PhD Dissertation

International Doctorate School in Information and Communication Technologies

DIT - University of Trento

Computational Models for Analyzing Affective Behavior and Personality from Speech and Text

Firoj Alam

Advisor

Prof. Giuseppe Riccardi	University of Trento

Examining Committee:

Prof. Nicu Sebe	University of Trento
Prof. Anna Esposito	Seconda Università di Napoli
Prof. Marco Cristani	Università degli Studi di Verona

January, 2017

Abstract

Automatic analysis and summarization of affective behavior and personality from human-human interactions are becoming a central theme in many research areas including computer and social sciences and psychology. Affective behavior are defined as short-term states, which are very brief in duration, arise in response to an event or situation that are relevant and change rapidly over time. They include *empathy*, *anger, frustration, satisfaction*, and *dissatisfaction.* Personality is defined as individual's longer-term characteristics that are stable over time and that describe individual's true nature. The stable personality traits have been captured in psychology by the Big-5 model that includes the following traits: *openness, conscientiousness, extraversion, agreeableness* and *neuroticism.* Traditional approaches towards measuring behavioral information and personality use either observer- or self- assessed questionnaires. Observers usually monitor the overt signals and label interactional scenarios, whereas self-assessors evaluate what they perceive from the interactional scenarios. Using this measured behavioral and personality information, a typical descriptive summary is designed to improve domain experts' decision-making processes. However, such a manual approach is time-consuming and expensive. Thus it motivated us to the design of automated computational models. Moreover, the motivation of studying affective behavior and personality is to design a behavioral profile of an individual, from which one can understand/predict how an individual interprets or values a situation. Therefore, the aim of the work presented in this dissertation is to design automated computational models for analyzing affective behavior such as *empathy*, *anger*, *frustration, satisfaction,* and *dissatisfaction* and Big-5 personality traits using behavioral signals that are expressed in conversational interactions.

The design of the computational models for decoding affective behavior and personality is a challenging problem due to the multifaceted nature of behavioral signals. During conversational interactions, many aspects of these signals are expressed and displayed by overt cues in terms of verbal and vocal non-verbal expressions. These expressions also vary depending on the type of interaction, context or situation such as phone conversations, face-to-machine, face-to-face, and social media interactions. The challenges of designing computational models require the investigation of 1) different overt cues expressed in several experimental contexts in real settings, 2) verbal and vocal non-verbal expressions in terms of linguistic, visual, and acoustic cues, and 3) combining the information from multiple channels such as linguistic, visual, and acoustic information.

Regarding the design of *computational models of affective behavior*, the *contributions* of the work presented here are

1. analysis of the call centers' conversations containing agents' and customers' speech,
2. addressing of the issues related to the segmentation and annotation by defining operational guidelines to annotate empathy of the agent and other emotional states of the customer on real call center data,
3. demonstration of how different channels of information such as acoustic, linguistic, and psycholinguistic channels can be combined to improve for both conversation-level and segment-level classification tasks, and
4. development of a computational pipeline for designing *affective scenes*, i.e., the emotional sequence of the interlocutors, from a dyadic conversation.

In designing *models for Big-5 personality traits*, we addressed two important problems; *computational personality recognition*, which infers self-assessed personality types, and *computational personality perception*, which infers personalities that observers attribute to an individual. The *contributions* of this work to personality research are

1. investigation of several scenarios such as broadcast news, human-human spoken conversations from a call center, social media posts such as Facebook status updates and multi-modal youtube blogs,
2. design of classification models using acoustic, linguistic and psycholinguistic features, and
3. investigation of several feature-level and decision-level combination strategies.

Based on studies conducted in this work it is demonstrated that fusion of various sources of information is beneficial for designing automated computational models. The computational models for affective behavior and personality that are presented here are fully automated and effective - they do not require any human intervention. The outcome of this research is potentially relevant for contributing to the automatic analysis of human interactions in several sectors such as customer care, education, and healthcare.

Keywords

Affective Behavior, Personality (Big-5), Vocal-Nonverbal Cues, Computational Models

Acknowledgements

I would like to thank my supervisor, Prof. Giuseppe Riccardi, for his support and guidance throughout this journey. I must admit, the opportunity he has given me to learn, and the freedom to work with other people in and outside of the research lab brought me where I am today.

Throughout my PhD, I have had the opportunity to work with so many wonderful people. I must mention the name of Morena Danieli from whom I learned so much about the theory of affective behavior. Her cooperative nature made my time with her abundantly resourceful and I had a great experience while working with her. I could never forget the support I received from Evgeny A. Stepanov during the stressful time of my PhD journey. I still remember the early days of my PhD, specifically the day when he was helping me to understand the "KS test" and assisting me in solving the problem. Many of my research ideas came from the discussions we had during our coffee breaks. Of course, it has also been a wonderful experience to work with other members of the lab - Fabio, Arindam, Carmelo, Orkan, Michael, and Shammur. Special thanks to Michael for proof reading many of my work.

I would like to thank Andrea and Francesca and other members of ICT doctoral school for all the administrative supports I received during the period. I would like to thank all of the secretaries of our lab - Katalin, Helena, Anna, Carolina, and Piera for making our regular activities easier and manageable. During my research work, I have been misusing the SISL server extensively. I would like to thank Veronica for taking care of such chaos all the time. I also have to mention that the *Welcome office* at the University of Trento. The support they provide has been indispensable.

I am grateful for the opportunities and knowledge I received while working with many NLP experts at FBK. My special thanks goes to Bernardo Magnini, Roberto Zanoli, Alberto Lavelli and Faisal Chowdhury from whom I learned the fundamentals and gained NLP experiences.

At the end of this journey, I must thank Prof. Mumit Khan, my undergraduate supervisor and former chairperson of CSE department, BRAC University, who taught me the alphabets of the research. I have to admit, I am lucky to get his guidance and him as my mentor, which shaped my research carrier.

I would like to thank all my friends who were part of this journey. Special thanks goes to Hasnat and Murtoza, for listening to my complaints, providing mental support, and helping me in so many different ways. I would also like to thank all members of the BDUnitn family for making Trento memorable for me.

The path to PhD is always very challenging, there are many emotional ups and downs. It is Shammur, my beloved wife, and a good friend, who has been there to support and inspire me to tackle those emotional imbalances. I am also thankful to her for understanding, holding my hand, being supportive and cherishing each and every moment with me. Shammur, you know, your understanding and encouragement in many difficult moments have been a great support. I would also like to thank my in-laws. I wish I could spend more time with them, but nonetheless they still believed in me, understood me.

Finally, I would like to thank my parents and family members, for their unconditional support and for the belief they have in me for the things I wanted to do. I am grateful to my elder sister for her support in difficult situations. I found her on my side whenever I needed. My father has given up many things and my mother taught me to be the meaning of persistence and patience. The debt that I owe you both is immeasurable. At this point, I owe you my deepest gratitude and love for your dedication and support.

To my parents
Shamsul Huda Patwary and *Balayeter Nesa*

To my wife
Shammur Absar Chowdhury

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Understanding the human behavioral characteristics is central in many different fields such as psychology, sociology, and other behavioral sciences. The goal of these fields is to understand behavioral characteristics to make better decisions in different communicative scenarios. The traditional approach is to manually code interactive scenarios with behavioral characteristics by observing and listening to overt and covert cues from audio or audio-visual recordings of social interactions. Such coding helps in quantitatively measuring and preparing a concrete summary out of it. This approach is largely manual done by trained experts. It is time consuming, and also expensive, and the resulted coded data may also have a high degree of variability due to the inter-coder, i.e., annotator, differences. The research field named *affective computing* emerged with a goal of providing a computing system with abilities such as: 1) to understand users' affective and other behavioral characteristics, and 2) to manage its own cognitive and rational mechanisms while interacting with users [5]. Such abilities can also help in automatically detecting, analyzing, and interpreting human behavioral characteristics from large-scale conversations, which will reduce time and effort. It can also help in designing more human-like interactive systems. Providing such abilities to the computer requires computational models to deal with different *social interactions*, where we express our behavior using different type of *overt* and *covert* cues [6].

Examples of *social interactions* include synchronous/asynchronous interactions in social media, dyadic phone conversations, face-to-face and face-to-machine, i.e., video-blog, interactions. The *covert* cues include physiological signals such as heart-rate, respiratory activity, and electrodermal activity. The *overt* cues include different verbal (linguistic expressions) and non-verbal (paralinguistic information) phenomena, facial expressions, gestures, and postures, which are displayed, expressed and observable.

In Figure 1.1, an example of a dyadic spoken conversation with the associated behavioral cues is presented. By exploiting these behavioral cues, one can design computational models to automatically detect different functional aspects of behavioral phenomena such as affective behavior, personality, conflicts, stance, and engagement, confidence or certainty in a task or activity.

