpart. This creates a rich and dynamic social interaction between humans. It is argued that humans also impose emotional aspect when interacting with computers and machines. However, the majority of existing works have not yet fully considered the two-way role of emotion: the influence of others on how a person's emotion changes and fluctuates in an interaction. In this work, we attempt to utilize such knowledge to elicit positive emotion in Human-Computer Interaction (HCI). We identify the processes that amounts to the social-affective loop: emotion expression, recognition, emotional triggers, and response. Accordingly, we approach our objective incrementally through three main tasks: 1) Recognizing affective states, 2) predicting social-affective events, and 3) eliciting a positive emotional response. Furthermore, as emotional corpus is pre-requisite in each of these tasks, we construct a corpus of spontaneous social-affective interaction between humans. We successfully construct the components to perform each of the mentioned tasks. These components will serve as the building blocks for future integration into a system that will be able to offer emotional support through real-time interaction with users.

Keywords:

emotion, social-affective dialogue, human-computer interaction

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1551128, March 16, 2017.

Acknowledgements

I would like to express a heartfelt gratitude to the members of Augmented Human Communication Laboratory (AHC Lab), for taking me in and showing me the curious world of scientific research, right at the time of a major life-decision crossroad: what is next after my bachelor graduation. I would not imagine myself in this position if it had not been for that exciting summer research internship back in 2013. I thank the Japanese Ministry of Education, Culture, Sports, Science and Technology for providing support throughout this study through MEXT scholarship.

AHC Lab and the faculty have provided a highly research-conducive environment, for which I am grateful. I would like to thank Professor Satoshi Nakamura, for giving the balance between guidance and creativity in his research supervision, making the journey always challenging, interesting, and collaborative. My direct supervisor, Assistant Professor Sakriani Sakti, for the continuous support, insightful comments, and lively discussion, even during lunch breaks and off office hours. Assistant Professor Koichiro Yoshino, and former Assistant Professor Graham Neubig, for the discussion and constructive feedback on my progress reports and written works. Member of the thesis committee, Professor Yuji Matsumoto, for the questions and comments on my thesis and annual seminars. Each of the faculty I have worked with at NAIST has given me inspirations that shape the kind of researcher I hope to become in the future.

I would also like to thank Ms. Manami Matsuda, for her cheerful presence at AHC Lab, always with support and encouragement, and a helping hand (or even two) for a vast assortment of problems a foreign student can expect to encounter in Japan. The numerous interesting people I've met and became friends with, thank you for making the mundane much more bearable. The long-time friends, with whom I've had friendships old enough to enter school, thank you for sticking around. My family, a mix of strikingly different personalities, yet unanimous when it comes to support, thank you for letting me pursue this, even when it means being a few thousand kilometers apart. And Michael, for being not at all the shabbiest partner; for severely overfitting, and happily so.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Social-Affective Human Communication

The *appraisal theory of emotion* argues that most of our emotional experiences are the result of a cognitive process, unconscious or controlled, of evaluating situations and events [1,2]. Among the contributing factors, the social world is argued to be one of the important aspects that influence our appraisal processes [3]. Emotion of others can pose as clue as to how we should appraise a situation, or as an additional stimulus in how we appraise a situation [4].

As argued by Scherer [5], emotional competence can be broken down into three lower level competences that interact and depend on one another: appraisal, regulation, and communication competences.

1. **Appraisal competence** refers to the person's ability to accurately evaluate a situation. There are two sides of appraisal competence: 1) emotion differentiation, which is the ability to tell various kinds of emotion apart, and 2) internal emotion elicitation, which is the ability to appraise the appropriate emotional response, or the absence thereof, in a given situation.

2. **Regulation competence** refers to a person's ability to appropriately modify their raw emotion in an effective manner. This modification can be influenced by a number of complex factors, such as strategic intention, societal rules, or re-appraisal of the situation.

3. **Communication competence** refers to a person's ability to encode and

decode emotion into and from communication clues. This competence dictates the ability of someone to convey their feelings into others so as to be understood, as well as understanding others' emotional states.

In the social sphere, these competences govern two main processes: emotion perception and production.

1. **Emotion perception** refers to the process of recognizing emotion and understanding its implication. Two competences that play important roles in this process are 1) *communication competence* to decode an emotional state based on social clues, and 2) *appraisal competence* to relate it to their environment and situation, and appraise the resulting emotion accordingly.

2. **Emotion production** refers to the adaptive function of emotion that is essential in coping with events related to a person's well being [5]. Two main competences for this action are 1) *regulation competence* to efficiently modify the raw emotion according to re-appraisal, social rules, or strategic intentions, and 2) *communication competence* to actualize the processed emotion and project it to the environment.

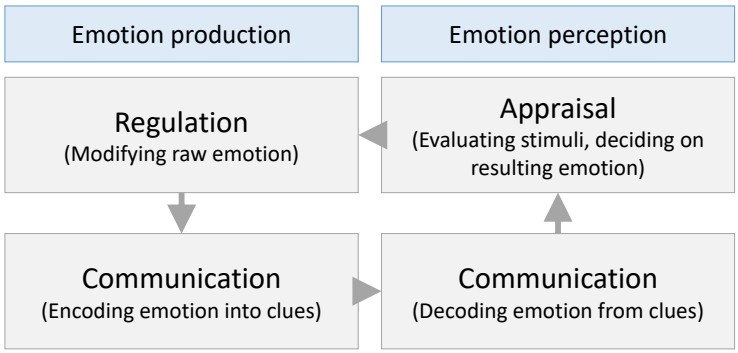


Figure 1.1: Emotional competences in emotion processes

Figure 1.1 illustrates these competences and processes and their underlying loop during social-affective communication. Conscious or not, we are constantly required to utilize these competences in any social interactions.

1.2. Affect in Human-Computer Interaction

It is argued that humans also impose the emotional aspect of social communications in their interaction with computers and machines [6]. They treat them politely, laugh with them, and sometimes get angry or frustrated at them. To mimic human interaction and benefit from its emotion-related potential, e.g. to provide emotional support, many works and studies have attempted to equip computers with emotional capabilities to reciprocate with humans in this regard.

The field of affective computing relates to, arises from, or influences emotion [7]. As in humans, it is believed that emotional competence in computers will enhance its quality of decision making and assistance. As such, research in this field is primarily dedicated to the incorporation of emotion into HCI. This main goal of emotional reciprocation demands the integration of many capabilities from a range of research topics, such as speech recognition, natural language understanding, emotion recognition, speech synthesis, and computer graphics. Over the years, many advancements have been made toward achieving this goal.

1.2.1 Related Works

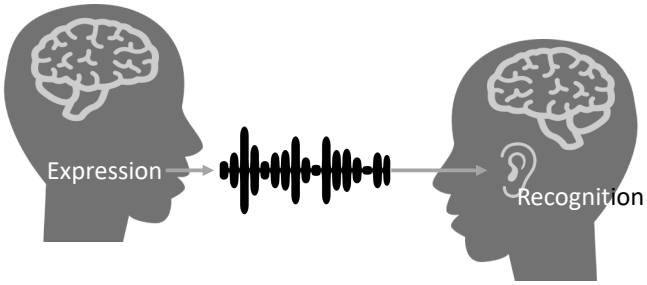


Figure 1.2: Traditional emotion works

As illustrated in Figure 1.2, two of the most studied issues in affect for HCI are:

- **Emotion recognition**, which allows a system to discern the user's emotions and address them in giving a response or performing their tasks [8,9]. In relation to emotional competences, this is the equivalent to communication competence on emotion perception, where we decode verbal and nonverbal cues into the underlying emotional state of the speaker [5] and use

the information accordingly. A study showed that when a tutoring system takes information of user's emotional state into account, task success rate can be significantly increased [10].

- **Emotion expression**, which helps convey a message to the user through emotional nuance. This is also equivalent to the communication competence, where we encode our feelings and emotional reaction into clues such that the listener can understand what we are feeling. In a listening-oriented system, this has been shown to increase closeness and satisfaction [11].

Recently, there has also been an increase of interest in the problem of emotion elicitation, an affective action prevalent in providing an emotionally beneficial interaction.

- **Emotion elicitation**, or **emotional triggers**, concerns eliciting a certain emotion from the user using the system's response. A recent study by Hasegawa et al. addresses this issue by predicting and eliciting emotion in online conversation [12]. The model is reported to be able to elicit a number of emotion classes properly by utilizing Twitter data and statistical machine translation techniques.

Figure 1.3 illustrates how this issue extends the traditional works on emotion.

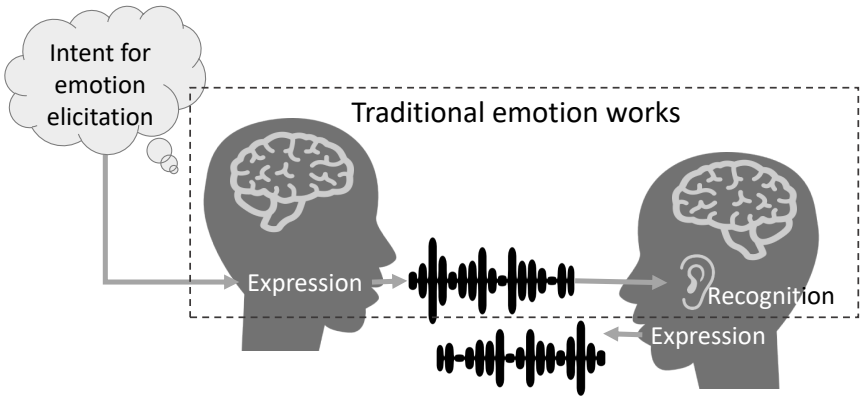


Figure 1.3: Emotion elicitation

At the core of the majority of the efforts in affective computing is the goal to help users meet their emotional needs through HCI, by taking into account

their emotional nature [13]. Many critics question the possibility of achieving this goal, given our lack of complete understanding of how emotion works in the first place. Picard et al. [13] argued against this skepticism, noting how humans routinely meet some of these needs through non-humans, for example pets, which have presumably less examined understanding of human emotions than we have today.

Indeed, a number of computational models of appraisal, personality, and relations are revealed to be successful in conveying emotional competences to the user. Skowron et al. constructed a dialogue system with positive, negative, and neutral affective profiles, showing consistent effect to the user compared to the respective profiles in humans [14]. Similarly, an evaluation of a conversational companion reported that the users felt that the companion had a personality; polite, friendly, and patient [15]. Even further, Bickmore et al. [16] reported affirming results in building and maintaining long-term human-computer relationships.

1.2.2 Challenges

To a large extent, emotion determines our quality of life. However, it is often the case that emotion is overlooked or even treated as an obstacle. The lack of awareness of the proper care of our emotional health has led to a number of serious problems, including incapabilities in forming meaningful relationships, sky-rocketing stress level, and a large number of untreated cases of emotion-related disturbances. In dealing with each of these problems, outside help from another person is invaluable.

The emotion expression and appraisal loop between interacting people creates a rich, dynamic, and meaningful interaction. When conducted skillfully, as performed by experts, a social-affective interaction can provide social support, reported to give positive effect with emotion-related problems [17]. Unfortunately, an expert is a limited and costly resource that is not always accessible to those in need. In this regard, an emotionally-competent computer agent could be a valuable assistive technology in addressing the problem. Bearing in mind the positive impact it could bring in real life application, this thesis attempts to focus on this issue.

As previously reviewed, a number of works have attempted to equip automated systems with the notion of emotion. These systems are able to communi-

cate emotion to and from the user through the usage of personalities, visual cues, as well as speaking styles. Furthermore, the incorporation of emotion is shown to help the systems perform their tasks better.

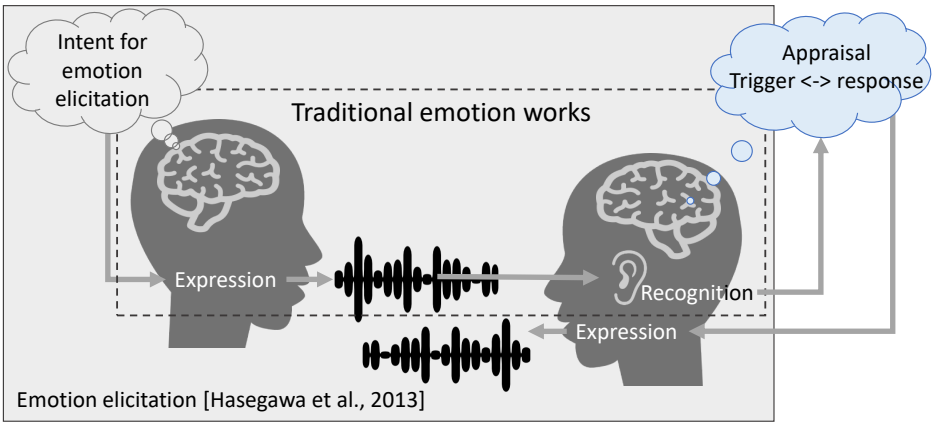


Figure 1.4: The role of appraisal in emotion elicitation

However, the majority of existing works have only focused on the felt part of emotion through communication competence. In other words, they have not yet observed the appraisal competence of humans during emotion perception. This entails the relationship between an utterance by the speaker, which acts as stimuli evaluated during appraisal, and the resulting emotion by the end of appraisal. Figure 1.4 illustrates this process in relation to existing works.

By examining the appraisal competence of humans, it would be possible to reverse the process and determine the appropriate trigger to a desired appraised emotion. The underlying knowledge and process of emotional triggers and responses is yet to be fully addressed, even though it has great potential in making a more emotionally positive HCI. High emotional competence is key in having an emotionally supportive interaction, the kind that benefits the participants and promotes positive emotion changes.

1.3. Thesis Objective and Contribution

This work aims to observe, learn, and utilize the social-affective knowledge of emotional triggers and responses to provide emotional support through HCI. Such

ability would be invaluable in assisting with emotion-related problems, as previously discussed in Section 1.2.2. As emotional support has a wide range of forms, in this thesis we focus on providing one by means of eliciting a positive emotional response.

Inspired by human communication, we approach this objective incrementally by addressing the processes that create the social-affective loop. As such, this thesis focuses on three main tasks:

1. **Recognizing affective states.** *How do we recognize emotion based on communication clues?* We attempt to recognize affective states using multimodal cues commonly available in social settings. This ability is essential in considering emotion in any interaction to provide the emotion context in further decision making.

2. **Recognizing social-affective events.** *How does emotion change? What causes this change?* We attempt to analyze social-affective conversations to study how emotion fluctuates in human conversation. We observe the change of emotion (emotional response) as well as its cause (emotional triggers) and use this information to predict their occurrence in new data.

3. **Eliciting a positive emotional response.** *Can we elicit a positive affective state in a dialogue system?* Lastly, using statistical approaches, we attempt to utilize the knowledge of emotional triggers-responses patterns to elicit a positive emotional response in the user.