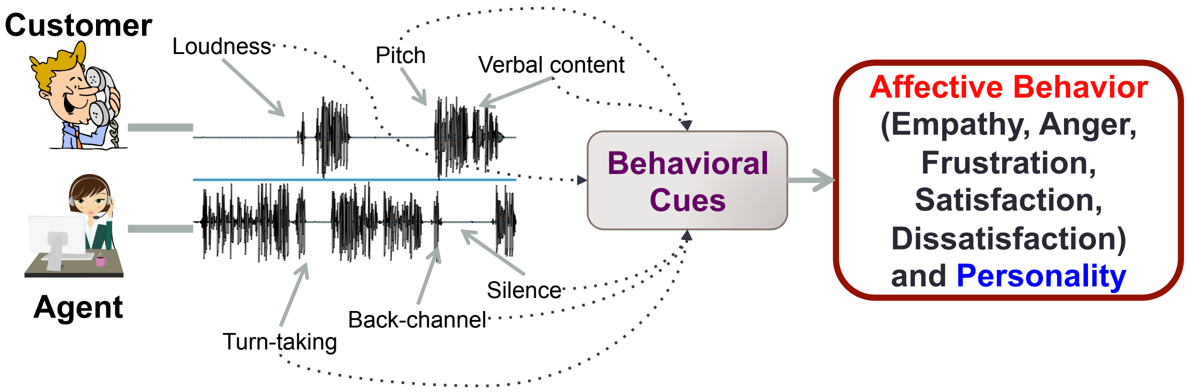


Figure 1.1: Behavioral cues that can be analyzed from spoken conversations to automatically detect affective behavior and personality. Other functional aspects can also be modeled. Affective behavior and personality are shown here for the sake of simplicity.

In Table, 1.1, we present an excerpt of a conversation between agent and customer. It presents different lexical and vocal non-verbal cues. The example of vocal non-verbal cues includes turn-taking, backchannel, and overlap. The example of lexical cues include *diffida* (warning), and *certo* (of course).

Among many behavioral characteristics, modeling *affective behavior* and *personality* are two important areas. In literature, affective behavior are defined as short-term states that fluctuate over time, which include *empathy* and other *basic* and *complex emotions*, whereas personality is defined as long-term traits of an individual that remain constant over the lifetime [7]. There

Table 1.1: An excerpt of a conversation between agent and customer, containing turn-taking, back-channel and lexical cues, e.g., *diffida* (warning). OV: Overlapping turn, BC: Backchannel, BC-OV: Backchannel in a form of overlap. A: Agent, C: Customer. Each row represents a turn in the conversation.

A:	[negare quello che lei ci]$_{OV}$ *[deny what you us]*
C:	[no no no non so non]$_{OV}$ *[no no no I do not know not]*
C:	andate in diffida perché *gone on a warning because*
A:	[una signora io il sistema davanti quindi]$_{OV}$ *[one, Madame, I (am in front of) the system, then]*
C:	[io ho chiamato ho mandato anche una raccomandata]$_{OV}$ [anche una raccomandata mi scusi mi faccia]$_{OV}$ *[I have called (I) have send also a registered (mail)] [also a registered (mail) excuse me let me]*
C:	[parlare]$_{BC-OV}$ *[speak]*
A:	[eh]$_{BC-OV}$ *[huh]*
A:	eh però sul se scadesse al *huh however on if expires at*
A:	[sistema risultano in diffida]$_{OV}$ *[(the) system (they)appear (to be on) warning]*
C:	[sì un computer sono andate in]$_{OV}$ *[yes a computer (they) are gone in]*
C:	diffida quelle lì non io non le ho assolutamente pagate perché il la lettura del *warning those ones (I) do not have paid them absolutely because the reading of the*
C:	[del del]$_{BC-OV}$ *[of the of the]*
A:	[eh]$_{BC-OV}$ *[huh]*
C:	contatore *meter*
C:	[è stato fermo per sì]$_{OV}$ [*(it) was stopped for yes]*
A:	[certo lei]$_{OV}$ *[of course you]*
A:	non le ha pagate quindi sono andate in diffida quindi c è stato un interesse di mora *did not payed them so they had gone on warning so there has been a default interest ...*

are many theories of personality; however, Big-5 traits are the most widely used, in which five traits include *openness, conscientiousness, extraversion, agreeableness* and *neuroticism*. In addition, the studies of personality are mainly focused on two different perspectives: personality perception, which infers personality traits that observers attribute to an individual, and personality recognition, which infers self-assessed personality.

Many state-of-the-art studies of affective behavior proposed computational models, which were designed from the analysis of acted and prototypical data. This characteristic impacts on two important aspects: on the one hand it leads to overestimation of system performance; on the other hand, it is very difficult to extend them in real life applications [8]. Compared to studies on basic and complex emotion there has not been any study on empathy focusing on call center domain. In addition, there has been very few work considered the dynamic aspects of emotional manifestations of the interlocutors in dyadic conversations [9]. This thesis focuses on these aspects in order to overcome the limitations of the current state-of-the-art.

In personality computing research, most of the studies mainly focused on either single modality with prototypical data or multi-modality in a few cases. In the study on personality traits, this thesis focuses on the use of real-life data set such as broadcast news [10], human-human spoken conversations [11] and social media [12, 13]. We investigated different set of features such as acoustic, lexical, parts-of-speech, psycholinguistic and audio-visual. In addition, it also focuses on both personality perception and recognition tasks.

The study of designing automatic computational models poses many challenges (see Section 1.1.1), and at the same time, the outcome of those models has many application sectors such as customer care, education, and health-care (1.1.2).

1.1.1 Research Challenges

The patterns of human behavioral characteristics are complex and multifaceted, which are coupled with heterogeneity and variability. These complex phenomena make the design of a *'universally useful computational system'* a very challenging task. Hence, the current state-of-art focuses on the design of a domain or application specific system(s) with a goal of a single and specific behavioral aspect in mind.

A complete pipeline to the design of a computational model includes designing an experimental scenario, collecting data by capturing expressed cues in the forms of audio/video/physiological recordings, and then extracting the patterns to design the model using machine learning algorithms. There are several challenges in each step of this pipeline.

At first, setting up an experiment to collect ecologically valid[1] data and obtaining a representative number of samples is a major problem and often impossible to get. Then, the collected data needs to be annotated by experts or crowds with the predefined labels. Each individual differs due to the inherent nature of subjectivity and variability. Therefore, in many cases, there are disagreements between expert annotators in identifying and labeling the data (e.g., *speech segment*) with the manifested behavioral expressions. We express behavior by overt and covert cues and if only overt cues are considered it also has many channels such as audio, and visual. Considering only one channel makes the computational task a difficult problem. In many interaction scenarios only an spoken channel is used such as telephone conversations and broadcast radio news. Hence, investigations are necessary to deal with such scenarios. After that, challenges remain to the design of computational models.

One of the important problems that has been hardly addressed in the lit-

[1] Ecological validity often refers to the relation between real-world phenomena and the investigation of these phenomena in experimental contexts [14].

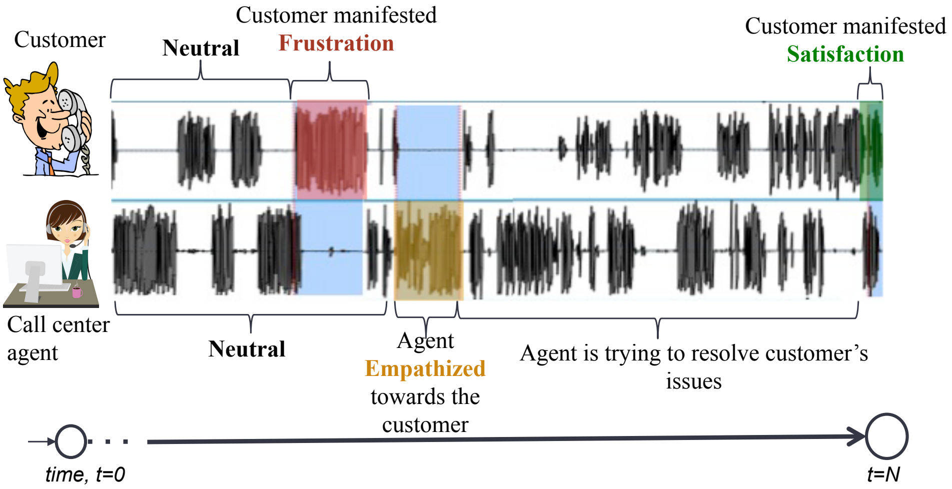


Figure 1.2: Flow of emotional manifestations in a dyadic phone conversation between an agent and a customer.

erature is modeling the dynamics (i.e., the flow) of emotional manifestations between interlocutors. The flow of emotional states has been addressed in Scherer appraisal theory[2]. The theory states that an emotional state, e.g., anger, arises when an individual evaluates a situation and in turn responds to that situation. As depending on the response, the situation might change, which can lead to a different emotional manifestation, e.g., satisfaction. The whole process can be repeated, which is extensively addressed in Gross's *modal model*[3].