Furthermore, as an emotional corpus is pre-requisite in performing these tasks, we construct a corpus of spontaneous social-affective interaction between humans. These tasks amount to a novel effort in considering emotional triggers and responses for an emotionally more positive HCI.

1.4. Thesis Overview

Figure 1.5 illustrates the overview of this work. Prior to conducting the experiments, we carry out preparatory tasks to obtain the necessary data and method. These are reported subsequently in Chapter 2 and 3. In Chapter 2 we discuss the construction of emotionally rich data used in this study. Afterwards, in Chapter

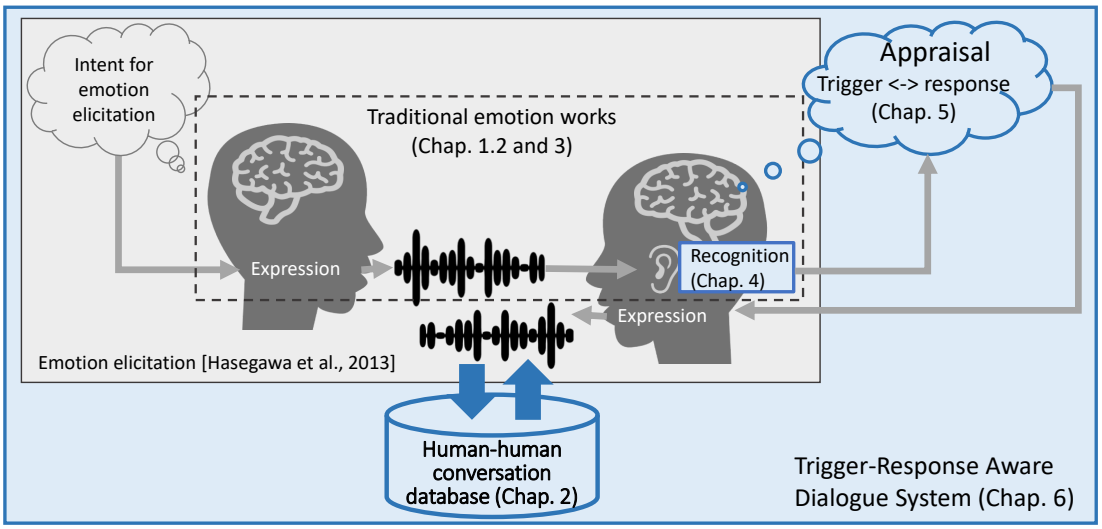


Figure 1.5: Thesis overview in relation to emotion processes and competences

3 we discuss the computational approach employed in modeling the problem and the solution.

We proceed by addressing our objectives through experiments, one by one in Chapters 4 through 6. Firstly, in Chapter 4, we construct an automatic emotion recognizer to demonstrate the ability to distinguish emotions. Secondly, in Chapter 5, we study the social-affective events of emotional triggers and responses by performing analysis and prediction. Afterwards, in Chapter 6, with a dialogue system we attempt to utilize the knowledge of social-affective events to elicit positive emotion in HCI. Each chapter starts with a discussion about the problem and related research in the field and concludes with a summary. We close the thesis with a conclusion and a discussion on future direction of the study in Chapter 7.

Chapter 2

Data Construction

Observation of complex phenomena of emotion requires rich sets of labeled data of natural interaction. Currently, many of emotionally colored speech corpora are constructed from acted speech or simulation [18, 19]. While this provides prominent emotion content, it does not represent natural occurrence of emotion. Furthermore, isolated emotional speech does not allow for the observation of emotion dynamics in an interaction.

To study emotion in social situation, we utilize spontaneous human interaction data. In particular, two kinds of interaction are examined: 1) induced emotion in laboratory environment, and 2) spontaneous emotion in the wild. In this section, we elaborate on the construction and properties of these corpora and argue for its suitability for this study.

2.1. Induced Emotion in Laboratory Environment: SEMAINE

2.1.1 About the Data

The SEMAINE Database is an annotated multimodal records of emotionally colored conversation between a person and a limited agent [19]. The corpus is collected from conversations in Wizard-of-Oz setting between two participants, one acts as the user and another acts as a wizard, posing as a Sensitive Artificial Listener (SAL). During the interaction, one restriction of the wizard is that it is unable to answer questions from the user.

A SAL is a limited agent designed to give the impression of attentive listening through verbal and non-verbal cues. In the corpus, there are four characters of SAL: cheerful Poppy, angry Spike, depressed Obadiah, and sensible Prudence. Each SAL responds to the user according to their characteristics, eliciting different reactions from the user, thus yielding an emotionally-colorful conversation.

The emotion contained in the corpus arises spontaneously, induced by the way the SAL behaves. In contrast to human-human interaction, the data shows how humans treat and react to the shortcomings of an automated agent. The corpus is designed specifically to represent interaction between a human and an automated agents. As such, the nature of the data provided is highly suitable for research in affect for HCI.

A participant posing as human user interacts with all 4 SAL characters, recorded in the span of 4 sessions (one session for each SAL). There are 24 participants posing as the user, and 4 participants posing as SAL. In total, 95 sessions of interaction is recorded. The sessions amounts to 45 hours of transcribed and annotated material.

2.1.2 Annotation

The majority of the recordings in the SEMAINE Database are fully transcribed, with time alignment according to the turn taking changes. Disfluencies (e.g. em, uh) are annotated as is, while laughter are assigned a special tag.

On the other hand, the emotion occurrences are annotated using the FEELtrace system [20] to allow recording of perceived emotion in real time. As an annotator is watching a target person in a recording (i.e. visual and audio information), they would move a cursor along a linear scale on an adjacent window to indicate the perceived emotional aspect (e.g. valence or arousal) of the target. This results in a sequence of real numbers ranging from -1 to 1, called a *trace*, that shows how a certain emotional aspect fall and rise within an interaction. The numbers in a trace are provided with an interval of 0.02 seconds.

The amount of annotation for the user and the SAL's clips differ in number. A small number of the SAL's clip are annotated by one to three annotators. On the other hand, for the user's clips, the majority have been annotated by 6 raters. The core annotation includes five emotion dimensions: valence, arousal, power, expectation, and intensity. Two reliability analyses, Quantitative Agreement

(QA) and correlational analysis, performed by the authors shows that 2/3 of the traces pass the stringent criterion of either analysis, and about 80% reach the level that are normally regarded as acceptable.

2.2. Spontaneous Emotion in the Wild: Television Talk Shows

2.2.1 About the Data

Although there has been an increase of interest in constructing corpora containing social interactions [21,22], there is still a lack of spontaneous and emotionally rich corpora. To bridge this gap, we construct a corpus of spontaneous social-affective interaction in the wild. We utilize various television talk shows containing natural conversations and real emotion occurrences. Interactions in talk show setting well represents typical social conversation, where small number of speakers are involved and various emotion-inducing topics are discussed. We construct our corpus both in English and Indonesian. As the data are of similar nature and setup, the inclusion of two languages allows the observation of social-affective communication across different cultures.

2.2.2 Data Collection

In English, we collect the data from three episodes of two of the most popular American television talk shows world wide. One of the shows talks about life experience of public figures, while the rest contain discussion of the struggles of families and negotiation in overcoming issues. Similarly, in Indonesian, we select three episodes from different kinds of talk shows to cover a broader range of emotions. The selected talk shows are very popular in the country, with discussions on engaging and interesting topics that trigger various emotions from the speakers. The first show is contains discussion focusing on politic related subjects. The second show concerns with topics in the area of humanities. The third show is has a lighter focus in celebrities, their career, and life. In each language, the different topics are expected to provide varied emotion content in the collected data.

In English, there are 12 speakers in total; 4 male speakers and 8 female. In Indonesian, there are 18 speakers in total; 12 male speakers and 6 female. In total, we collected English conversational data consisting of 1 hour, 2 minutes, and 19 seconds, and Indonesian consisting of 1 hour, 34 minutes, and 49.7 seconds of speech.

2.2.3 Annotation

In this section, we explain the annotation procedure of the corpus. We impose rigorous quality control to ensure the consistency of the results. We annotate the corpus in terms of emotion, dialogue act, speaker information, as well as speech transcription.

Procedure

In annotating the corpus, we bear in mind that language and culture affect how emotion is perceived and expressed in an interaction. We carefully select 6 annotators for the task, 3 for each language. Every annotator is required to be (1) a native speaker of the language used in the show, and (2) knowledgeable of the culture in the interaction of the show. With these requirements, we try to ensure that the annotators can observe emotion dynamics of the interaction to the furthest extent. To ensure consistency, we have each annotator annotate the full corpus.

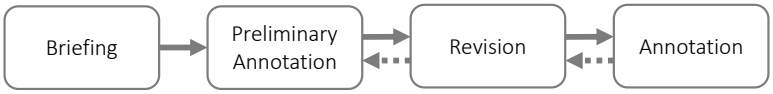


Figure 2.1: Overview of annotation procedure

Figure 2.1 gives an overview of the annotation procedure. Before annotating the corpus, the annotators are briefed and given a document of guidelines to get a clearer picture of the task and its goal. The document provides theoretical background of emotion and dialogue acts in discourse as well as a number of examples.

After the annotators are briefed, firstly, we ask them to do preliminary annotation by working on a small subset of the corpus. This step is done to let

them get familiar with the task. Furthermore, with the preliminary result, we are able to confirm whether the annotators have fully understood the guidelines, and verify the quality and consistency of their annotations.

We manually screen the preliminary annotation result and give feedback to the annotators accordingly. They are asked to revise inconsistencies with the guidelines if there are any. This process is repeated until the quality of the preliminary annotation is sufficient. Once their results are verified, the annotators are authorized to work on the rest of the corpus. We perform the same screen-and-revise process on the full corpus annotation to achieve a tenable result.

Dialogue Act Annotation

We perform the annotation on the sentence level[1] by determining its dialogue act. A dialogue act represents the meaning of an utterance at the level of illocutionary force [23]. The dialogue act annotation will provide information about the discourse structure, showing the relationship between dialogue turns.

We define a set of dialogue acts adapted from [24] to describe the structure of discourse. We reduce the original set of labels from 42 to 17 by grouping together similar labels, such as Yes-No-Question and Declarative Yes-No-Question. The 17 dialogue act labels are given in Table 2.1.

Table 2.1: Dialogue act labels

id	Dialogue Act	id	Dialogue Act
stat	Statement	rept	Repeat Phrase
opi	Opinion	ack	Acknowledgement
back	Backchannel	thnk	Thanking
Qyno	Yes-No Question	apcr	Appreciation
Qopn	Open Question	aplg	Apology
Qwh	Wh Question	hdg	Hedge
Qbck	Backchannel Question	drct	Directive
conf	Agree/Confirm	abdn	Abandoned
deny	Disagree/Deny		

[1] a sentence utterance refers to a continuous utterance of a speaker until 1) a full sentence is produced, or 2) proceeded by a different speaker

Emotion Annotation

As elaborated in Section 3.1, in this thesis we follow the circumplex model of affect as the computational model of emotion [25]. We define two emotion dimensions as the descriptor of felt emotion; valence and arousal. Valence measures the positivity or negativity of emotion while arousal measures the activity. With these axes, two sets of emotion labels are defined to allow observation from different perspectives. Following this model, the emotion annotation of the corpus consists of the level of valence (`val`) and (`aro`). The value of each dimension can be as low as -3 and as high as 3. This provides a granularity that balances between details of information and cognition load for the annotators.

2.2.4 Analysis

We inspect three properties of the corpus to gain better insight of the data contained within. We look over the conversational aspect of each language through the composition of dialogue acts in the corpus. Furthermore, we examine the quality of emotion annotation by looking at the inter-annotator label correlation.

It is important to keep in mind that these analyses do not provide conclusive differentiation between English and Indonesian languages due to the limited amount of data at hand and the differences in conversation topics between the two languages. However, they may give some idea about the different phenomenon and tendency that occur between the collected American-English and Indonesian television talk show broadcasts.

Length of Segments

We plot the distribution of number of segments according to their duration on Figure 2.2. In the figure, the y-axis shows the number of segments with respect to the x-axis, which shows the duration. Different trends can be observed in each language. The line graph for the English talk shows exceeds that of Indonesian for shorter durations, and then decreases heavily and has the value 0 from 16 seconds throughout the end. On the other hand, a long tail can be seen in the line graph for Indonesian.

The plot on Figure 2.2 shows that English speech segments are shorter on average compared to that of Indonesian. Respectively, the average durations of

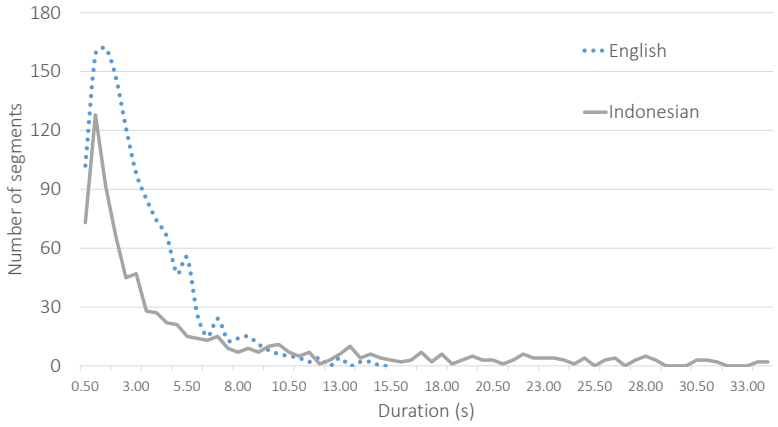


Figure 2.2: Number of segments in respect to the duration

segments for English and Indonesian are 2.93 and 7.57 seconds. This indicates that a dialogue turn in Indonesian tend to last longer than that in English. This difference of duration potentially affects the way emotion is communicated in a conversation.

Dialogue Act Labels

To analyze the consistency of annotation, we calculate Fleiss' kappa κ of the three annotators' results. κ measures the inter-annotator agreement of nominal variables when more than two annotators are employed. Respectively, English and Indonesian dialogue act annotations have κ of 0.54 and 0.45. According to interpretation of κ [26], both of the annotations are considered to have moderate agreement.

Figure 2.3 shows the composition of dialogue acts in the social-affective interaction on television talk shows. Statements dominate the collected conversations in both languages. This is not a surprising finding, as information exchange often happens in the form of a statement. However, it can be observed that in Indonesian, the composition is less dominated with statements than English. This could indicate more social-affective feedback and activity in the interaction, for example, in form of questions, confirmation, and back channel.