To have a clear understanding we present an example of a dyadic interaction with a flow of emotional manifestations in Figure 1.2. As it can be seen in this scenario, when the customer was manifesting frustration, the agent was trying to understand the customer's emotional state. Then, the agent responded by empathizing towards the customer and was trying to resolve the

[2] Appraisal theory [15] states that emotional states can change due to the underlying appraisal (evaluation) and reaction processes of individuals. Scherer describes the dynamics of emotion as a component process model.

[3] The key idea of the modal model is that emotional states unfold over time, and their response may change the environmental stimuli, and that may alter the subsequent instances of that and other emotional states

customer's issues. Finally, customer manifested satisfaction. Both interlocutors were attended and evaluated their interaction and repeatedly responded in this scenario. This dynamic flow of emotional manifestations varies a lot in the time scale, depending on how situation-context evolves during the interaction. Automatically finding this kind of emotion flow, i.e., emotional sequence, is another technical challenge, which this thesis investigates.

1.1.2 Possible Application Domains

The computational models of affective behavior and personality traits have many application sectors. An overview is given below, which we do not claim as a complete list.

Affective Behavior: The sectors that have been investigated for the automatic analysis of affective behavior, particularly emotion from speech, include call center, education such as intelligent-tutoring, healthcare, such as therapist empathy, well-being, such as counseling of a distressed married-couple [8, 16, 17].

(a) **Call-center:** It is one of such sectors, which this thesis mainly focuses on studying affective behavior. An example is shown in Figure 1.3. Some application examples are: "Affective mirror" in which system provides emotional information of the human operator from their voice, Jerk-O-Meter – a system that monitors activity and stress from phone conversations and provides feedback to the user, T-System – emotionally aware voice portal, which aims to provide conciliation strategies by detecting caller's emotional states (see [16] and the references therein). Riccardi et al. [18] investigated the effect of caller's emotional states on the accuracy of spoken dialog systems. They report that automatic detection of caller's emotional states, such as anger, may be beneficial for the adaptation of the system's dialog strategies.

(b) **Education:** In an intelligent-tutoring system with an interaction between child and computer, typically, the goal is to design models to compute the child's uncertainty and engagement [19]. The other behavioral manifestations that have been studied in this domain include confidence, certainty, frustration, engagement, joy and agreement-disagreement [20–24].

(c) **Healthcare:** In healthcare, many application scenarios can be imagined for identifying different neurological and psychiatric diseases, which include Schizophrenia, Alzheimer disease, Bipolar disorder, Parkinson's disease and Autistic Spectrum Disorder (ASD) [25].

(d) **Other** application scenarios include emotionally aware in-car systems [26], emotional music player, emotionally aware avatar-based chat system, surveillance such as surgeries, crisis management, summarizing conversations in a meeting, media retrieval, games and human-robot interaction [8, 17, 27, 28].

Personality Traits: Computational models for personality traits recognition may be applied to many sectors too. Currently, many companies are trying to understand their customer preferences in buying products [29–32]. Marketers are trying to personalize the products and services based on the customer's personal characteristics. For example, the extraversion trait is related to the different type of music preferences such as country, pop, religious, and soundtrack [33, 34].

(a) **Education:** In an educational context, psychologists are using traits' measures to understand the barriers to learning and performance. For example, research suggests that extroverts do better in school, whereas introvert characteristics are the advantages of an individual at the university [35].

(b) **Healthcare:** Personality traits have many application scenarios in healthcare. Matthews et al. [35] report that personality traits are associated with different mental disorders and can be used as predictors. The therapist usually uses the trait information in order to aid the diagnosis. It helps the therapist in identifying the patient's likely characteristics. As an example, it is highly unlikely that a person with a high score in agreeableness has antisocial personality disorder [35].

(c) **Professional carrier:** Studies have also proved that personality traits are associated with job performance, for example, extroverts characteristics are better for sales positions [35]. The conscientiousness trait is the most important indicator for the job performance, which is associated with the integrity and desirable work behavior. In many industries, personality assessment is commonly used during the selection process. It is also reported that it can be measured automatically during the job interview [36, 37]. Depending on the profession, the positive and negative association varies for each trait. For example, conscientiousness is positively correlated with creative and artistic professionals whereas it is negatively correlated with health-service professionals. Those who has a high score in neuroticism and low in conscientiousness are most likely to change the profession.

(d) **Communicative scenario:** The style of communications like emails, blog entries [38] and the choice of particular parts-of-speech depends on the author's personality. Hence, automatically customized email response can be written by matching the individual's personality.

(e) **Features in Other applications:** The personality traits can be used as features for many other tasks in social media, for example, Celli et al. used them as features in agreement *vs* disagreement classification task [39].

In addition to the application sectors discussed above, a descriptive summary of affective behavior and personality traits can be utilized in many areas to predict the characteristics of an individual.

1.2 Addressed Research Problems

Developing computational models for different behavioral constructs are a challenging problem. It is due to the variability in the manifestations of behavioral patterns, diverse definitions for the same concept such as empathy, and the variety of domain and application scenarios. As mentioned earlier, the traditional approach of measuring behavioral constructs is done manually by trained experts [6]. It is reported in the literature that human experts can only analyze less than 1% of the data in call centers [40] as it is a time-consuming and a labor-intensive task. In order to facilitate the domain experts,[4] while counseling, consulting and providing services, and scaling up the processes, we investigated the following research question:

Can we summarize conversations with a description of personality and affective behavior using verbal and vocal non-verbal cues? Can we design affective scene from affective behavior?

Such a research question is very broad due to the multifaceted and heterogeneous nature of behavioral manifestations. Therefore, we narrowed down our research focus, and specifically investigated *affective behavior such as empathy, basic and complex emotions*[5] and *personality traits* using verbal and vocal non-verbal cues. While figuring out the answer to this research question we had an application scenario in mind as depicted in Figure 1.3. In this application scenario, agent and customer are interacting in a call center, and the idea is to automatically analyze the conversation and prepare a

[4] By domain experts, we refer to the call center managers, therapist and decision makers for example.

[5] We use the term *basic emotion* for anger and *complex emotion* for frustration, satisfaction, and dissatisfaction.

descriptive summary using the information of affective behavior and personality traits. The descriptive summary can facilitate domain experts such as call center managers, decision makers, or it can also help the agent in real time too.

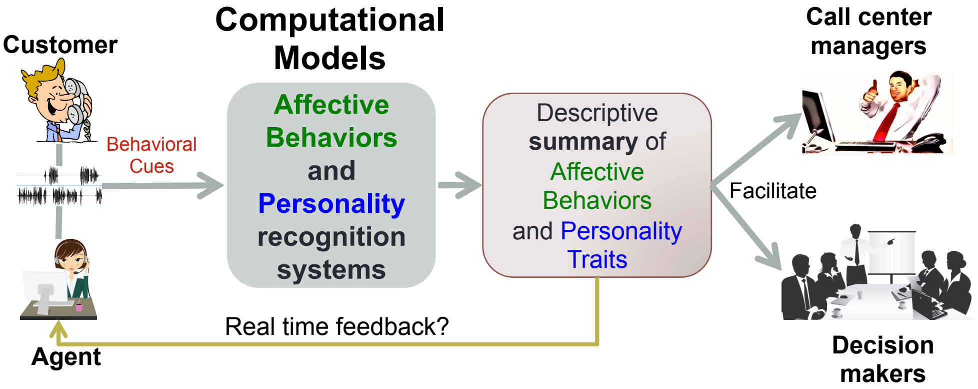

Customer
Computational Models
Behavioral Cues
Affective Behaviors and Personality recognition systems
Descriptive summary of Affective Behaviors and Personality Traits
Call center managers
Facilitate
Decision makers
Real time feedback?
Agent

Figure 1.3: An application scenario for the call center.

Designing computational models involves the following challenges:

- Annotation of ecologically valid data with real behavioral expressions requires an operational definition and guidelines. For example, there has not been any operational definition for annotating and modeling *empathy* for the call-center scenario.
- Annotation of affective behavior in a continuous time scale is another important problem due to the variability and disagreement of the segment boundary of an emotional manifestation.
- Automatically generating the emotional sequence poses different challenges such as segmentation of spoken conversations and assigning a label to the corresponding segments.
- In any ecologically valid dataset, the class imbalance is an important problem for designing the computational model, which requires the investigation of different sampling techniques.
- Since any behavioral construct is manifested using different verbal and

vocal non-verbal cues, therefore, it is necessary to investigate each linguistic and acoustic information independently and in combination. This requires the investigation of different combination strategies at the feature- and decision- level.

- The way of human interaction differs in different communicative scenarios such as human-human, human-machine. Hence, it is necessary to investigate the capability of the automatic system in different scenarios.
- The design of a complete automated pipeline where no human intervention is required.

1.3 Contributions

This thesis focuses on designing computational models for *affective behavior* and *personality traits* using expressed behavioral cues, i.e., overt cues, such as linguistic, and paralinguistic information by investigating ecologically valid real call center and social media conversations.

1.3.1 Affective Behavior

The contributions of affective behavior include the study of *empathy*, of a call center's agent, and basic and complex *emotional states* of the customer such as *anger*, *frustration*, *satisfaction* and *dissatisfaction*. From the emotional manifestation of the agent and customer, we designed *affective scene*. The *computational models for classifying the manifestation of empathy and the design of affective scene are the first contributions in the field of affective computing research.*

- We investigated a large set of, ecologically valid and inbound, call center phone conversations. In a typical conversation, the customer interacted with the call center agent in order to solve a problem or ask for information. An operational annotation guidelines have been defined by following the principle of Gross modal model [4]. Annotators, expert

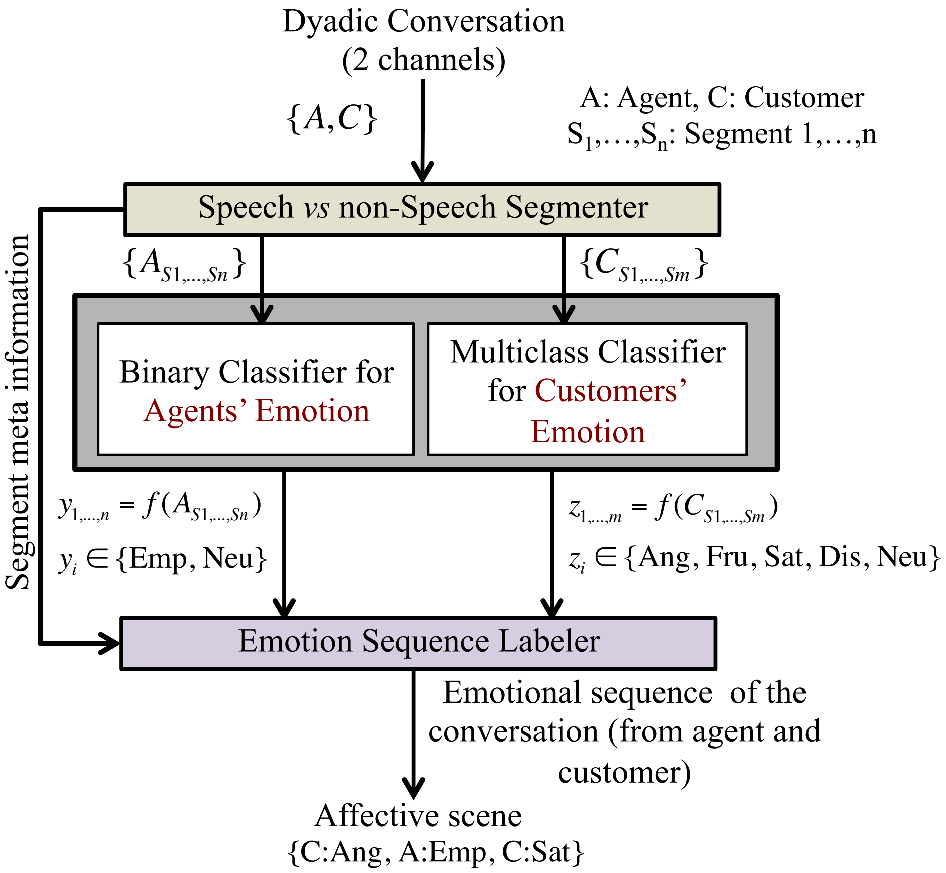

Dyadic Conversation
(2 channels)
{A,C}
A: Agent, C: Customer
S1,…,Sn: Segment 1,…,n
Speech vs non-Speech Segmenter
Segment meta information
{A_S1,...,Sn}
{C_S1,...,Sm}
Binary Classifier for Agents' Emotion
Multiclass Classifier for Customers' Emotion
y_1,...,n = f(A_S1,...,Sn)
y_i ∈ {Emp, Neu}
z_1,...,m = f(C_S1,...,Sm)
z_i ∈ {Ang, Fru, Sat, Dis, Neu}
Emotion Sequence Labeler
Emotional sequence of the conversation (from agent and customer)
Affective scene
{C:Ang, A:Emp, C:Sat}

Figure 1.4: System for generating affective scene, i.e., emotional sequence, for the whole conversation. Agent's emotional states empathy (Emp) and neutral (Neu). Customer's emotional states anger (Ang), frustration (Fru), satisfaction (Sat), dissatisfaction (Dis) and neutral (Neu).

psychologists, have used the guidelines to annotate the manifestation of empathy and other emotional states in a continuous time scale.

- We designed computational models both at the conversation- and segment-level and investigated the different type of features such as acoustic, linguistic, and paralinguistic. We also conducted experiments with the feature- and decision- level combinations. Our findings suggest that decision-level combination is better than the feature-level combination.
- In any real setting, the manifestation of an emotional state is typically less frequent than a neutral state. It results in a skewed label distribu-

tion, which is a challenging problem for designing computational models using machine learning techniques. Our contributions to this problem include the investigation of down-sampling the majority class at the data-level and up-sampling the minority class at the feature-level.

- Another important problem is the mismatch between manual *vs* automatic segment boundaries, which poses a great challenge for the evaluation of the system. We propose algorithmic steps to solve this problem.
- Finally, we present that the segment-level classification model can be used to generate affective scene, i.e., emotional sequence, as depicted in Figure 1.4. The research question that we investigate for studying the emotion sequence is "*who* expresses *what* type of emotional state *when*".

1.3.2 Personality Traits

Contributions to the design of classification models for personality traits include the investigation of diverse corpora that differs in terms personality traits *perception*[6] and *recognition*[7] tasks. The choice of tasks also varies in terms of source speaking styles and experimental context. There have been many theories and models for personality, and for this research we have used Big-5 personality traits model, which is widely used and most influential in the literature [41]. Both perception and recognition tasks have been set by designing binary classifiers for each personality trait. Our contributions regarding the study of personality trait include the investigation of:

- different feature representations of lexical features such as boolean, frequency and tf-idf representation,
- feature type such as acoustic, lexical, parts-of-speech, psycholinguistic, emotional, traits labels as features,
- feature- and decision- level combination/fusion, and

[6] Infers personality traits that observers attribute to an individual.

[7] Infers self-assessed personality traits.

- different feature selection and classification algorithms.

1.4 Publications Relevant to the Thesis

The following publications are relevant to this thesis, which are revised in the preparation of the thesis.

- Firoj Alam, Morena Danieli, Giuseppe Riccardi, Annotating and Modeling Empathy in Spoken Conversations, *submitted to IEEE Transactions on Affective Computing*, 2016.
- Fabio Celli, Arindam Ghosh, Firoj Alam, Giuseppe Riccardi, In the Mood for Sharing Contents: Emotions, Personality and Interaction Styles in the Diffusion of News, Information Processing & Management, Elsevier, 2016.
- Firoj Alam, Fabio Celli, Evgeny A. Stepanov, Arindam Ghosh, Giuseppe Riccardi, The Social Mood of News: Self-reported Annotations to Design Automatic Mood Detection Systems, Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, Co-located with COLING 2016, Osaka, Japan, 2016.
- Firoj Alam, Shammur Absar Chowdhury, Morena Danieli, Giuseppe Riccardi, How Interlocutors Coordinate with each other within Emotional Segments?, COLING 2016, Osaka, Japan, 2016.
- Firoj Alam, Morena Danieli, Giuseppe Riccardi, Can We Detect Speakers' Empathy?: A Real-Life Case Study, 7th IEEE International Conference Cognitive InfoCommunications, 2016.
- E. A. Stepanov, B. Favre, F. Alam, S. A. Chowdhury, K. Singla, J. Trione, F. B'echet, G. Riccardi, Automatic Summarization of Call-center Conversations, IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015), Scottsdale, Arizona, USA, 2015.