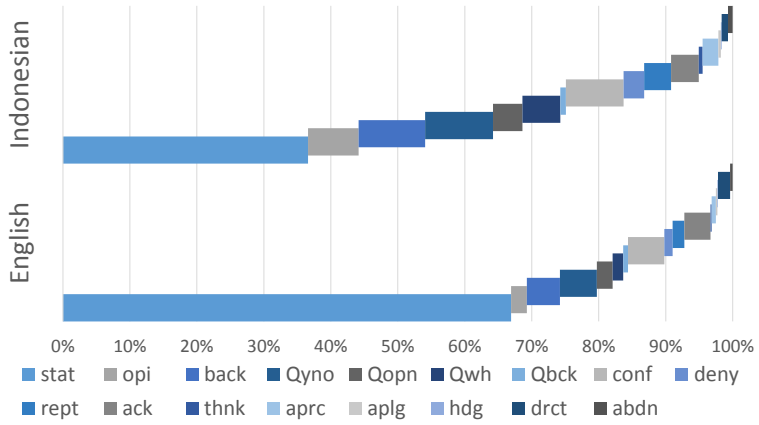


Figure 2.3: Composition of dialogue act labels

Emotion Labels

To analyze the annotation consistency, we calculate mean Pearson's correlation coefficients r of the three annotators for each language. Pearson's r measures the strength and direction of linear relationship between two variables. An absolute value of r between 0.0 and 0.3 is interpreted as weak correlation, greater than 0.3 up to 0.5 as moderate correlation, and higher than 0.5 as strong correlation. We observe moderate and strong correlation in the result of emotion annotation.

Table 2.2: Correlation coefficients of the emotion annotations

	`aro`	`val`
Indonesian	0.36	0.32
English	0.32	0.64

In general, the rates of correlation for the labels seem to be comparable for both languages, except for valence in English, which has significantly stronger correlation than the rest. Similar level of agreement for both dimensions in Indonesian suggests uniform capabilities in perceiving both dimension in the sentence segments. On the other hand, perception of valence in English appears to be stronger and more uniform across the annotators.

2.3. Summary

In this chapter, we discussed two emotionally rich corpora to be used in the following tasks. The first corpus, SEMAINE, contains induced emotion in a laboratory setting collected from Wizard-of-Oz interaction between a user and a wizard listening agent [19]. As such, the corpus contains data limited to the specific laboratory condition it was recorded in. On the other hand, the second corpus contained spontaneous real emotion occurrences collected from television talk show recordings. The data contains more variety, both in terms of emotion and content. The different nature of emotions contained within the corpora is beneficial for task-specific utilization, depending on the nature and the envisioned application of the resulting component.

Chapter 3

Computational Approaches for Affective Computing

3.1. Computational Models of Affect

A number of families of emotion theories that are proposed in literatures in psychology and neuroscience. Generally, these theories differ in terms of the aspects of emotion they include and highlight. An understanding of varying views of emotion will be invaluable in deciding for a model that is compatible to the problem that we are trying to solve.

Categorical emotion. The first family of models of emotion is *categorical*. In these models, a finite number of emotion categories are defined. One of the most adapted set of categories was proposed by Ekman, including happiness, anger, sadness, disgust, fear, surprise, and neutral [27]. These emotion are argued to be the most basic emotion that are universal regardless of culture or other social influences. On the other hand, Plutchik propsosed the wheels of emotion, containing 8 basic emotion and their derivatives, which are the secondary and tertiary emotions [28].

Dimensional emotion. The second family is *dimensional*, where emotion is seen as a point in an n-dimensional space, described using affective dimensions as the axes. The longest established affective dimensions are valence and arousal, proposed by Russel as the circumplex model of affect [25]. Since then, it has been argued that additional dimensions are needed to better distinguish certain types of emotions, e.g. fear and anger: power and expectancy [29]. However, the

decision on which affective dimensions to use remains strongly tied to the task at hand.

In this work, we define the emotion scope based on the *circumplex model of affect* [25]. Two dimensions of emotion are defined: *valence* and *arousal*. Valence measures the positivity or negativity of emotion; e.g. the feeling of joy is indicated by positive valence while fear is negative. On the other hand, arousal measures the activity of emotion; e.g. depression is low in arousal (passive), while rage is high (active). Figure 3.1 illustrates the valence-arousal dimension in respect to a number of common emotion terms.

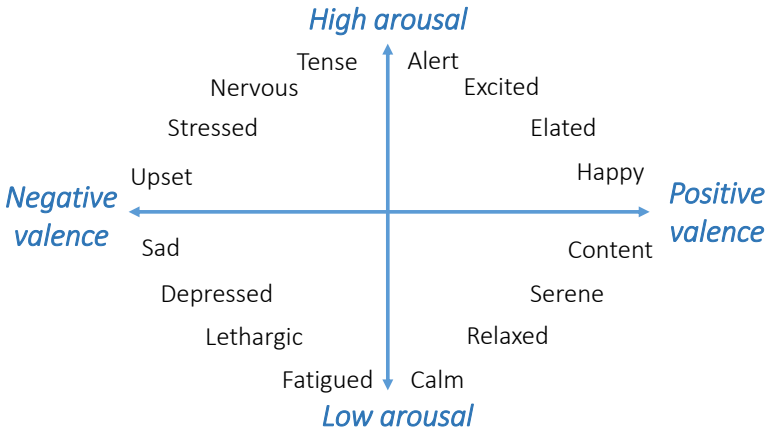


Figure 3.1: Emotion dimensions and common terms

This model highlight the perceived form of emotion, and is able to represent both primary and secondary emotion. An advantage of this model is that it is intuitive and also easily adaptable and extendable to either discrete or dimensional emotion definitions. The long established dimension are core to many works in affective computing and potentially provides useful information even at an early stage of research. Furthermore, this allows useful comparison to a wide range of related works.

3.2. Emotion-Rich Features

Humans channel their emotion through various modalities of communication. In this section, we discuss the features of each of these channels commonly used

for automated emotion-related tasks, in particular emotion recognition. In emotion recognition works, each of these channel of emotion-rich information is often referred to as a modality. Thus, emotion recognition with only one single modality, such as speech-based emotion recognition, is also called unimodal emotion recognition. In contrast, when two or more modalities are combined, it is called bimodal or multimodal emotion recognition, respectively.

3.2.1 Acoustic

One of the earliest effort in speech-based emotion recognition utilized only pitch-related features to distinguish between 4 emotion classes [30]. The majority of works in speech-based emotion recognition have utilized utterance level feature of a segment of speech as opposed to frame level [31, 32]. The extraction of such feature is largely based on the findings of Schuler et al. [33]. First, acoustic features, often called the Low Level Descriptor (LLD), is extracted at frame level for the entirety of the speech. Secondly, statistics of the frame level features, e.g. mean, standard deviation, etc., often called functionals, are computed to make up the global feature of the utterance. The acoustic features commonly used include spectral, e.g. Mel-Frequency Cepstral Coefficient (MFCC); cepstral, e.g. Mel-spectrum bins; energy; and voice related informations, e.g. F0, F0 envelope.

3.2.2 Semantic

Aside from acoustic information, textual features have also been utilized in emotion recognition tasks [34, 35]. Spoken content is widely known to contain information of the speaker's emotional state [36]. The usage of bag-of-words, bag-of-N-grams have been shown to work on emotion recognition task [37]. More recent works expand this approach by using lower level representation of the words, such as word vectors [38], or affective information of words, such as and WordNet affect [39].

3.2.3 Visual

Another salient clue about a person's emotional state lies on their facial expression. It is repeatedly observed that emotion-rich information can be obtained

through facial landmarks, especially the eyebrows, eyes, and mouth area. Commonly, these landmarks are represented by a set of points, and the position of these points in the face would differ according to the facial expression the person has, thus allowing the distinction of various emotional expressions. With video segments, the feature on each frame is commonly stacked, e.g. in [40] to produce the segment level features. Other approaches attempt to select the frames containing the emotional information prior to further processing [41–43].

Furthermore, Ekman et al. established the Facial Action Coding System (FACS) as a system to measure facial expressions a human being produces based on the anatomy of facial muscles and their movements [44]. With FACS, we can deconstruct any facial expression into specific Action Units (AU), which describe the contraction or relaxation of specific set of facial muscles. AUs can be thought of as higher level form of facial landmarks information.

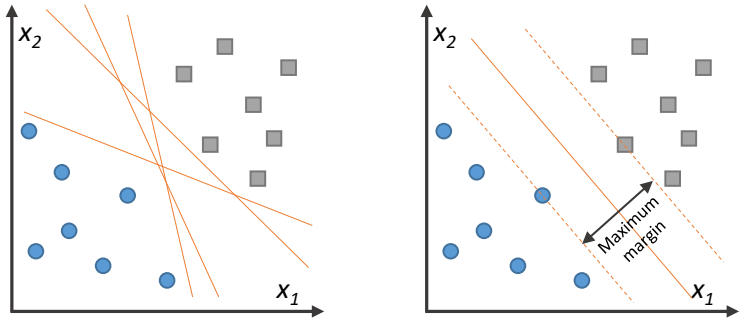
3.3. Support Vector Machines

A Support Vector Machine (SVM) is a discriminative classifier that separates data into two classes by using hyperplanes. An SVM is trained supervisedly, which is to say that the training data should contain information of the correct label. Given a training dataset, the algorithm outputs an optimal separating hyperplane that leaves as wide a margin as possible between the classes. The class of a new data will be decided according to where it falls in relation to the separating hyperplane.

For example, let's take a look at data points in a two-dimensional space, as illustrated in Figure 3.2 (a). As we can see, there are many ways to separate the data points into two classes. However, the optimal separation, as illustrated in Figure 3.2 (b), is one where the distance between the two closest data points (one from each class) to the separating hyperplane is as wide as possible. The training examples closest to the hyperplane is also known as the *support vectors*.

In practice, there often exists no separating plane that can provide a clear separation between classes in a dataset. In such cases, a new feature vector representation, commonly in a very high dimensional space, of the data is computed to create better separation of the data points. In this data projection process, a kernel function is used to provide a fast and efficient way of computation.

Instead of computing the new high-dimensional feature vector, the kernel



(a) Possible separations (b) Optimal separation

Figure 3.2: Separating plane in 2D space

function computes similarity between the data points and a number of selected landmarks. The kernel function allows the model to operate in a high-dimensional space by using implicit features, i.e. without having to actually compute the position of the data points in such space, but instead by simply computing the inner product of all pairs of the data. A commonly used kernel function is the Gaussian kernel or the Radial Basis Function (RBF) kernel.

The RBF kernel on two data points $\mathbf{x}$ and $\mathbf{x}'$, represented as feature vectors in some input space, is defined as

$$K(\mathbf{x}, \mathbf{x}') = \frac{\exp\left(\|\mathbf{x} - \mathbf{x}'\|\right)^2}{2\sigma^2} \tag{3.1}$$

where σ is a free parameter. The feature space of the RBF kernel has an infinite number of dimensions.

SVMs are naturally two-class classifiers. However, it is still possible to utilize the algorithm on multi-class problems. A common way to adapt the algorithm to such problems is to build one-against-all classifiers for each of the class. One-against-all classifiers distinguish one class from the rest, which is to say they classify data into two classes: belonging to the class or not. This approach turns the multi-class classification problem into several binary ones. A new data is classified into the class where to which hyperplane it has the largest margin. Another approach is to build one-against-one classifiers, and then choose the class to which a new data is classified to most often.

Furthermore, SVM can also be extended to solve regression problems [45]. This model is often called Support Vector Regression (SVR). In SVR, the input is first mapped into n-dimensional feature space with non-linear kernel mapping.

Afterwards, a linear model is constructed in this feature space. SVR maintains all the main features of the maximal-margin algorithm of SVM. In SVR, a new type of loss function is introduced which is not dependent on the dimensionality of the space.

3.4. Artificial Neural Networks

An Artificial Neural Network (ANN, or NN) is a system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic response to external inputs [46]. This network is inspired by the biological brains of humans, which are made up of a large amount of neurons forming numerous clusters. Each neuron is connected to many other, and its state affects the activation states of other units connected to it. Similarly, an ANN is made up of layers of perceptrons that are interconnected to one another. Essentially, this model is a function that maps an feature vector x to its label y.

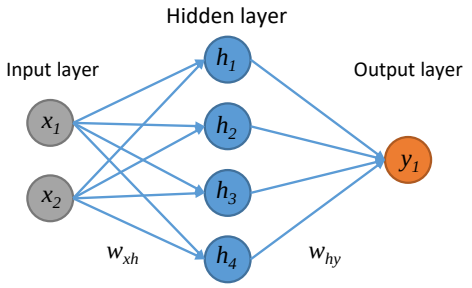


Figure 3.3: A fully connected neural network with one hidden layer

The input layer consists of the input vector $\mathbf{x} = \{x_1, x_2, ..., x_K\}$. The hidden layer consist of a collection of N neurons $\mathbf{h} = \{h_1, h_2, ..., h_N\}$. Similarly, the output layer consist of neurons that represent each element of the output vector $\mathbf{y} = \{y_1, y_2, ..., y_M\}$. Figure 3.3 illustrates a neural network with two units in the input layer, four units in the hidden layer, and a single unit as the output layer. In other words, $K = 2, N = 4$, and $M = 1$.

Every element in the input layer is connected to every element in the hidden layer, and so on for every two sequential layers. This constitutes a fully-connected ANN. Each of this connection is assigned a weight w_{ki}, where k is the unit index on

the lower layer, and i is the index of the higher[1] layer. All the weights in an ANN make up the weight parameters of the model $\mathbf{W} = \left\{\mathbf{w}_{12}, \mathbf{w}_{23}, ..., \mathbf{w}_{[L-2][L-1]}\right\}$ where L is the number of layers of the model. In the illustrated example on Figure 3.3, $L = 3$, layer 1 is indexed as x, 2 as h, and 3 as y

The output h_n of an arbitrary hidden unit u_n is

$$h_n = f(u_n) = f(\sum_{k=1}^{K} w_{xh,k} x_k) \tag{3.2}$$

where K is the number of unit on the previous layer, x_k k-th element of the input vector, and $w_{xh,k}$ is the weight of the connection from the k-th unit to the current unit. Often an additional value is added to the input vector with a corresponding additional weight, acting as a bias.

Similarly, the output y_m of an arbitrary output unit u'_m is

$$y_m = f(u'_m) = f(\sum_{n=1}^{N} w_{hy,n} h_n) \tag{3.3}$$

f is commonly referred to as the *activation function*. An activation function is a pre-defined function that helps provides a smooth transition according to the change of input values. An example of such function is the hyperbolic tangent, which squeezes the input into $\{-1, 1\}$, or the sigmoid function, which squeezes the input into $\{0, 1\}$.