- Morena Danieli, Giuseppe Riccardi, Firoj Alam, Emotion Unfolding and Affective Scenes: A Case Study in Spoken Conversations, Proceedings of the International Workshop on Emotion Representations and Modeling for Companion Technologies, pp. 5-11, ACM (ICMI), 2015.
- Morena Danieli, Giuseppe Riccardi, Firoj Alam, Annotation of Complex Emotions in Real-Life Dialogues: The Case of Empathy, Clic-it14, 2014.
- Firoj Alam, Giuseppe Riccardi, Predicting Personality Traits using Multimodal Information, Proceedings of the 2014 ACM Multimedia on Workshop on Computational Personality Recognition, pp. 15-18. ACM, 2014, Orlando, USA.
- Firoj Alam, Giuseppe Riccardi, Fusion of Acoustic, Linguistic and Psycholinguistic Features for Speaker Personality Traits Recognition, ICASSP2014 - Speech and Language Processing (ICASSP2014 - SLTC), 04-09 May 2014, Florence, Italy.
- Firoj Alam, Giuseppe Riccardi, Comparative Study of Speaker Personality Traits Recognition in Conversational and Broadcast News Speech, Interspeech-2013, 25-29 August 2013, Lyon, France.
- Firoj Alam, Evgeny A. Stepanov, Giuseppe Riccardi, Personality Traits Recognition on Social Network - Facebook, WCPR-2013 (ICWSM-13), Cambridge, MA, USA.

1.5 The Structure of the Thesis

This thesis addresses speaker/users affective behavior and personality traits from their conversations, which are particularly two independent tasks. Therefore, the chapters are organized into two parts, I) affective behavior such as empathy, basic and complex emotions, II) the recognition and perception of personality traits.

The *first part* presents the work on affective behavior. In Chapter 2, we present a review of historical traces and contemporary research of emotion in general, empathy and also research in affective computing. The details of corpora, which have been used for the study of automatic classification experiments is presented in Chapter 3. For the experiments, we used SISL behavioral corpus of call-centers' phone conversations and FAU-Aibo robot corpus of child-computer interactions. In Chapter 4, we present the features, classification algorithms and evaluation approaches that we investigated throughout the research work. We present the study of empathy in Chapter 5, which includes an in-depth analysis at the segment level, classification experiments at the conversation- and segment- level. In Chapter 6, we discuss our study on the classification of basic and complex emotions at the conversation- and segment- level. Towards segmenting and labeling emotional states of conversations, we investigated the HMM-based sequence labeling approach, which we present in Chapter 7. We present the study of *affective scene*, i.e., emotional sequence, in Chapter 8. Then, in Chapter 9 we discuss our study of cross-corpus emotional classification. For the cross-corpus study, we experimented three different settings such as intra, inter and mixed by utilizing acoustic features. We will conclude this part with a brief summary in Chapter 10.

In the *second part* of the thesis, we present our investigation to the design of computational models for the personality traits that varies in terms of experimental context, communication types such as social media, dyadic interaction, face-to-machine interaction, i.e., video-blog, and broadcast news. The variation of the tasks also includes personality traits perception and recognition. In Chapter 11, we present a review of the current state-of-the-art of the study of personality both in psychology and affective computing. In Chapter 12, we present the study of personality traits recognition using Facebook dataset, which contains users' "Facebook status updates" annotated with self-assessment. Our goal was to find the lexical evidence in discriminating different traits while we explored different classification algorithms. In Chapter 13, we report our study of personality traits perception and recognition using corpora that differ in source speaking style, and experimental context. The utilized corpora include Speaker Personality Traits corpus collected from the recordings of broadcast news and SISL-Persia (SIS Lab Personable and Intelligent Virtual Agents) corpus, which is an Italian human-human spoken dialog corpus. We investigated the different type of features, feature selection, feature and decision level combination/fusion. We present our study of a multi-modal corpus in Chapter 14, where we investigated the spoken and visual cues independently and combinedly. We present our study of mood, communication style and personality traits in Chapter 15. After that with a brief summary in Chapter 16, we will conclude this part. Finally, in Chapter 17, we conclude the thesis with a summary and discuss the limitation and future directions.

1.6 Terminology

In this Section, we highlight the terminology and concepts that are relevant for our study.

Behavior: It is defined as "... quite broadly to include anything an individual does when interacting with the physical environment, including crying, speaking, listening, running, jumping, shifting attention, and even thinking." [42].

Behavioral Signals/Cues: Signals that are direct manifestations of individual's internal states being affected by the situation, the task and the context. Cues are patterns of the signals and they can be overt or covert. Examples of overt cues are changes in the speaking rate or lips getting stiff. Examples of covert cues are changes in the heart-rate or galvanic skin response.

Affect: It is an umbrella term that covers a variety of phenomena that we experience such as emotion, stress, empathy, mood, and interpersonal stance [43, 44]. All of these states share a special affective quality that sets them apart from the neutral states. In order to distinguish between each of them, Scherer [2] defined a design-feature approach, which consists of seven distinguished dimensions, including intensity, duration, synchronization, event-focus, appraisal elicitation, rapidity of change, and behavioral-impact.

Affective Behavior: The component of the behavior that can be explained in terms of affect analysis.

Emotion: There is a variety of definitions of this concept. A few of them are reported below.

According to Scherer, emotion is a relatively brief and synchronized response, by all or most organismic subsystems, to the evaluation of an external or internal stimulus.

Gross's [45] definition of emotion refers to its *modal model*, which is based on three core features such as 1) what gives rise to emotions (when an individual attend and evaluate a situation), 2) what makes up an emotion (subjective experience, behavior, and peripheral physiology), and 3) malleability of emotion.

According to Frijda "emotions are intense feelings that are directed to someone or something." [43, 46].

Emotional State: The state of an individal's emotions. An emotional state is a product of the psychological and physiological processes that generate an emotional response, and that contextualize, regulate, or otherwise alter such responses [47].

Mood: It is "a feeling that tend to be less intense than emotions and that often lacks contextual stimulus" [43, 48].

According to Scherer [2] mood is a "diffuse affect state most pronounced as change in subjective feeling, of low intensity but relatively long in duration, often without apparent cause." It includes cheerful, gloomy, irritable, listless, depressed, and buoyant.

Empathy: According to Hoffman [49],
"Empathy can be defined as an emotional state triggered by another's emotional state or situation, in which one feels what the other feels or would normally be expected to feel in his situation."

By McCall and Singer [50], empathy is defined based on four key components.: "First, empathy refers to an affective state. Secondly, that state is elicited by the inference or imagination of another person's state. Thirdly, that state is isomorphic with the other person's state. Fourthly, the empathizer knows that the other person is the source of the state. In other words, empathy is the experience of vicariously feeling what another person is feeling without confounding the feeling with one's own direct experience."

Perry and Shamay-Tsoory [51] "... denotes empathy as our ability to identify with or to feel what the other is feeling."

Affective Scene (Emotional Sequence): It is defined as "an emotional episode where an individual is interacting, in an environment, with a second individual and is affected by an emotion-arousing process that (a) generates a variation in their emotional state, and (b) triggers a behavioral and linguistic response" [52]. The affective scene extends from the event-triggering of the 'unfolding of emotions' throughout the closure event when individuals disengage themselves from the communicative context.

Big-5: The "Big-5" factors of personality are five broad dimensions of personality that are used to describe unique individuals [53]. Each trait in "Big-5" are defined with adjective markers as follows:

Openness (O): Artistic, curious, imaginative.

Conscientiousness (C): Efficient, organized, responsible.

Extraversion (E): Energetic, active, assertive.

Agreeableness (A): Compassionate, cooperative, friendly.

Neuroticism (N): Anxious, tense, self-pitying. The opposite direction is referred to as Emotional Stability

Part I

Affective Behavior

In this part of the thesis, we discuss our work on designing computational models for affective behavior. We designed and evaluated the models using real call center's dyadic spoken conversations. We mainly focused on studying empathy, designing affective scene from the whole conversation, investigating conversation and segment level cues in terms of acoustic, lexical and psycholinguistic cues. In Chapter 2, we present the state-of-the-art that are relevant with our line of research, then in Chapter 3 we provide a detail study of the dataset. We present the details of the feature extraction, classification and evaluation methods in Chapter 4. Classification study of empathy and other emotional states is presented in Chapter 5 and 6. In order to design emotion sequence we also investigated a generative approach, which we present in Chapter 7. The complex dynamics of emotional manifestations in a conversation appears in the form of emotion sequence, which we define in terms of the affective scene. We present our study of affective scene in Chapter 8. One of the great challenge is to design system that is usable across domain/language. We present our study of cross-language emotion classification study in Chapter 9. A summary of the study of affective behavior is presented in Chapter 10.

Chapter 2

State-of-the-Art: Affective Behavior

Understanding the manifestations of affective behavior and emotional states has a long story, and still it is an ongoing research. In this chapter, we will provide a brief overview of the research in psychology and affective computing with a particular focus on empathy and other emotion-related states.

We will start by providing the introduction of the terminology that we will be using throughout this thesis.Then in the subsequent sections, we will discuss some relevant studies belonging to psychology and affective computing scientific domains. We make no attempt to extensively review the massive literature focusing on the different aspects of the psychology of emotion as that has not been our main goal. With our effort, we attempted to highlight the notable reviews and work both in psychology and affective computing. In Figure 2.1 we present a rough timeline. We do not claim that it is exhaustive and complete, and we are only reporting it to have a bird's-eye view of the historical traces of the study of emotion.

2.1 A Brief History

The historical attempt to understand emotion can be traced from the study of the Greek philosopher Plato (428-347 B.C.), who suggested that the mind has the tripartite structure such as *cognition*, *emotion* and *motivation* [2, 54]. There have been debates that emotion does not exist in these three aspects, rather, there are a reason, spirit, and appetite. Emotion is not only divided between spirit and appetite, but it also remains in reason as well. This ancient debate is currently known as "cognition-emotion" debate.

The Greek philosopher Aristotle (384-322 B.C.) supported the idea of Plato and attempted to define taxonomies in which emotion include anger,

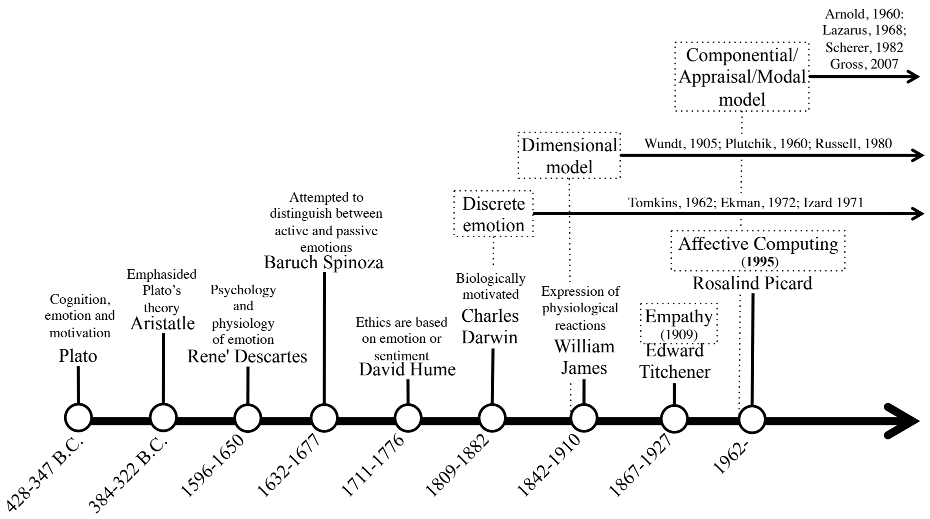


Figure 2.1: Historical records of the study of emotion in psychology and affective computing. Reported most notable traces, which are based on the study of Scherer [2]. Dotted box represents the contemporary research of emotional models.

fear, pity and opposite of these. Aristotle extensively discussed certain emotions, such as anger, but has not defined what are the opposite emotions. His analysis of anger includes cognitive components, social context, behavioral tendency and a recognition of physical arousal. He also noted that any physical or psychological discomfort may lead to anger.

The tripartite structure of Plato is currently adopted by many modern psychologists and putting more emphasis on those three aspects.

French philosopher, René Descartes (1596-1650), mainly focused on mental and physicological processes. In his view, emotions involve not only sensations caused by the physical agitation but also perceptions, desires, and beliefs. Descartes, who is often called the founder of modern psychology, defined emotion as a type of passion, in which passion is defined as "the perceptions, feelings or emotions of the soul which we relate specifically to it, and which are caused, maintained, and fortified by some movement of the animal sprits" [54]. In his value-oriented analysis, he identified six simple and primitive passions such as wonder, love, hatred, desire, joy and sadness.

He also argued that all others are composed of these six. Since his study, the controversy of “mind-body” debate remains till now, in terms of mind-body interaction and mental causation. An example of the present controversy includes finding the relationship between physiological patterns and specific emotional states.

Baruch Spinoza’s (1632-1677) study of emotion includes how to attain happiness, and finding the distinction between active and passive emotions. He argued that happiness can be achieved only once we get straight our thinking around the world. In his opinion, the active emotions are those that are rationally understood whereas the passive is not.

David Hume (1711-1776) defined emotion as a certain kind of sensation or what he called an “impression”, which is physically stimulated by the movement of the “animal spirits” in the blood. His study suggests that there are good emotions such as pride, a bad emotion such as humility.

Charles Darwin’s (1809-1882) book “The Expression of the Emotions in Man and Animals” has been a highly cited research for emotions. His study of emotion mainly concerned about the universality of emotion and emotional expressions. His work on emotion is one of most influential work in modern psychology. He provided a strong emphasis on the expressions of emotion in the face, body, and voice. His study also includes intercultural and developmental approaches. According to Darwin, emotional expressions are evolved and adaptive. In his view emotions are useful as they help people to solve problems, motivates people to engage in actions that are important for survival. Following the principles of Darwin’s theory, there has been a significant amount of work towards finding the universality of emotions and their association with facial expressions. The studies include Ekman’s six basic emotion [55], Izard’s ten universal facial expressions [56, 57], Plutchik eight basic bipolar emotions [58], Tomkins nine emotions [59]. In the Darwinian perspective (Darwin 1872), emotions are evolved phenomena, which are important survival functions.

American psychologist and philosopher William James (1842-1910) were mainly concerned about understanding the nature of emotional experience. He suggested that emotion is the perception of differentiated bodily changes, which are specific for each emotion.

2.2 Contemporary Research

Current theories of emotion greatly differ based on the number of emotions and the idea that evoked the differentiation. The following overview is based on the study of Scherer [2], where he classified the currently used models into four categories. In addition to defining theories by contemporary theorists, the other focuses include 1) the elicitation and differentiation of emotion based on an antecedent evaluation, 2) finding the emotion-specific response patterns by studying different modalities, and 3) finding the effects of emotion on another type of psychological functioning such as memory, learning and thinking.

2.2.1 Dimensional Models

The dimensional approach to emotion conceptualizes emotions by defining where they lie in different dimensions based on the degree of excitation, pleasantness, or relaxation. Dimensional theories are mainly concerned with the subjective feeling components and its verbal reflection. Over the time, several dimensional models have been developed.

Unidimensional models:

Theorists of the unidimensional models agree that one dimension is sufficient to make a basic distinction between different emotions. The dimensions include *activation/arousal* i.e., the physiological and psychological state of being active or inactive, or *valance* i.e., the subjective feeling of pleasantness or unpleasantness. Arousal was defined as the degree to which individuals use the level of subjective arousal denoted by affect words when labeling their subjective emotional states. Whereas, valence was operationally defined as the degree to which individuals use the pleasantness or unpleasantness denoted by affect words when labeling their subjective emotional states. The major difference in activation/arousal dimension is the degree of arousal such as low, high or in between. Many psychologists argued that pleasantness-unpleasantness dimension, as referred as a valance, is the most important determinant for emotional feeling. Based on the degree of pleasantness, they can be grouped into two poles such as unpleasant, bad and disagreeable in one pole, and good, agreeable and pleasant in another pole. Using this dimension one can also distinguish positive and negative emotions.

Multidimensional models:

Wundt [60], one of the pioneers, who proposed the multidimensional model to represent the emotional states. He suggested a three-dimensional model, which include pleasantness-unpleasantness, rest-activation, relaxation-attention. Later Plutchik [58, 61] and Russell [62] contributed to the advancement of the multidimensional models. Plutchik offers a three-dimensional model, in which he identified eight primary emotions and represented them in the circumplex in eight sectors. They are sadness, surprise, fear, trust, joy, anticipation, anger, and disgust. In the circumplex, he placed similar emotions close together and opposites 180 degrees apart with complementary colors. The emotions with no color represent the mixtures of the primary emotions. In this model, the inner circles represent more basic and outer circles more complex emotions. Arousal and valence represent the vertical and the horizontal axis respectively. The third dimension represents the intensity of the emotions. It is needed to mention Harold Schlosberg (1941) is the one who first introduced circumplex model at first.

Russell proposed a two-dimensional model of valence and arousal, in which he also placed emotions on a circumplex. Arousal and valence represent the vertical and the horizontal axis respectively, and the center of the circle represents a neutral valence and a medium level of arousal.

2.2.2 Descrete Emotion Models

Discrete emotion theorist argues that there are a small number of discrete emotion exist, which include 3 to 14 basic or primary emotions. The theories of discrete emotion mainly focused on the study of facial and motor expression, and action system.

Circuit models:

Circuit emotional models are based on a neuropsychological approach to emotion, which argue that emotions and their distinguishing features are determined by neural circuits. Neuroscientist Panksepp suggests that there are four fundamental neural circuits that are responsible for producing well-organized behavioral sequences for neural stimulation such as rage, fear, expectancy and panic [63]. For example, a result of this line of research showed that different neural circuits are implicated in the processing of arousal and valence of interpersonal versus non-interpersonal emotions [64].