The main task in training an ANN is to optimize the weight parameters $\mathbf{W}$ to minimize the error between the prediction $\mathbf{y}$ and the correct label $\mathbf{t}$. Commonly, a loss function E, e.g. the mean squared error, is defined to formally measure the performance of the network. An algorithm is then used to train the model by minimizing the loss function E in respect to the weight parameters $\mathbf{W}$ of the model. One of the most popular algorithms to train a ANN is the Stochastic Gradient Descent (SGD) algorithm with backpropagation.

In SGD, we perform iterative update of the weight parameters $\mathbf{W}$ in the direction of the gradient of the loss function until a local minimum is reached. At each iteration, a single data point is randomly chosen as a reference in computing the loss function's gradient.

[1]In this thesis, we refer to the input layer as the lowest layer, and output layer as the highest.

To update the weights in each layer, we need to compute the gradient of E with respect to each layer's set of weights. In the example in Figure 3.3, we need to compute $\frac{\partial E}{\partial w_{xh}}$ and $\frac{\partial E}{\partial w_{hy}}$. By employing the chain rule, we can calculate this recursively starting from the highest layer.

From the chain rule,

$$\frac{\partial E}{\partial w_{hy}} = \frac{\partial E}{\partial y_m} \cdot \frac{\partial y_m}{\partial u'_m} \cdot \frac{\partial u'_m}{\partial w_{hy}}, \tag{3.4}$$

similarly,

$$\frac{\partial E}{\partial w_{xh}} = \sum_{m=1}^{M} \left(\frac{\partial E}{\partial y_m} \cdot \frac{\partial y_m}{\partial u'_m} \cdot \frac{\partial u'_m}{\partial h_n} \right) \cdot \frac{\partial h_n}{\partial u_n} \cdot \frac{\partial u_n}{\partial w_{xh}}. \tag{3.5}$$

In general, the update of the weights takes form as

$$w_{l-1,l}^{new} = w_{l-1,l}^{old} - \eta \cdot \frac{\partial E}{\partial w_{l-1,l}} \tag{3.6}$$

where η is the learning rate of the model, determining the size of the step, or parameter change, we take in the direction of the gradient. As this process is done backwards, from the highest layer to the lowest, it is known as *backpropagation*.

Chapter 4

Recognizing Affective States

As discussed in Chapter 1, emotion recognition is fundamental in an affective HCI. Without the ability to track user's emotional state, incorporation of emotion could instead hurt the user interaction experience and result in erroneous actions of the system. In this chapter, we present an experiment in recognizing affective states as the first step in studying social-affective aspects in human communication.

Early works on emotion recognition mainly focus on speech features, i.e. acoustic information [30]. Over time, in the effort to extend these approaches, researchers starts to work on emotional expression recognition [47]. Knowing that these communication clues convey different information, potentially complimentary, the effort to combine multiple modalities into a single emotion recognizer appeared soon thereafter [48].

Combination of multiple modalities is beneficial for emotion recognition. In humans, emotion is often expressed through multiple clues, for example, speech intonation, volume, facial expression, as well as word content. Observation of a single modality in isolation will provide incomplete clue about emotion that occurs.

In the following experiments, we train unimodal classifiers and fuse them into a multimodal one by combining their predictions. We consider acoustic, semantic, and visual information. By including features of different modalities, we hope to be able to observe the fuller picture of emotion occurrences. This constitutes the first multimodal emotion recognition effort in Indonesian.

4.1. Proposed Approach: Multimodal Emotion Recognition

We attempt to train a high-accuracy emotion recognizer by performing a feature combination of different modalities. Figure 4.1 visualizes the combination scheme proposed in this work.

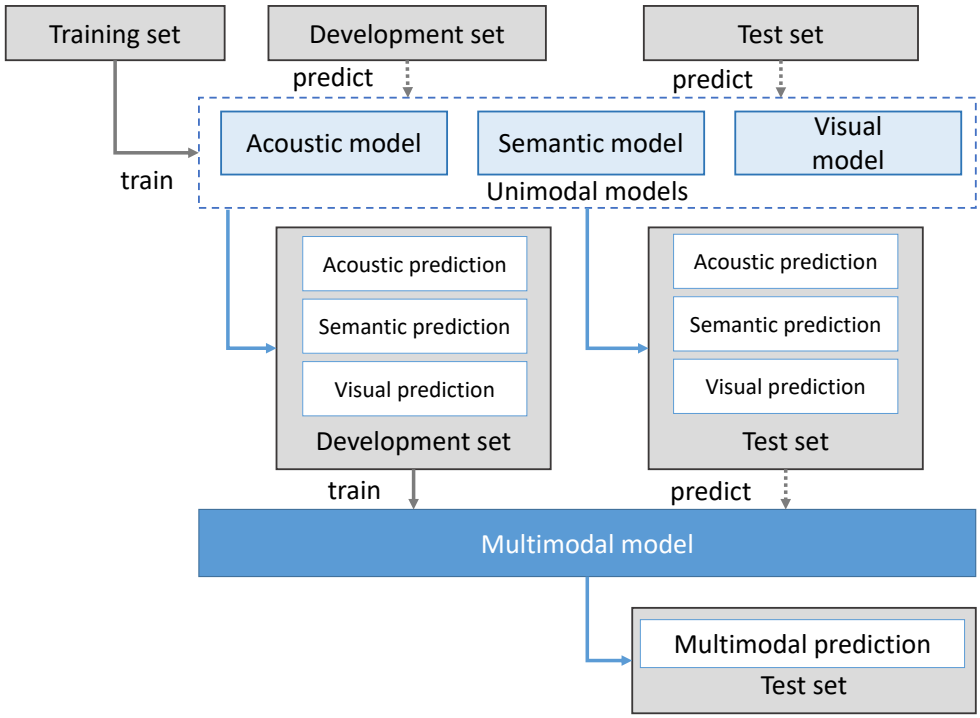


Figure 4.1: Combination scheme of multiple modalities

The model construction starts with the training of several classifiers using various unimodal features. We then utilize these unimodal models to make prediction on subsets of the corpus. We utilize their predictions, in the form of probability distributions, as classification features for the multimodal model.

The use of semantic or lexical information for emotion recognition commonly requires a large, often knowledge-based, language-specific database, i.e. word2vec [38] or WordNet Affect [39]. The construction of such model can be problematic for many languages where resources are scarce, such as Indonesian. In this work, in addition to the acoustic and visual features, we utilize semantic information

extracted purely from the available emotion corpus, without the need of additional data. We exploit the TF-IDF score to weight the words to form the semantic representation.

In this experiment, we perform feature combination at decision level as described above, instead of at feature level by concatenating the features, due to several reasons. Feature-level fusion increases the number of features during training, even though the amount of data remains the same. Furthermore, any modification of features will lead to retraining of the entire model. On the other hand, the decision level fusion treats the modalities in a more modular way, allowing modification without the need of full model retraining.

4.2. Experimental Set Up

In this experiment, we utlize the TV talk show corpus described in section 2.2 in English and Indonesian to train the emotion recognizer in both languages, separately. We use sentence-level segments to avoid segments to allow recognition at turn level in a conversational data.

4.2.1 Label

We simplify the emotion recognition problem by discretizing the affective dimensions values into three classes: positive, neutral, and negative. The reason is twofold: 1) To capture the global value of the utterance, 2) to avoid class sparsity in the data, as the amount is limited. This simplification is a trade-off decision between recognition granularity and model complexity. We believe that incremental steps in tackling a problem will result in a better recognizer in the long run, thus we decide on the less complex problem at this stage of the research.

The emotion label of each segment is decided by majority voting. We first discretize the annotation into positive, neutral, and negative, and perform majority voting afterwards. Segments with three different votes are excluded to avoid potentially ambiguous emotion occurrences. Table 4.1 summarizes the number of final segments for each emotion recognition tasks. We divide the total with 80:10:10 ratio into training, development, and test sets.

Table 4.1: Total number of segments for emotion recognition tasks

Language	Valence	Arousal
English	1158	1105
Indonesian	805	820

4.2.2 Features

We extract several feature sets to represent three modalities of communication: acoustic, semantic, and visual.

Acoustic. First, we extract global features of each utterance using the openSMILE toolkit [49]. Three different acoustic feature sets are selected: INTERSPEECH 2009 baseline features (IS09) and extended Geneva Minimalistic Acoustic Parameter Set (eGemap)s feature sets. The IS09 feature set is described in Table 4.2. This feature set is widely used in emotion recognition research, thus providing comparability to extensive related works.

Table 4.2: Baseline feature of INTERSPEECH 2009 emotion challenge

LLD ($16 \cdot 2$)	Functionals (12)
($\triangle$) ZCR	mean
($\triangle$) RMS Energy	standard deviation
($\triangle$) F0	kurtosis, skewness
($\triangle$) HNR	extremes: value, rel. position, range
($\triangle$) MFCC 1-12	linear regression: offset, slope, MSE

On the other hand, the eGemaps feature set is proposed as reduced acoustic feature set, containing only knowledge-based selected features that are 1) highly potential in indexing affective signals, and 2) proven to be effective in previous studies [50]. This feature set includes parameters related to frequency (pitch, jitter, formant), energy (shimmer, loudness, HNR), spectral (alpha ratio, Hammarberg index, MFCC 1-4, spectral scope, format relative energy and bandwidth, spectral flux). Reduction of the total numbers of features is mostly contributed by the much fewer number of functionals, mostly only involving arithmetic mean and coefficient variation, with additional ones for selected features.

Semantic. Secondly, to extract the semantic features of the utterances, we

compute the TFIDF weighted vector of its transcribed speech. The TFIDF weight of each term t in sentence s is computed as:

$$\text{TFIDF}(t, T) = F_{t,T} \log \frac{|T|}{DF_t}, \tag{4.1}$$

where $F_{t,T}$ is defined as term frequency of term t in a sentence T, and DF_t as total number of sentences that contains the term t, calculated over the full database. Thus, the vector for each sentence s is the size of the corpus term vocabulary, with each term weighted according to Equation 4.1.

Visual. Lastly, we extract facial features of each frame with the OpenFace toolkit [51]. We extract 2D position of the facial landmarks, as well as AU intensities, and treat them as two separate feature sets. We exclude frames with low face detection confidence. To obtain the segment level features, we calculate 4 statistics of the frame level features: mean, standard deviation, minimum, and maximum value.

Table 4.3 summarizes the total number of features in each feature set.

Table 4.3: Number of features in each feature set

Modality	Feature set	Number of features
Acoustic	gemaps	62
	IS09	384
Visual	AU	72
	landmarks	336
Semantic	TFIDF	2801

4.2.3 Model

We train an SVM classifier for each feature set and all combinations of the three modalities. We use the libSVM library [52] in our experiments. Prior to training, we scale the features into $\{0, 1\}$ range to avoid overpowering of features with high values range. Furthermore, we perform parameter optimization with grid search to optimize the SVM models. These steps are recommended in [53] and has been shown to be effective in SVM classification experiments. For all models, unimodal or multimodal, we compare their performance by using the recognition accuracy on the test set. The 3 classes in the data amounts to a chance rate of 33.33%

4.3. Result

4.3.1 Unimodal Emotion Recognition

Table 4.4 presents the performance of unimodal emotion recognition for 4 emotion recognition tasks: Arousal and valence level predictions in English and Indonesian. The unimodal recognition result shows that in 3 of the 4 tasks, with the exception of valence in English, acoustic signals appear to be the most discriminative classification feature, signaling that the richest emotion clue in the conversational data is carried through speech. Interestingly, regardless of the much bigger number of features, the IS09 set outperforms the eGemaps set. This suggest that some features that are present in IS09 but absent in eGamps are helpful in recognizing emotion.

Table 4.4: Unimodal emotion recognition accuracy on test set. Highest number on each task is boldfaced

Language	Modality	Feature	Arousal	Valence
English	Acoustic	eGemaps	0.738	0.344
		IS09	**0.765**	0.379
	Visual	AU	0.693	0.387
		landmarks	0.702	0.189
	Semantic	TFIDF	0.693	**0.551**
Indonesian	Acoustic	eGemaps	0.902	0.839
		IS09	**0.926**	**0.864**
	Visual	AU	0.829	0.814
		landmarks	0.829	0.814
	Semantic	TFIDF	0.829	0.530

On the other hand, the positivity or negativity of emotional state in the English conversational data seem to be conveyed the strongest in the content of speech, i.e. the words. This is the opposite of that of Indonesian, where valence level prediction with TFIDF features is significantly lower than the other modalities.

Lastly, the recognition performance using visual features is slightly underperforming that of acoustic in all of the tasks. Overall, the AU features seem to be

performing better than landmarks, given the identical results in Indonesian, the slight gap in arousal in English, and the wide gap in valence in English.

4.3.2 Multimodal Emotion Recognition

Table 4.5 summarized the result of multimodal emotion recognition. We experimented with all possible combinations of the three modalities. The numbers suggest that the determining factor in the result of feature combination is the highest performing feature among the unimodal ones.

Table 4.5: Multimodal emotion recognition accuracy on test set. Highest number on each task is boldfaced. *: higher than unimodal best

Language	Feature			Arousal	Valence
	Semantic	Acoustic	Visual		
English	TFIDF	eGemaps	AU	**0.756**	0.637*
			landmarks	**0.756**	0.637*
		IS09	AU	0.729	0.637*
			landmarks	0.738	0.637*
Indonesian	TFIDF	eGemaps	AU	0.829	0.839
			landmarks	0.829	0.839
		IS09	AU	**0.926**	**0.938***
			landmarks	**0.926**	0.925*

For example, in valence in English, all other other features is overpowered by TFIDF, which was the highest performing unimodal feature of the task. In reverse, the decision of which visual features to use in the combination does not affect the performance of the end multimodal model, most probably due to its suboptimal performance in comparison to other unimodal features. In both languages, in the arousal tasks the best combination of the modalities yields comparable number as that of the unimodal. On the other hand, in the valence task, the combination is able to improve the recognition rate significantly.

To understand this difference better, we analyze the result of the unimodal and multimodal recognition. For each pair of the best performing multimodal sets (i.e. ABC becomes pairs AB, AC, and BC), we compute the correlation coefficient of the probability distributions to see the similarity of the predictions. We find

that on the arousal task in both languages, the unimodal prediction outputs highly correlated predictions, making most mistakes on the same examples. In consequence, they do not provide complimentary information to one another when combined. In contrast, for the valence task in both languages, we observe variance and different trends of prediction. This means the modalities contain different, presumably complimentary, information that is useful in better modeling the problem and classifying the data.

4.4. Summary

In this chapter, we presented emotion recognition experiments on spontaneous social-affective data collected from TV talk shows. We consider emotion dimension level estimation, discretized into three classes: positive, neutral, and negative. We experimented with a total of five feature sets in three modalities: acoustic, semantic, and visual.

We combine the predictions of unimodal features by combining their probability estimates as new features used in the multimodal experiment. When compared to the unimodal result, in the multimodal feature combination, we attain comparable performance for the arousal classification task and a significant improvement for the valence classification task.

Improvement of the performance can be achieved by increasing the amount of data, or by considering other feature sets that better represent the emotional clues in human communication. In the future, we hope to examine a more fine-grained emotion dimension level estimation, e.g. by considering a finer level of separation (5 classes, 7 classes, etc) or by performing regression task to predict the real value. We also hope to integrate the emotion recognizer into a working system to allow real-time emotion recognition in HCI.

Chapter 5

Analysis and Automatic Prediction of Social-Affective Events

In addition to the more traditional works on emotion recognition and simulation, there has recently been an increasing interest in *emotion elicitation*, or *emotional triggers*, in which a computer system attempts to trigger a certain emotion from the user through the interaction. Previous work by Hasegawa et. al. [12] is reported to be able to elicit basic emotions properly by training individual models that "translates" user's input into the eliciting response.

However, the model have not yet observe the relationship between an utterance by the speaker, which acts as stimuli evaluated during appraisal, and the resulting emotion by the end of appraisal. This is illustrated in Figure 1.4. As discussed in Section 1.2.2, observation and incorporation of the underlying process that causes change of emotion can provide useful information in eliciting positive emotion in any interaction.

In this chapter, we extend previous user-centered study by analyzing and predicting emotional responses in a social-affective conversation by examining the relationship between emotional triggers and responses. We observe how emotion fluctuates and its connection to actions taken in discourse. The analysis is intended to provide the knowledge to perform prediction of emotional triggers and responses. On the other hand, the prediction is aimed to accommodate two of the most important social-affective abilities: (1) to decide on an emotion triggering

action, and (2) to be able give an appropriate emotional response.

5.1. Proposed Approach: Analyzing and Predicting Social-Affective Events

5.1.1 Triturns and Social-Affective Events

To properly analyze the fluctuation of emotion in a conversation, it is necessary to ensure that the observed sequences of conversation are in response to each other, as natural conversation can be unordered and disconnected from one turn to the next. To do this, we group consecutive sequences of conversation into a unit called a triturn [54]. Three consecutive sequences of speech in a conversation is considered a triturn when the second sequence is in response to the first, and the third is in response to the second.

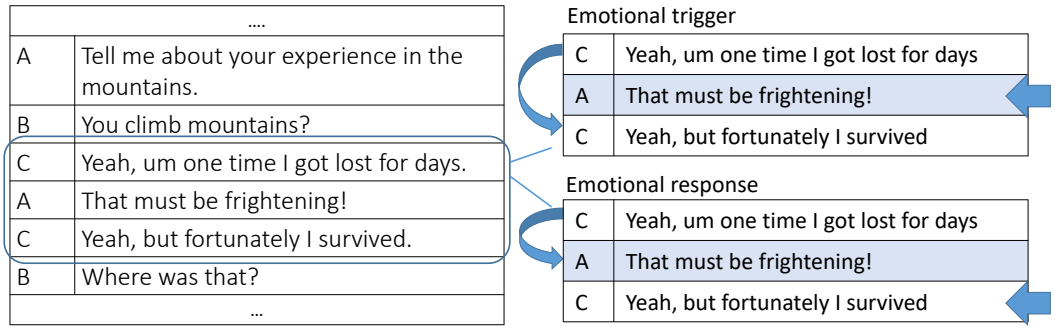


Figure 5.1: Triturn and social-affective events in a conversation

On each triturn, we refer to the change of emotion from the first turn as *emotional response*. As the responses can occur simultaneously, we examine valence and arousal separately. We categorize changes of each emotion dimension into three events: rise, drop, or constant. On the other hand, we consider the second turn to be the *emotional trigger* of the emotional response. We consider the set of 17 dialogue acts in Table 2.1 described in Section 2.2.3 as emotional triggers.

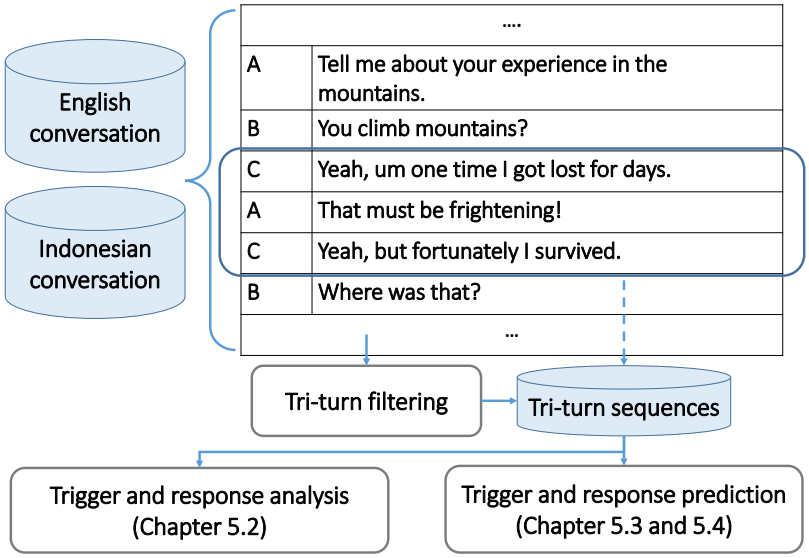


Figure 5.2: Overview

5.1.2 Analysis and Prediction of Social-Affective Events

Figure 5.2 presents the overview of this task. We collected tri-turns from a conversational corpus to perform social-affective events analysis and prediction. In the analysis, we compute statistical connection between emotional responses and triggers. We investigate: 1) whether there are dialogue actions that heavily characterizes any emotional response, and 2) whether the emotion of the trigger is correlated with the emotional response.

In emotional triggers prediction, given the first and last sequence of a triturn as an emotional response, we try to automatically predict what action takes place as the trigger. This answers the following question: *Which dialogue act triggered this emotional response?* On the other hand, in emotional response prediction, given two consecutive turns, we try to predict the emotional response that will occur next. This answers the following question: *Will the observed emotion dimension rise, drop, or stay constant?* We utilize the acoustic information and the dialogue act of the respective two sequences as classification features to capture emotion as well as dialogue-related information of the triturn.

5.2. Data Analysis

We perform analysis on the pre-processed data by correlating dialogue acts and the occuring emotional responses. This section presents the result of the analysis and as a potential knowledge for emotional awareness in HCI.

5.2.1 Emotional Triggers

To examine the connection between dialogue acts and changes of emotion occurring in conversation, we adapt the Term Frequency-Inverse Document Frequency (TF-IDF) formula as written in Equation (4.1) to measure the importance of each dialogue act in triggering a certain emotion event. The adapted formula for dialogue acts is written as (5.1)

$$tfidf(d, t, T) = f(d, t) \times \log \frac{\{t \in T\}}{1 + \{t \in T : d \in t\}}, \tag{5.1}$$

where d is the dialogue act, t is the emotion event, and T is the collection of events. $f(d, t)$ denotes the raw frequency of d in t. This score is calculated for each dialogue act on each emotion event. This score can inform us if a particular dialogue act is characteristic of a particular emotion event.

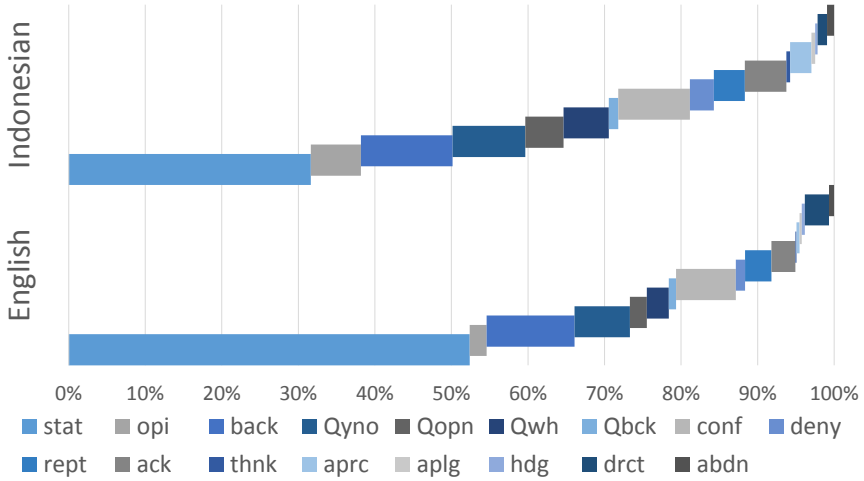
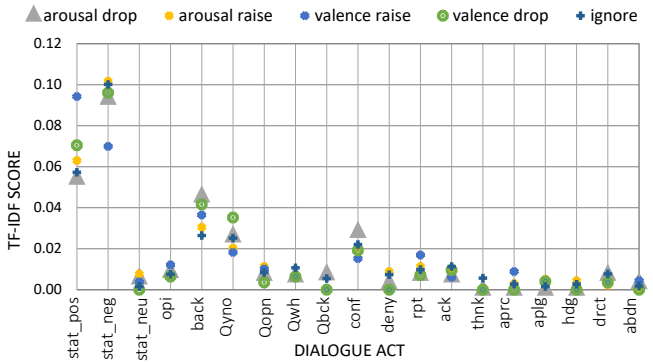


Figure 5.3: Dialogue act frequency on triggers of all emotion events

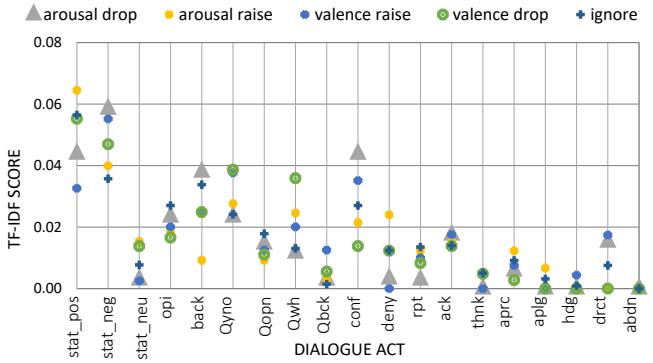
First, we take a look at all the emotional triggers, regardless of the emotion event they elicit, and see the frequency of dialogue acts. Figure 5.3 visualize the

percentage of dialogue act frequencies for both languages. Statement has the highest frequency for both languages.

Second, we look into details for each emotion event. For each event, we calculate the TF-IDF score of each dialogue act using Equation (5.1). As the frequency of `stat` is high for all types of change, we separate statements according to their emotion (i.e. positive, negative, and neutral), and calculate their scores accordingly.



(a) English



(b) Indonesian

Figure 5.4: Dialogue acts scores for all emotion events

Figure 5.4 visualizes the TF-IDF scores of the dialogue acts, where a point in the graph corresponds to the TF-IDF score for a dialogue act on an emotion event. A large number of overlapping dots for a dialogue act means its TF-IDF score is similar for all emotion events. When this is observed, we can conclude that the dialogue act does not characterize a particular emotion event.

Comparison of Figures 5.4(a) and 5.4(b) tells us that the emotion events have

different characteristics between the two languages when viewed from the action that triggers them. In English, the dots overlap one another with only slight differences, signaling that dialogue acts weakly characterize emotion events. On the other hand, in Indonesian, fewer overlapping dots are observed. This means, in English, emotion events are not attributed to the act taken in discourse, as opposed to Indonesian.

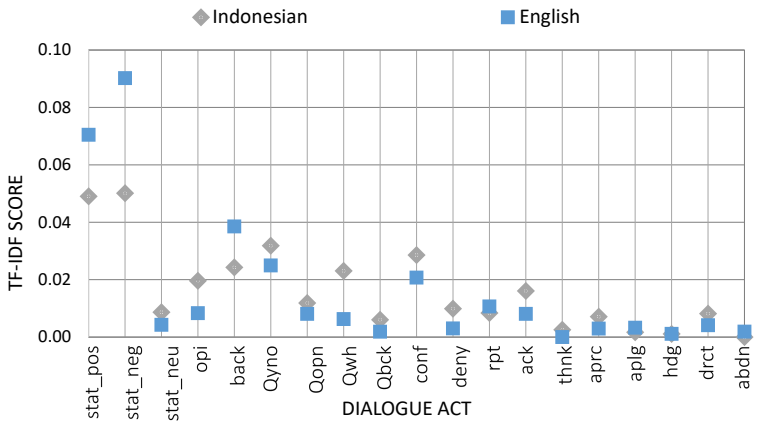


Figure 5.5: Average scores of dialogue acts in English and Indonesian

Figure 5.5 shows the TF-IDF score of the dialogue acts in English and Indonesian averaged over all emotion events. From this figure, we can identify significant dialogue acts in triggering emotion events. In English, the top five dialogue acts are `stat`, `back`, `Qyno`, `conf`, and `rpt`. In Indonesian, the top five dialogue acts are `stat`, `Qyno`, `back`, `conf`, and `Qwh`.

This finding suggests that showing interest in the conversation through questions and backchannels is a way to emotionally engage with the counterpart. Furthermore, providing new information in conversation can also elicit an emotional response from the counterpart. Table 5.1 includes an example of conversation taken from the corpus that demonstrates this finding. Looking at the level of valence and arousal, we can notice an increase for both the host and the guest.

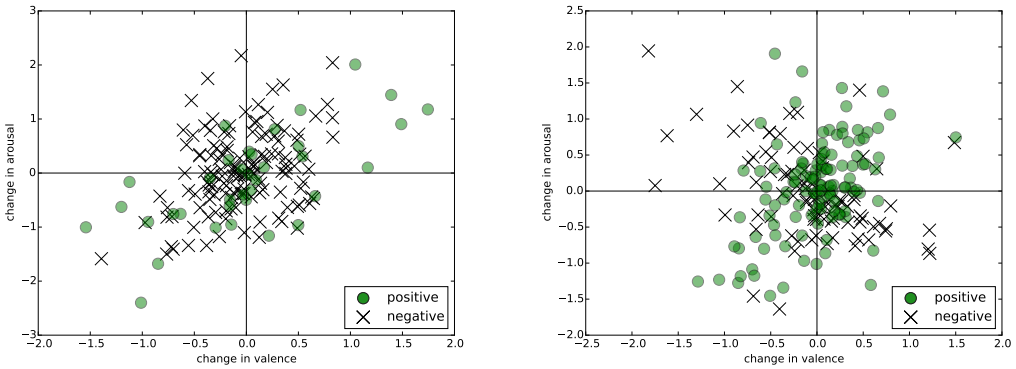
5.2.2 Emotional Response

Statements make up a large part of human conversation. Therefore, we attempt to take a closer look at statements as emotional triggers by grouping them ac-

Table 5.1: Example of conversation

Speaker	Transcription	act	aro	val
Guest	Well I still have a lot of clothes in my closet I really shouldn't have	stat	0	-1
Host	Yeah	back	-1	0
Guest	But—yeah	conf	1	1
Host	Why?	Qwh	0	1
Guest	Just [inaudible] I want to say just in case but I don't think so 'cause I really think I got it conquered this time	opi	2	2

cording to the emotion of the statement and plot them according to the emotional response. This distribution is shown in Figure 5.6.