Basic and complex emotions:

The biological version of basic emotion theories pioneered by Darwin [65], Tomkins [66], Ekman [67] and Izard [68] suggest that there are a small set of basic emotions. The basic or fundamental emotions include anger, fear joy, sadness, and disgust, and these are developed during the course of evolution. They also suggest that each of these fundamental emotions has their own specific physiological and behavioral reaction patterns. Since it is difficult to represent the large variety of emotions with the small number of emotions, therefore, the theorists in this tradition suggested more complex model such as Plutchik's circumplex model [58, 61].

In Plutchik's circumplex model, we can find the distinction between the basic and complex model regarding their intensities such as rage vs. annoyance, respectively. The study of Oatley and Johnson-Laird [69, 70], suggest that there is a distinction between basic and complex emotion, i.e., the later is derived from the former. Basic emotions are innate distinguished by their distinctive signals in the brain [71] and each of them has their universal nonverbal expressions. These emotions can arise as a result of basic appraisals [70]. Basic emotions are the biological basis for the complex emotions [54]. It is reported that complex emotions integrate a basic emotional signal and a conscious cognitive appraisal [54].

2.2.3 Meaning Oriented Models

Lexical models:

The basic assumption of the theorist of lexical models is that the inherent quality of language such as emotional lexicon might be useful to determine the underlying structure of emotional phenomena. The study of Ortony et al. mainly focused on the structural analysis of emotional lexicon to present the underlying semantic implicational structure [72]. The other theorists used cluster analysis in the form of a tree structure to present the classification of emotional states [73].

Social constructive models:

The theorists in this tradition claim that the meaning of emotion is constructed based on socioculturally determined behavior and value patterns. They embraced the importance of emotion lexicon to reflect the emotional meaning structure in the respective culture.

2.2.4 Componential/Appraisal/Modal Models

One of the major goals of componential theories is to find the link between the elicitation of emotion and their response patterns. The componential models assume that emotions are elicited by a cognitive evaluation of an antecedent event or situation, or based on the individual's subjective appraisals. Also, the pattern of the emotional reactions in different response domain is determined based on the outcome of the evaluation process. The term **appraisal** first coined by Arnold (1960) and Lazarus (1966), and then more extensive studies have been done in the early 1980's [2, 74], leading to the development of *appraisal theory*. On the basis of appraisal theory and some other features (see in Section 2.2.4) of emotion *modal model* has been defined by Gross [4]. *Primary appraisal* is a kind of assessment based on how significant an event is for a person such as whether it is a threat or an opportunity. *Secondary appraisal* is a kind of assessment, which considers the ability of a person to cope or take advantage of the consequences of a situation.

Appraisal models:

Appraisal models of emotion suggest that organisms evaluate events/situations based on the appraisal[1] process in order to determine the nature of ensuing emotion [2]. In the view of appraisal models, the emotion-antecedent evaluation process can occur in an automatic and largely unconscious way. Also, it occurs at different levels of the central nervous system. Moors et al. [74] defined *appraisal* as appraisal processes, which works like a function and produces appraisal values, i.e., appraisals - our evaluations, interpretations, and explanations of events. The appraisals usually lead to different emotional reactions, which is different to different people. Lazarus also defined primary and secondary appraisals (see in [75]).

It is well agreed among the appraisal theorists that emotions are multi-componential episodes and appraisal is a component in these episodes. Their main goal is to find the appraisal factors and values that may determine the cause of emotion elicitation, their differentiation, intensity, and quality.

Modal model:

The *modal model* theory defines emotion based on three core features such as "when gives rise to emotions" and "what makes up the emotion", and "the

[1] Appraisal is an act or a process in order to develop an opinion, assessment, or judgment of an event or situation

malleability of emotion" [76]. The *first feature* state that emotion arises when an individual participate and evaluate a situation by focusing currently active goal(s). The goal and the situation give rise to an emotion. Whenever, the meaning of the goal or situation change over time the emotion might also change. The *second feature* state that emotions involve changes in the domain of subjective experience, behavior and peripheral physiology. The *third feature*, 'malleability' refers to the fact that emotions can be modified as they arise and then play themselves. For example, emotions frequently interrupt our current activity and lead us to think, feel and behave differently than what we are doing. According to the modal model, *"emotions involve person-situation transections that compel attention, have meaning to an individual in light of currently active goals, and give rise to the coordinated yet flexible multisystem responses that modify the ongoing person-situation transection in crucial ways"*. The key idea of the modal model is that emotional responses may change the environmental stimuli, and that may alter the subsequent instances of that and other emotions.

OCC model

Focusing on designing the human-computer interaction systems, Ortony et al. developed a computationally tractable model of emotion [72, 77]. This model mainly adopted by computer scientists in order to synthesize emotions in artificial agents. OCC model suggests that the elicitation of emotion mainly depends on the three aspects of the environment such as events that concern oneself, agents that one considers responsible for such events, and objects of concern. The OCC model defines 22 emotion categories, divided into three classes, and in six groups. It considers emotions as a valenced reaction to events, agents, and objects, and postulate that valenced reactions are the keys to differentiating between emotions and non-emotions.

Summary

As the main goal of studies is to predict the emotional responses of an individual in a particular event/situation/stimulus, therefore, it appears that componential models particularly *modal model* may serve a useful basis to achieve that kind of goal. Because it focuses three essential features that are important to achieve the goal. In our study, we utilized the modal model of emotion in order to define annotation guideline and annotate emotional states.

2.2.5 Research on Empathy

Historical traces:

The traces of linguistic root of empathy can be found from the work of David Hume (1711-1776), who has basically used the term “sympathy” to its semantic meaning of sharing the sentiment. Philosopher Adam Smith (1723-1790) stressed more about the study of sympathy further and mentioned that by imagination we place ourselves in other situation and try to feel other’s feeling in some degree. Charles Darwin (1809-1882) considered sympathy as an essential part of social instinct. Theodor Lipps (1851-1914) transformed the term Einf´uhlung (empathy) from a concept of philosophical aesthetics into a central category of the philosophy of the social and human sciences [78]. Later, the word *empathy* was coined by Edward Titchener (1867-1927) in 1909 [79] as a translation of the German term *Einfühlung*. Since then this concept has been widely used to refer to a wide range of pro-social emotional behavior ranging from sympathy to compassion, and including an accurate understanding of the other person’s feelings. Carl Rogers (1902-1987) is a clinical psychologist and therapist, who is one of the influentials for studying the empathy and founder of client-centered therapy. In his work [80], he suggests that the ideal therapist is, first of all, empathic and it is a central part of the success of a therapist.

Recent studies:

Recently, the hypothesis is that one's empathy is triggered by understanding and sharing others' emotional states has found neuroscientific underpinnings in the discovery of the mirror-neurons system. This is hypothesized to embody the automatic and unconscious routines of emotional and empathic behavior in interpersonal relations. These include action understanding, attribution of intentions (*mind-reading*), and recognition of emotions and sensations [81]. In everyday life, empathy supports important aspects of inter-personal communication to an extent where some psychic diseases that affect the relationship with other persons, such as autism and Asperger syndrome, are explained regarding impairment of empathic ability [82].

As reported in [51], there are two separate brain systems for empathy: an emotional (affective) system and a cognitive system. Emotional (affective) empathy involve several underlying processes such as emotional contagion, emotion recognition, shared the pain and be aware of the emotional feelings of another individual [51, 83], which leads to the way we understand other's mind as a "simulation theory". The cognitive empathy involves identifying and understanding what another individual is thinking and/or feeling without a necessary affective response [51, 83].

There are different empathy-arousing modes as identified by Hoffman [84], which includes mimicry, conditioning, direct association, verbally mediated association and perspective taking. These modes can operate alone or in combination. As reported by Hoffman [84], multiple modes allow us to respond empathically whatever the distress cues are present. For example, facial, vocal and postural cues are expressed through mimicry; situational cues through conditioning and association; distress expressed orally, in writing or by someone else. The manifestation of multiple modes not only enable us to instantly and automatically respond empathically with or without conscious awareness but also compel us to do that.

Emotion regulation:

The traces of emotion regulation study has been found from the study of Freud (1959), Lazarus (1966), Mischel et al. (1989) [85] and very recently there has been a dramatic increase in its theoretical and empirical advances.

Emotion regulation refers to "shaping which emotion one has when one has them, and how one experiences or expresses these emotions". [4, pp-6]. This regulation may be controlled or automatic, and it could be conscious or unconscious [85]. It is one of the essential skill for prosocial behavior. Moreover, it is also an important aspect of empathy because if one can understand and feel the emotions of others but not able to regulate those emotions by oneself, then it would be difficult for the person to empathize properly [86].