(a) English (b) Indonesian

Figure 5.6: Emotion of statements with respect to the emotional response it triggers

Different tendencies can be observed from Figure 5.6. In English, the emotion triggering statements have an even distribution of positive and negative emotion. Statements with positive emotion spread to the upper right and lower left quadrants, meaning they cause an increase or decrease of both valence and arousal at the same time. On the other hand, statements with negative emotion give opposing effects to valence and arousal.

Unlike in English, in Indonesian, negative emotion dominated the emotion triggering statements. However, the few statements that have positive emotion

seem to have a bigger effect compared to the negative ones.

5.3. Experimental Set Up

In this experiment, we utilize the TV Talk Shows corpus discussed in Section 2.2 as it contains spontaneous social-affective interaction between humans. To form triturns, from the sentence-level segments we first concatenate consecutive sentences of the same speaker until a change of speaker occurs to form dialogue-turn segments. Afterwards, we form triturn unit from dialogue turns with A-B-A pattern, as illustrated in Figure 5.1. In English, 626 tri-turns are collected and in Indonesian 567. We parition the data with 80:20 ratio to serve as training and test set.

To gather information from the speech, we extract acoustic features of each turn as defined in the INTERSPEECH 2009 emotion chalenge [55] using the openSMILE feature extractor [49]. As previously discussed, this feature set includes the most common yet promising feature types and functionals covering prosodic, spectral, and voice quality features.

On each triturn, we stack the features of the two corresponding turn sequences to gather information of the context, i.e. the first two turns for emotional response prediction, the first and last turns for emotional triggers prediction. To balance the number of instances and features, we perform correlation-based feature extraction [56] and linear discriminant analysis of our feature set. After reducing the dimension, we train a deep neural network classifier using Theano and the PDNN toolkit [57].

5.4. Result

Figure 5.7 summarizes the result of the social-affective events prediction experiments. the chance rate of the trigger prediction task is 5.88%, given the 17 classes of dialogue acts. Our models achieve the accuracy of 53.97% for English, and 31.58% for Indonesian.

The lower performance for Indonesian is suspected to be the implication of the dialogue act analysis in Section 5.2.1. In Indonesian, different emotion events can be triggered by different actions. Learning of these more complex patterns is likely

to require a larger amount of data than currently present. On the other hand, for English, triggers of a certain emotion event aren't strongly characterized by the dialogue act. In other words, the occuring emotion event only weakly affects the dialogue act, while other features such as acoustic features and dialogue act helps make the prediction.

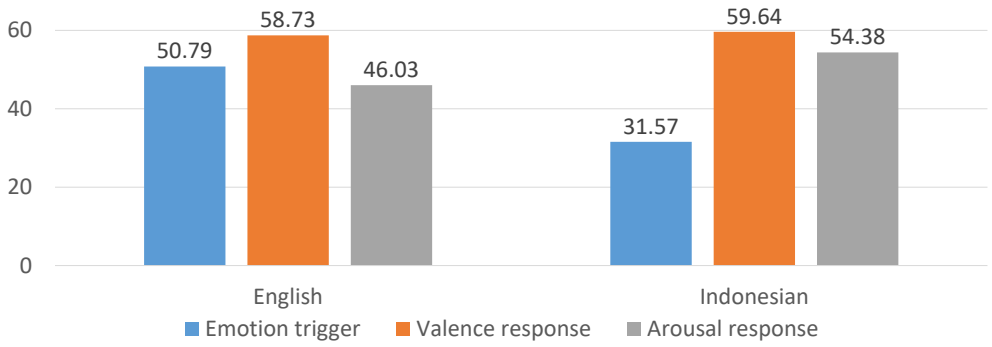


Figure 5.7: Accuracy of social-affective events prediction

With three classes of events, the chance rate on the responses prediction task is 33.33%. As shown in Figure 5.7, compared to the other tasks, the prediction regarding valence has the highest performance. This suggests a stronger pattern between change of valence, action in discourse, and speech characteristics, compared to that of arousal. The same occurrence is seen in Indonesian data

The suboptimal performance is likely due to the limited amount of data used in this study. Inherently, there are numerous factors that leads to a change of emotion in a conversation. This means that to properly recognize patterns for such events, a large number of features are required. It is likely that if we could prepare more data in the future, the accuracy will increase significantly.

5.5. Summary

This chapter presented a study on social-affective events, including analysis and prediction of emotional triggers and responses. We examine natural conversations in English and Indonesian and analyze the emotion events that occurred within. Each language shows different tendencies and characteristics of the social-affective events, suggesting that they are influenced by language and culture. We look forward to affirm this finding on a larger amount of data.

The experiment on automatic prediction offers an approach in equipping conversational agents and dialogue systems with social-affective capabilities. In a later stage of the study, we hope to include more modalities of interaction in observing the dynamics of emotion, such as textual and visual features. A fuller picture of the occurring events is highly potential in increasing the performance of the prediction models.

Chapter 6

Eliciting Positive Emotional Impact

As discussed in Chapter 1, previous works in emotion elicitation have not yet addressed the appraisal process that gives rise to the change of emotion. Previous work by Hasegawa et al. [12] proposed statistical response generators using a machine translation technique to elicit various emotion. Each response generator has an emotion target; the response "translated," i.e. generated, from the model that targets "sadness" is expected to elicit sadness. On the other hand, Skowron et al. [14] approached emotion elicitation by using various affective personalities in the system. In other words, these approaches have only focused on the intent of elicitation itself. Furthermore, these studies are limited to textual conversation, leaving a gap between the result and natural human communication.

In this chapter, we attempt to elicit a positive emotional change in HCI by exploiting examples of appraisal in human spoken dialogue. We collected dialogue sequences containing emotional triggers and responses to serve as examples in a dialogue system. Subsequently, we augment the traditional response selection criterion with emotional parameters: 1) user's emotional state, and 2) expected[1] future emotional impact of the candidate responses. These parameters represent parts of the information that humans use in social-affective interactions.

The proposed system improves upon the existing studies by harnessing information of human appraisal in eliciting user's emotion. We eliminate the need

[1]Within the scope of the proposed method, we use the word *expected* for its literal meaning, as opposed to its usage as a term in probability theory.

of multiple models and the definition of emotion targets by aiming for a general positive affective interaction. The use of data-driven approach rids the need of complex modeling and manual labor. Text-based human subjective evaluation with crowdsourcing shows that in comparison to the traditional response selection method, the proposed one elicits an overall more positive emotional impact, and yields higher coherence as well as emotional connection.

In essence, we aim to provide an emotionally positive interaction between the user and the system. Striving for natural human communication, we utilize human spoken interaction to build the system.

6.1. Example-based Chat Oriented Dialogue System

In this section, we explain the traditional approach in building a chat-oriented dialogue system and how it addresses the difficulties commonly faced in dialogue system construction.

Example-Based Dialogue Management (EBDM) is a data-driven approach of dialogue modeling that uses a semantically indexed corpus of *query-response*[2] pair examples instead of handcrafted rules or probabilistic models [58]. At a given time, the system will return a response of the best example according to a semantic constraint between the query and example query. This circumvents the challenge of domain identification and switching—a task particularly hard in chat-oriented systems where no specific goal or topic is predefined beforehand. With the increasing amount of available conversational data, EBDM offers a straightforward and effective approach for deploying a dialogue system in any domain.

Lasguido et al. have previously examined the utilization of cosine similarity for response retrieval in an example-based dialogue system [54]. In their approach, the similarity is computed between TF-IDF weighted sentence vectors, as described in Section 4.2, of the query and the examples.

Cosine similarity between two sentence vectors S_q and S_e is computed as:

[2]In the context of dialogue system, we will use term *query* to refer to user's input, and *response* to refer to system's output

$$\cos_{sim}(S_q, S_e) = \frac{S_q \cdot S_e}{\|S_q\| \, \|S_e\|}, \tag{6.1}$$

Given a query, this cosine similarity is computed over all example queries in the database and treated as the example pair scores. The response of the example pair with the highest score is then returned to the user as the system's response. This is one of the main approaches in chat-oriented dialogue systems. This process is illustrated in Figure 6.1. The traditional EBDM response selection method serves as the baseline of this study.

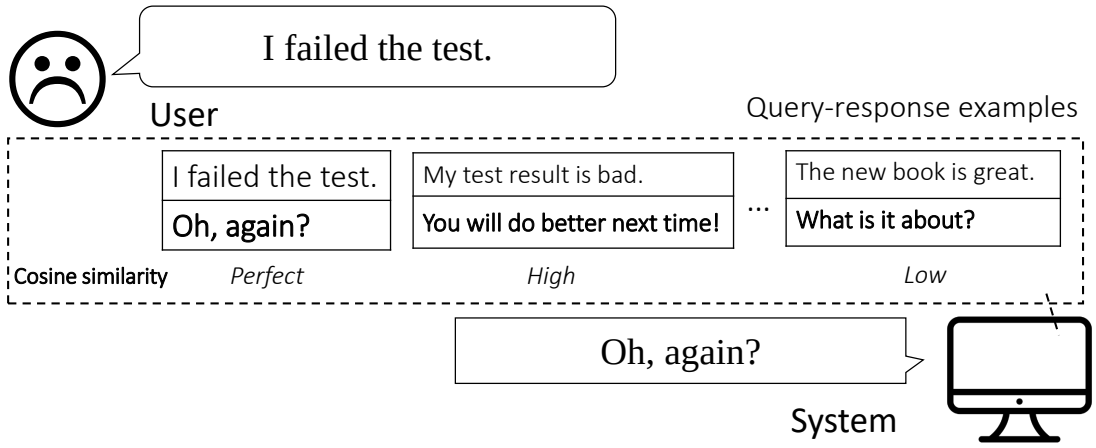


Figure 6.1: Response selection on EBDM

6.2. Proposed Approach: Eliciting Positive Emotional Impact in Dialogue Response Selection

6.2.1 Triturns as Example Database

To allow the consideration of the emotional parameters aforementioned, we make use of *tri-turn* units in the selection process in place of the *query-response* pairs in the traditional EBDM approach. As discussed in Section 5.1.1, a tri-turn consists of three consecutive dialogue turns that are in response to each other.

This format has been previously utilized in collecting *query-response* examples from a text-based conversational data to ensure that an example is dyadic [54].

In this work, we instead exploit the tri-turn format to observe emotional triggers and responses in a conversation. The triturn format allows the observation of the future response, i.e. the user response to the system response, in the examples. We believe that expected future impact is an aspect that should not be overlooked. This common knowledge is prevalent in humans and strongly guides how we communicate with other people—for example, to refrain from provocative responses and to seek pleasing ones.

Within this chapter, the first, second, and third turns in a tri-turn are referred to as *query*, *response*, and *future*, respectively. The change of emotion observed from *query* to *future* can be regarded as the impact of *response*.

6.2.2 Response Selection Method

We are fully aware that the *future* of each tri-turn is not a definite prediction of user response in real interaction—we are not considering it as one. Within the dialogue system, each tri-turn is essentially an example of a human's response in a conversation with certain semantic and emotional context. Our hope is that, in real interaction, given a similar semantic and emotional context as an example tri-turn's *query*, when the system outputs the *response*, the user will experience an emotional change consistent with that of the *future*.

Thus, in addition to semantic constraint as described in Section 6.1, we formulate two types of emotional constraints: (1) emotion similarity between the query and the example queries, and (2) expected emotional impact of the candidate responses. Figure 6.2 shows how considering expected emotional impact could potentially elicit a more positive effect in the real interaction.

To measure emotion similarity, we compute the Pearson's correlation coefficient of the emotion vector between the query and the example queries. Correlation r_{qe} between two emotion representation vectors for query q and example e of length n is calculated using Equation 6.2.

$$r_{qe} = \frac{\sum_{i=1}^{n}(q_i - \bar{q})(e_i - \bar{e})}{\sqrt{\sum_{i=1}^{n}(q_i - \bar{q})^2}\sqrt{\sum_{i=1}^{n}(e_i - \bar{e})^2}}, \tag{6.2}$$

This similarity measure utilizes real-time valence-arousal values instead of discrete

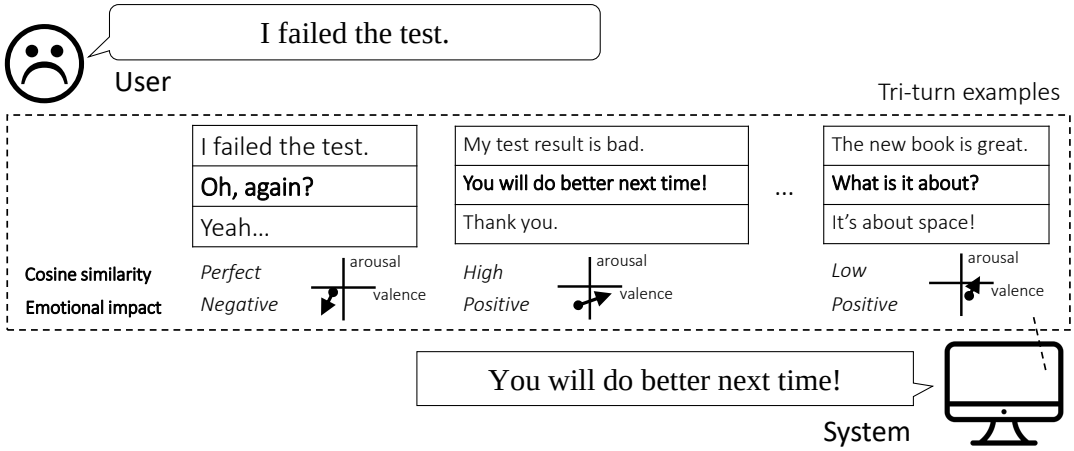


Figure 6.2: Considering expected emotional impact in dialogue response selection

emotion label. In contrast with discrete label, real-time annotation captures emotion fluctuation within an utterance, represented with the values of valence or arousal with a constant time interval, e.g. a value for every second.

As the length of emotion vector depends on the duration of the utterance, prior to emotion similarity calculation, sampling is performed to keep the emotion vector in uniform length of n. For shorter utterances with fewer than n values in the emotion vector, we perform sampling with replacement, i.e. a number can be sampled more than once. The sampling preserves distribution of the values in the original emotion vector. We calculate the emotion similarity score separately for valence and arousal, and then take the average as the final score.

Secondly, we measure the expected emotional impact of the candidate responses. In a tri-turn, emotional impact of a *response* according to the *query* and *future* is computed using Equation 6.3.

$$\text{impact}(response) = \frac{1}{n}\sum_{i=1}^{n} f_i - \frac{1}{n}\sum_{i=1}^{n} q_i, \tag{6.3}$$

where q and f are the emotion vectors of *query* and *future*. In other words, the actual emotion impact observed in an example is the expected emotional impact during the real interaction. For expected emotional impact, we consider only valence as the final score

Figure 6.3 illustrates the steps of response selection of the baseline and proposed systems. We perform the selection in three steps based on the defined

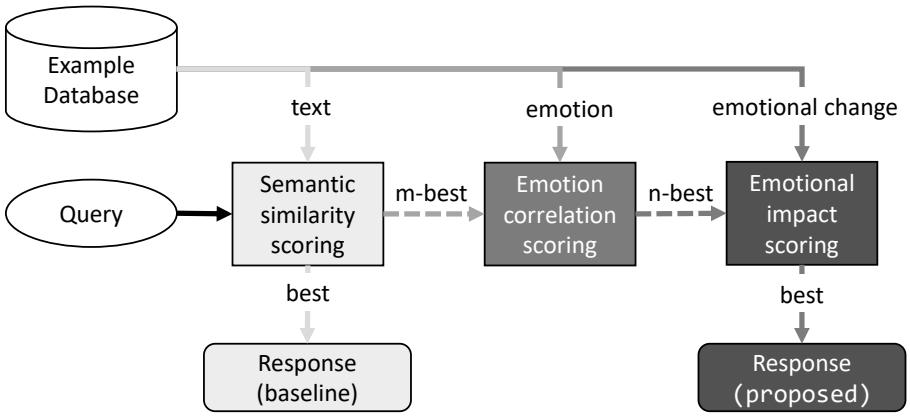


Figure 6.3: Steps of response selection

constraints. For each step, a new score is calculated and re-ranking is performed only with the new score, i.e. no fusion with the previous score is performed.

The baseline system will output the *response* of the tri-turn example with the highest semantic similarity score (Equation 6.1). On the other hand, on the proposed system's response selection, we pass m examples with highest semantic similarity scores to the next step and calculate their emotion similarity scores (Equation 6.2). From n examples with highest emotion similarity scores, we output the *response* of the tri-turn example with the most positive expected emotional impact (Equation 6.3).

The filtering on each step is done to ensure that the semantic and emotional contexts in the candidate examples match the real interaction such that the response yields as similar an impact as possible with the example. As emotion space is much smaller than that of semantic, many of the examples may achieve a high emotion similarity score. Thus, imposing the emotion constraints in the reduced pool will help achieve a more relevant result. Furthermore, this reduces the computation time since the number of examples to be scored will be greatly minimized. When working with big example databases, this property is beneficial in giving a timely response.

It is important to note that this strategy does not translate to selection of the response with the most positive emotion. On the other hand, it is equivalent to selecting the response that has the most potential in eliciting a positive emotional impact, given a semantic and emotional context. Even though there is no explicit dialogue strategy to be followed, we hope that the data reflects the appropriate

situation to show negative emotion to elicit a positive impact in the user, such as relating to one's anger or showing empathy.

6.3. Experimental Set Up

In this study, we consider the SEMAINE Database discussed in Section 2.1. This database is suitable for this particular task as it shows how user behave when they interact with an automated agent, albeit one played by a wizard. Furthermore, the characteristics of the agents has a great potential in conveying salient emotional triggers in responding to the user's query.

We selected 66 sessions from the full corpus based on transcription and emotion annotation availability; 17 Poppy's sessions, 16 Spike, 17 Obadiah, 16 Prudence. For every dialogue turn, we keep the speaker information, time alignment, transcription, and emotion traces.

In the SEMAINE Database, Poppy and Prudence have more positive emotional trends in their sessions, while Spike and Obadiah have rather negative trends. This means, Poppy and Prudence tend to draw the user into the positive-valence region of emotion as opposed to Spike and Obadiah. This resembles a *positive emotional impact*, where the final emotional state is more positive than the initial. Thus, we exclusively use sessions of Poppy and Prudence to construct the example database.

We partition the recording sessions in the corpus into training and test sets. The training set and test set comprise 29 (15 Poppy, 14 Prudence) and 4 (2 Poppy, 2 Prudence) sessions, respectively. We construct the example database exclusively from the training set, containing 1105 tri-turns. We sample the valence and arousal emotion traces of every dialogue turn into 100-length vectors to keep the length uniform.

In this phase of the research, we utilize the transcription and emotion annotation provided from the corpus as information of the tri-turns to isolate the errors of automatic speech and emotion recognition. For the n-best filtering, we chose 10 for the semantic similarity constraint and 3 for the emotion considering the size of the corpus.

6.4. Emotional Impact Analysis

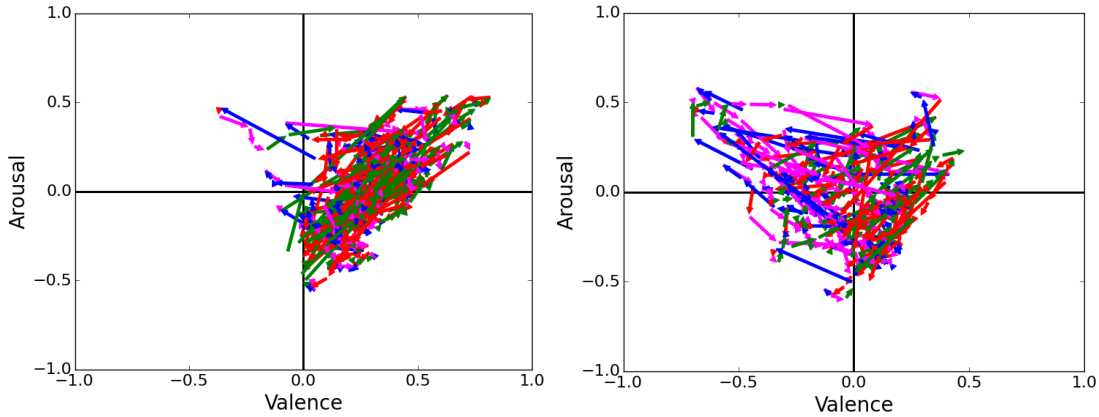
We suspect that the distinct characteristic of each SAL affects the user's emotional state in different ways. To observe the emotional impact of the dialogue turns in the data, we extract tri-turn units from the selected 66 sessions of the corpus. As SEMAINE contains only dyadic interactions, a turn can be assumed as a response to the previous one.

We investigate whether the characteristics of the SAL affect the tendencies of emotion occurrences in a conversation by analyzing the extracted tri-turns. From all the tri-turns extracted from the subset, we compute the emotional impacts and plot them onto the valence-arousal axes, separated by the SAL to show emotion trends of each one. Figure 6.4 presents this information. In the figure, the arrows represent emotional impact, with initial emotion as starting point and and final as ending. The direction of the arrows shows the emotional change that occurs. Up and down directions show the increase and decrease of arousal. Right and left directions show the increase and decrease of valence.

The figure shows different emotional paths taken during the conversation with distinct trends. In Poppy's and Prudence's sessions, most of the emotional occurrences and transitions happen in positive-valence and positive-arousal region with occasional movement to the negative-valence region. On the other hand, in Spike's sessions, movement to the negative-valence positive-arousal region is significantly more often compared to the others. The same phenomenon occurs with negative-valence negative-arousal region in Obadiah's. This tendency is consistent to the characteristic portrayed by each SAL, as our initial intuition suggests.

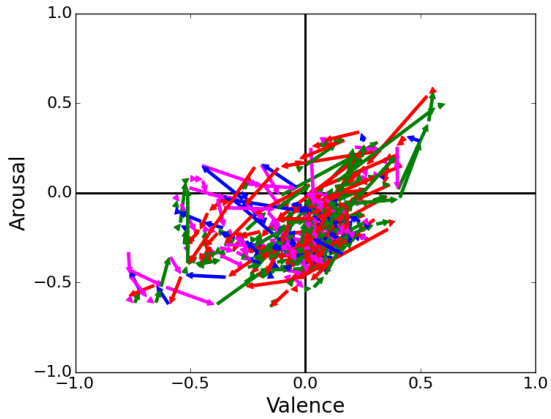
6.5. Human Evaluation: Procedure, Result, and Analysis

We perform an offline subjective evaluation to qualitatively measure perceived differences between the two response selection methods. From the test set, we extract 198 test queries. For each test query is, we generate two responses with the baseline and proposed dialogue systems. To emphasize the difference between the systems, we only make use of test queries that result in unique responses.

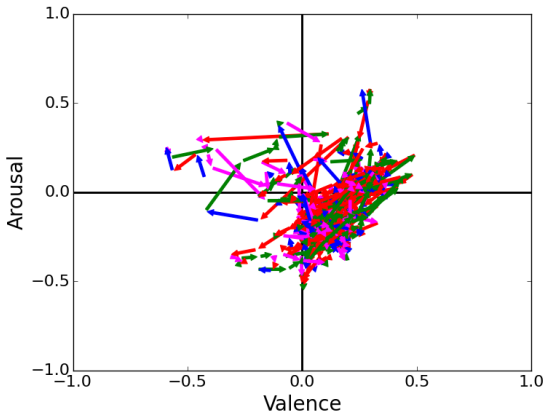


(a) Poppy

(b) Spike



(c) Obadiah



(d) Prudence

Figure 6.4: Emotional changes in SEMAINE sessions separated by SAL character

We further filter the queries based on utterance length and emotion labels to give enough context and variance in the evaluation. In the end, 50 queries are selected.

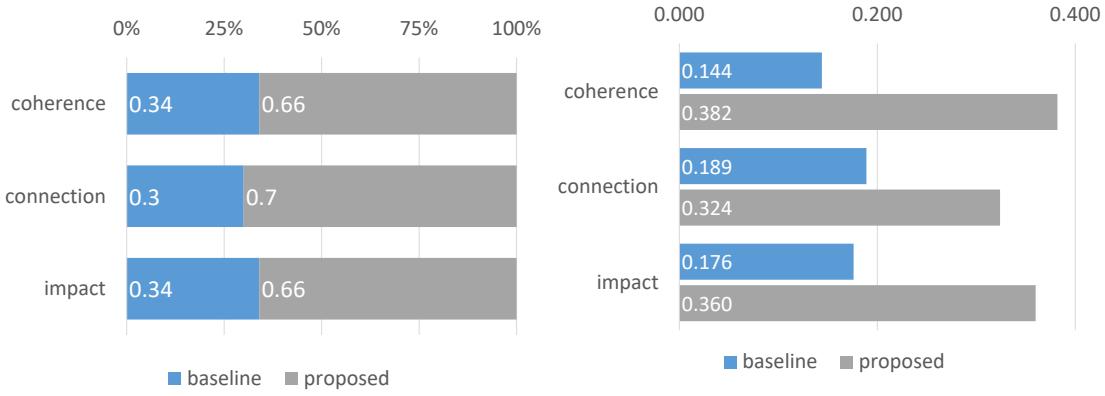
We perform subjective evaluation of the systems with crowdsourcing. We ask the evaluators to compare the systems' responses in respect to the test queries. The query and responses are presented in form of text. For each test query, the responses from the systems are presented as with random ordering, and the evaluators are asked three questions, adapted from [14]:

1. Which response is more coherent? Coherence refers to the logical continuity of the dialogue.

2. Which response has more potential in building emotional connection between the speakers? Emotional connection refers to the potential of continued interaction and relationship development.

3. Which response gives a more positive emotional impact? Emotional impact refers to the potential emotional change the response may cause.

50 judgements are collected per query. Each judgment is weighted with the level of trust of the worker[3]. In this evaluation, we employ workers with high-ranking level of trust. The final judgement of each query for each question is based on the total weight of the overall judgements—the system with greater weight wins.



(a) Winning percentages

(b) Agreement levels for different winning systems

Figure 6.5: Human evaluation result

From the subjective evaluation, we first calculate the winning percentage of each system for each criterion. This result is presented in Figure 6.5(a). It is shown that in comparison to the baseline system, the proposed system is perceived as more coherent, having more potential in building emotional connection, and giving a more positive emotional impact.

We investigate this result further by computing the agreement of the final judgement using the Fleiss' Kappa formula. This result is presented in Figure 6.5(b). We separate the queries based on the winning system and compute the overall agreement respectively. It is revealed that the queries where the proposed

[3]The level of trust is provided by the crowdsourcing platform we employ in this evaluation.

Table 6.1: Candidate responses re-ranking based on three consecutive selection constraints. *: baseline response, **: proposed response

Query: Em going to London tomorrow. *(valence: 0.39, arousal: -0.11)*			
Candidate responses	ranking steps		
	semantic	**emotion**	**impact**
* And where in Australia?	1		
[laugh]	2		
Organised people need to have holiday.	3	1	
It would be very unwise for us to discuss possible external examiners.	4		
[laugh]	5		
It's good that sounds eh like a good thing to do, although you wouldn't want to em overspend.	6		
That sounds interesting you've quite a lot going on so you need to manage your time.	7	2	
Yes.	8		
Mhm.	9		
** That sounds nice.	10	3	1

system wins have far stronger agreement than that where the baseline system wins. This suggests a higher confidence and certainty in cases where the proposed system wins.

We analyze the consequence of re-ranking and the effect of emotion similarity in the response selection using queries extracted from the test set. Table 6.1 presents the 10-best semantic similarity ranking, re-ranked and filtered into 3-best emotion similarity ranking, and the candidate response that passed the filtering with the best emotional impact. The table shows that the proposed method can select one of the candidate responses that even though is not the best in semantic similarity score, has a higher score in terms of emotion similarity and expected impact.

Table 6.2: Baseline and proposed responses for identical text with different emotional contexts. The proposed system can adapt to user emotion, while baseline method outputs the same response

Query : Thank you. *(valence: 0.13, arousal: -0.18)*
Baseline : Thank you very much that **Proposed :** And I hope that everything goes exactly according to plan.
Query : Thank you. *(valence: 0.43, arousal: 0.05)*
Baseline : Thank you very much that **Proposed :** It is always a pleasure talking to you you're just like me.

Furthermore, the proposed selection method is able to generate different responses to identical textual input with different emotional contexts. Table 6.2 demonstrates this quality. This shows system's ability to adapt to user's emotion in giving a response. These qualities can contribute towards a more pleasant and emotionally positive HCI.

6.6. Summary

We presented a novel attempt in eliciting postive emotional impact in dialogue response selection by incorporating emotional state and future potential impact in dialogue response selection. We use tri-turn units in place of the traditional query-response pairs to observe emotional responses in the example database.

The data-driven approach we employ in this study is straightforward and could be efficiently replicated and extended. With the increasing access to data and the advancements in emotion recognition, a large unlabeled corpus of conversational data could be used to extensively expand the example database.

We compared the proposed response selection method with the traditional method in terms of coherence, emotional connection, and emotional impact. Subjective evaluation showed that by incorporating emotional state and potential impact in selecting a response, we can elicit a more positive emotional impact in

the user, as well as achieve higher coherence and emotional connection.

In the future, we look forward to try a more sophisticated function for estimating the emotional impact. Further, we hope to test the proposed idea further in a setting closer to real conversation. For example, by using spontaneous social interactions, considering interaction history, and using real-time emotion recognition. We also hope to apply this idea to a more complex dialogue models, such as Partially Observable Markov Decision Process (POMDP), and train the system to learn an explicit dialogue policy using reinforcement learning.

Chapter 7

Conclusion and Future Works

7.1. Conclusion

In this thesis, we attempted to incorporate social-affective knowledge of human communication into HCI. Extending previous works mainly focusing on the communication competence of emotion, we focus on the pattern of emotional triggers and responses evaluated during appraisal process in shaping the emotional state of a person. We attempt to exploit this knowledge provide emotional support in HCI in the form of positive emotion elicitation.

Inspired by the two-way emotion loop, we divide our objective into three tasks: 1) Recognizing affective states, 2) recognizing social-affective events, and 3) eliciting a positive emotional response in HCI. Additionally, we also presented a corpus of spontaneous social-affective interaction in the wild from various television talk shows, containing natural conversations and real emotion occurrences.

We begin our study by constructing an automatic emotion recognizer. We combine the predictions from models of three different modalities to produce multimodal emotion recognizers. When compared to the unimodal result, the multimodal combination successfully attain comparable or higher classification accuracy. In predicting the levels of valence and arousal, we achieve accuracy of 63.7% and 75.6% in English, and 93.8% and 92.6% in Indonesian, respectively.

Secondly, in the social-affect events study, we presented analysis and prediction of emotional triggers and responses based on spontaneous human conversation. This task extends previous works that focuses solely on affective communication competences by analyzing emotional triggers and responses. The

automatic prediction offers an approach in equipping conversational agents and dialogue systems with social-affective awareness: 1) to be able to decide an emotion triggering action, and 2) to be able to predict the appropriate response to an emotion trigger. For English and Indonesian, we achieve an accuracy of 50.8% and 31.5% for the first task, and an average of 52.2% and 57.0% for the second.

Lastly, we attempt to elicit a positive emotional user response in an example-based dialogue system. We impose an emotionally sensitive response selection criterion by incorporating 1) emotional state, and 2) expected future potential impact. We use tri-turn units in place of the traditional query-response pairs to observe and exploit the information of emotional triggers and responses in the example database. Human subjective evaluation reveal that the proposed system is preferred at least 66.7% of the time compared to the baseline on all three evaluation metrics: coherence, emotional connection, and positive emotional impact.

To move closer towards real application of the idea we proposed in this study, we hope to construct a system that is able to perform real time interaction with a human user. As such, the components produced in this work should serve as building blocks for future integration into one working system.

7.2. Future Works

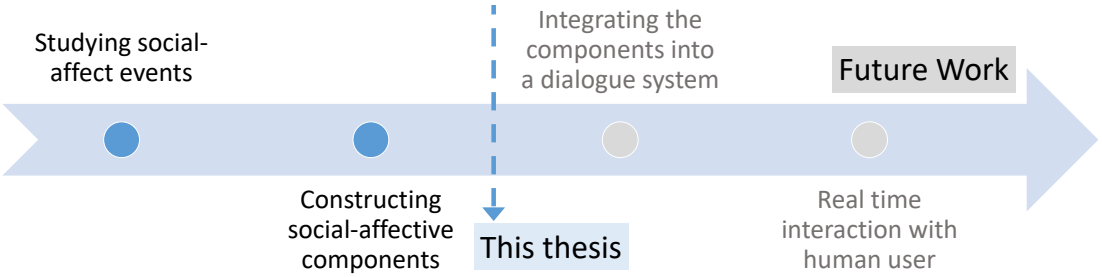


Figure 7.1: This thesis and its future work

This thesis provides initial attempts in utilizing computational approaches on social-affective data to transfer emotion knowledge into HCI. As the experiments are rather modular, in the future, we hope to unify the efforts into one working framework of a dialogue system. As HCI is tailored to the benefit of the user,

we hope to utilize the envisioned system to provide positive emotional support to the user.

It is natural that emotional competences vary from person to person. To the extent of machine learning, it becomes crucial that the model learns to handle user's emotion from a highly competent, i.e. an expert, source. In this regard, collaboration with psychologists, counselors, or therapists has high potential in providing an expert's point of view of the problem. In doing so, it is also important to keep the balance between human labor and automation in the solution. For example, as opposed to knowledge-based rules, a lower-level abstraction of the expert's knowledge, e.g. in forms of data, might be more advantageous in the long run.

Continuous effort should be done in improving the performance of each of the task carried in this thesis. Data is without doubt one of the most defining factor in producing a high-performing models that matches with real life emotion occurrences. Aside from its quantity, the content of the data should be tailored such that the gap between the contained data and real life emotion is as little as possible. The scope of the data reviewed in this thesis is yet to cover all facets of social-affective interactions. As such, it might be beneficial to take a look into other spontaneous dialogue situations in collecting additional data.

Uncertainty will likely always remain a problem in automated affective agents, due to the innumerable factors that can influence emotion. As with human abilities, it might be that perfect accuracy for automated emotion task is unattainable. Until a reliable performance can be achieved, the use of an automated agent should always be balanced with some degree of human intervention; the balance of which should be determined case-by-case according to the intended application of the agent. For example, medium reliability may be acceptable for chat-bots, but certainly not for at-home therapy agent. Furthermore, in case of long-term application, it is possibly the case that instead a general solution that might work for any user, a system that can adapt and improve its performance to a designated user is preferred.

List of Publications

In Collection

- Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, Ayu Purwarianti, and Satoshi Nakamura. "Emotion and its triggers in human spoken dialogue: Recognition and analysis." Situated Dialog in Speech-Based Human-Computer Interaction. Springer International Publishing, 2016. 103-110.

International Conferences

- Nurul Lubis, Dessi Lestari, Ayu Purwarianti, Sakriani Sakti, and Satoshi Nakamura. "Construction and analysis of Indonesian emotional speech corpus." Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA), 2014 17th Oriental Chapter of the International Committee for the. IEEE, 2014. Best Student Paper.

- Nurul Lubis, Dessi Lestari, Ayu Purwarianti, Sakriani Sakti, and Satoshi Nakamura. "Emotion recognition on Indonesian television talk shows." Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014. Best Rated Poster Presentation.

- Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. "Construction and Analysis of Social-Affective Interaction Corpus in English and Indonesian." Proceedings of the 18th Oriental COCOSDA, pp. 202–206, Shanghai, China, October, 2015.

- Nurul Lubis, Sakriani Sakti, Graham Neubig, Koichiro Yoshino, Tomoki Toda, Satoshi Nakamura. "A Study of Social-Affective Communication:

Automatic Prediction of Emotion Triggers and Responses in Television Talk Shows." Proc. of 2015 IEEE ASRU, December 2015.

- Nurul Lubis, Randy Gomez, Sakriani Sakti, Keisuke Nakamura, Koichiro Yoshino, Satoshi Nakamura, Kazuhiro Nakadai. "Construction of Japanese Audio-Visual Emotion Database and Its Application in Emotion Recognition." The Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), 2016

Technical Reports

- Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, Dessi Lestari, Ayu Purwarianti, and Satoshi Nakamura. "Recognition and Analysis of Emotion in Indonesian Conversational Speech." SLP, 2014-SLP-104(1), pp. 1-6. 2014

- Nurul Lubis, Sakriani Sakti, Graham Neubig, Koichiro Yoshino, Satoshi Nakamura. "Predicting Emotional Responses from Spontaneous Social-Affective Interaction Data." ASJ, pp. 7 - 8, Mar, 2016.

- Nurul Lubis, Randy Gomez, Sakriani Sakti, Keisuke Nakamura, Koichiro Yoshino, Satoshi Nakamura, Kazuhiro Nakadai. "Constructing a Japanese Multimodal Corpus from Emotional Monologues and Dialogues." IEICE Technical Report, Vol. 116, No. 378, pp. 9-10, Dec. 2016.

References

[1] Phoebe C Ellsworth and Klaus R Scherer. Appraisal processes in emotion. *Handbook of affective sciences*, 572:V595, 2003.

[2] Klaus R Scherer, Angela Schorr, and Tom Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

[3] Antony SR Manstead and Agneta H Fischer. Social appraisal: The social world as object of and influence on appraisal processes. *Appraisal processes in emotion: Theory, methods, research*, pages 221–232, 2001.

[4] Brian Parkinson, Agneta H Fischer, and Antony SR Manstead. *Emotion in social relations: Cultural, group, and interpersonal processes*. Psychology Press, 2004.

[5] KR Scherer. Component models of emotion can inform the quest for emotional competence. *The science of emotional intelligence: Knowns and unknowns*, pages 101–126, 2007.

[6] Byron Reeves and Clifford Nass. *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press, 1996.

[7] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[8] Sangdo Han, Yonghee Kim, and Gary Geunbae Lee. Micro-counseling dialog system based on semantic content. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 63–72. Springer, 2015.

[9] Myrthe Tielman, Mark Neerincx, John-Jules Meyer, and Rosemarijn Looije. Adaptive emotional expression in robot-child interaction. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 407–414. ACM, 2014.

[10] Kate Forbes-Riley and Diane Litman. Adapting to multiple affective states in spoken dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226. Association for Computational Linguistics, 2012.

[11] Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. Effects of self-disclosure and empathy in human-computer dialogue. In *Proceedings of Spoken Language Technology Workshop*, pages 109–112. IEEE, 2008.

[12] Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of Association for Computational Linguistics (1)*, pages 964–972, 2013.

[13] Rosalind W Picard and Jonathan Klein. Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with computers*, 14(2):141–169, 2002.

[14] Marcin Skowron, Mathias Theunis, Sebastian Rank, and Arvid Kappas. Affect and social processes in online communication–experiments with an affective dialog system. *Transactions on Affective Computing*, 4(3):267–279, 2013.

[15] David Benyon, Björn Gambäck, Preben Hansen, Oli Mival, and Nick Webb. How was your day? evaluating a conversational companion. *Transactions on Affective Computing*, 4(3):299–311, 2013.

[16] Timothy W Bickmore and Rosalind W Picard. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327, 2005.

[17] Amy N Cohen, Constance Hammen, Risha M Henry, and Shannon E Daley. Effects of stress and social support on recurrence in bipolar disorder. *Journal of affective disorders*, 82(1):143–147, 2004.

[18] Tanja Bänziger, Hannes Pirker, and K Scherer. Gemep-geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions. In *Proceedings of LREC*, volume 6, pages 15–019, 2006.

[19] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Transactions on Affective Computing*, 3(1):5–17, 2012.

[20] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[21] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. The humaine database: addressing the collection and annotation of naturalistic and induced emotional data. In *Affective computing and intelligent interaction*, pages 488–500. Springer, 2007.

[22] Fabien Ringeval, Andreas Sonderegger, Jens Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[23] John Langshaw Austin. *How to do things with words*, volume 367. Oxford university press, 1975.

[24] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

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

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

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

[28] Robert Plutchik. *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division, 1980.

[29] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

[30] Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing emotion in speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1970–1973. IEEE, 1996.

[31] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9):1162–1171, 2011.

[32] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech*, pages 223–227, 2014.

[33] Björn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, et al. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *INTERSPEECH*, pages 2253–2256, 2007.

[34] Gilly Leshed and Joseph'Jofish' Kaye. Understanding how bloggers feel: recognizing affect in blog posts. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1019–1024. ACM, 2006.

[35] Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)*, 5(2):165–183, 2006.

[36] Anton Batliner, Stefan Steidl, Björn Schuller, Dino Seppi, Kornel Laskowski, Thurid Vogt, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic

Kessous, et al. Combining efforts for improving automatic classification of emotional user states. *Proc. IS-LTC*, pages 240–245, 2006.

[37] Bjorn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Emotion recognition from speech: putting asr in the loop. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4585–4588. IEEE, 2009.

[38] Bai Xue, Chen Fu, and Zhan Shaobin. A study on sentiment computing and classification of sina weibo with word2vec. In *2014 IEEE International Congress on Big Data*, pages 358–363. IEEE, 2014.

[39] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.

[40] Mingli Song, Jiajun Bu, Chun Chen, and Nan Li. Audio-visual based emotion recognition-a new approach. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–1020. IEEE, 2004.

[41] Malika Meghjani, Frank Ferrie, and Gregory Dudek. Bimodal information analysis for emotion recognition. In *Applications of Computer Vision (WACV), 2009 Workshop on*, pages 1–6. IEEE, 2009.

[42] Sara Zhalehpour, Zahid Akhtar, and Cigdem Eroglu Erdem. Multimodal emotion recognition with automatic peak frame selection. In *Innovations in Intelligent Systems and Applications (INISTA) Proceedings, 2014 IEEE International Symposium on*, pages 116–121. IEEE, 2014.

[43] Yongjin Wang and Ling Guan. Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia*, 10(5):936–946, 2008.

[44] Paul Ekman, Wallace V Friesen, and J Hager. The facial action coding system (facs): A technique for the measurement of facial action. *Palo Alto: Consulting Psychologists*, 1978.

[45] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[46] Maureen Caudill. Neural networks primer, part i. *AI expert*, 2(12):46–52, 1987.

[47] Irfan A. Essa and Alex Paul Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):757–763, 1997.

[48] Lawrence S Chen, Thomas S Huang, Tsutomu Miyasato, and Ryohei Nakatsu. Multimodal human emotion/expression recognition. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 366–371. IEEE, 1998.

[49] Florian Eyben, Martin Wöllmer, and Björn Schuller. OPENsmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[50] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[51] Tadas Baltru, Peter Robinson, Louis-Philippe Morency, et al. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[52] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[53] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification. 2003.

[54] Nio Lasguido, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Utilizing human-to-human conversation examples for a multi domain chat-oriented dialog system. *Transactions on Information and Systems*, 97(6):1497–1505, 2014.

[55] Björn Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315, 2009.

[56] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[57] James Bergstra, Frédéric Bastien, Olivier Breuleux, Pascal Lamblin, Razvan Pascanu, Olivier Delalleau, Guillaume Desjardins, David Warde-Farley, Ian Goodfellow, Arnaud Bergeron, et al. Theano: Deep learning on GPUs with python. In *NIPS 2011, BigLearning Workshop, Granada, Spain*, 2011.

[58] Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, and Gary Geunbae Lee. Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication*, 51(5):466–484, 2009.