Emotion regulation is a complex process as it involves changes in the duration or intensity of behavioral, experiential and/or physiological responses [45]. Based on the *modal model* of emotion, Gross [45] proposed an emotion regulation framework, called *"process model of emotion regulation"*, which includes five families of emotion regulation processes such as situation selection, situation modification, attentional deployment, cognitive change and response modulation.

Eisenberg et al. [87] proposed a model focusing on the degree of emotion regulation, such as over, optimal and under, during a state of emotion arousal. The study of Lockwood et al. reports that the type of emotion regulation strategies such as cognitive reappraisal and expressive suppression also relates to the empathic response [88].

2.2.6 Study of Vocal Expressions for Emotion

Contemporary researchers have been trying to uncover the overt and covert signals associated with the emotional manifestations, which has been started mainly from the influential study of Darwin. He claims that "facial expressions are the residual actions of complete behavioral responses, and occur in combination with other bodily responses such as vocalizations, postures, gestures, skeletal muscle movements, and physiological responses" [89]. The study of emotional manifestations and the associated overt signals are mainly focused on two different traditions such as facial and vocal expressions. Research suggest that visual information may be more effective than auditory information, which might be the reason for the study of facial expressions and has been highly successful (see [44] and the references therein). Both channels, such as face and voice share some similarities as well as differences. The similarity includes, cues from both face and voice convey information to a perceiver to reliably differentiate discrete emotions such as anger and sadness. The differences are, 1) over a long distance communication or in dim light, the vocal expression may be more effective than facial expression, whereas facial expressions are effective in crowds of people i.e., cocktail party; 2) vocal expressions are relative and more time-dependent than visual expressions; 3) face is strongly related to visual arts and voice is strongly related to music.

This thesis mainly focused on utilizing vocal content, both verbal and non-verbal cues, therefore, below we highlight the studies that follow the tradition of vocal expressions.

Approach to investigate vocal communicaiton of emotion

Darwin's findings on the importance of vocal cues associated with emotional manifestations are largely consistent with the influential contemporary work by Scherer and his colleagues [2, 90–94]. To understand the process of the vocal communication of emotion, Scherer modified Brunswik's (1956) theory [92]. Brunswik model is based on the theory of perception, which considers biological function in a particular environment to a description of a psychological process that reflects and exploits that environment. Juslin and Scherer [44] report that this theory is suitable for three important reasons: 1) it is based on evolution perspective, 2) perceivers can make inferences about the objects or states of affairs based on a set of cues in the environment and it also important insights that can help us to explain the characteristics of the vocal channel, 3) the lens model of this theory is suited to study the vocal expression. For completeness, we present the modified version of Brunswik lens model from Scherer [92] in Figure 2.2.

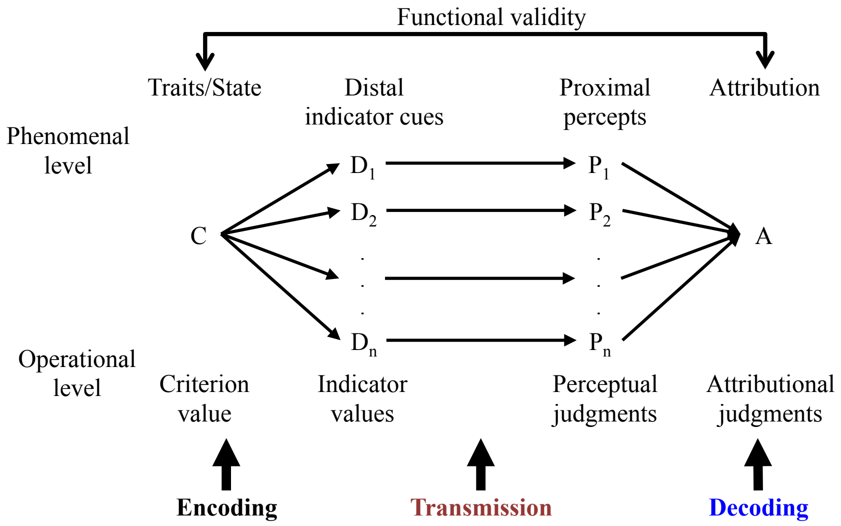


Figure 2.2: Modified version of Brunswik lens model from Scherer [3], which has been applied to the study of vocal expressions of affect.

This model include three distinguishing part: 1) *encoding* the long term traits e.g., personality and short term states e.g., emotion through distal cues e.g., vocal content from the sender side, 2) *transmission* in acoustic space and 3) *decoding* on the receiver side, which leads to the inference of the traits/states. From the research perspective, this model suggests each component of the complete process is important for the research.

Vocal cues

The studies of vocal cues have been investigated based on different theories of emotion such as discrete emotions, dimensional model, and componential model. From the discrete emotion perspective, one of the notable work done by Bense and Scherer [91], which investigated 14 emotions by 12 professional actors with 29 acoustic features. Their findings suggest that F0 and mean amplitude have the strongest association to the portrayed emotions. Using the acoustic features they report the statistical classification results of 40% in accuracy. A detailed review has been done by Juslin and Laukka [95] over hundred studies of vocal expressions associated with discrete emotions, where they report how different voice cues are correlated with emotional manifestations. For example, mean F0 is high for anger, fear, joy, surprise and stress. The studies vary in the way data has been collected such as portrayed, elicited or natural emotional expressions, which results in the variation in the results as well.

Steps to analyze vocal cues to affect

Juslin and Scherer [44] proposed a number of steps in order to conduct an experiment for vocal expressions of emotion, which are as follows:

1. Choosing the affect labels in terms of pragmatic and scientific considerations.
2. Obtaining a representative number of speech samples while considering the ways data has been collected. It includes portrayed, elicited/induced or natural emotional expressions.
3. Recording speech samples carefully as it is time-consuming and costly. The necessary issues that need to consider include designing the research setting; obtaining and training the necessary recording personnel; the additional time spent during each experiment in recording the information; the time spent storing and maintaining the recorded materials.
4. Segmenting speech samples for labeling. Two type of segmentation includes physical and perceptual. Physical segmentation means that boundaries are set based on the acoustic pattern of the speech samples, for example, silence periods. Perceptual segmentation refers that annotator set the boundaries of the segment based on some predefined criteria.

5. Finding the association of emotional manifestations and vocal cues, i.e., acoustic features, for example, F0, loudness, MFCC, and spectral related.

For our study, we followed every step and the only minor difference is that we could not maintain the order. The issue is that our data has been collected from real settings, which forced us to choose affective states after doing some preliminary analysis of the data.

Affect annotation using vocal cues

The typical approach of emotion inferences from vocal expression is to conduct judgment experiment, in which judges/annotators are asked to recognize the emotion expressed in the speech samples, using force-choice format, i.e., choosing one emotion label from a list of emotions. One of the important questions is that to what extent listener can infer emotion from the speech samples, which leads to labeling/annotating the data. Juslin and Scherer [44] review of 39 studies consisting of 60 listening experiments proved its significance. The other approaches to annotations include rate the speech samples in continuous scale or with a free description.

2.3 Affective Computing Research

The first investigation for speech emotion recognition was conducted in the mid-eighties using the statistical properties of certain acoustic features [96]. Since then, a considerable amount of work has been done on this domain. In spoken dialogue systems, people use emotion recognition module for better communication with the users. For example, in the projects "Prosody for dialogue systems" and "SmartKom", ticket reservation systems are developed that employ automatic emotion recognition module to be able to recognize the annoyance or frustration of a user and change their response accordingly [97]. Similar approaches have been applied in call center applications [98, 99]. Therapists also employed the emotion recognition system as a diagnostic tool in medicine. To extract real-time emotional characteristics of speech, emotional speech recognition methods have been used [100]. A notable overview of emotion recognition from speech has been conducted in [101], which includes an up-to-date record of the available data; most frequent acoustic features and classification techniques. The research community of this field has also made an evaluation campaign in the Interspeech 2009 emotion challenge [102] to recognize emotion by designing a benchmark dataset.

A Recent survey of Weninger et al. [8], reports the important challenges that needed to overcome in order to design real applications. These include collecting real data, solving data imbalance problem, issues of generalization across application scenarios and languages, requirements of real-time and incremental processing, dealing with acoustic variability, and evaluation methods.

2.3.1 Available Corpora

Notable overviews of emotional corpora can be found in [101, 103, 104], which varies based on modality, language, emotion elicitation approaches, categorical vs. dimensional, a number of subjects, domains, and purpose of the task, i.e., recognition vs. synthesis. The work in [103,104], a part of HUMAINE project, listed three different kinds of corpora focusing on modalities. It includes multi-modal, speech-only and facial expressions based corpora. Each review categorized corpora based on different characteristics such as availability i.e., public or copyrighted, type of emotion elicitation i.e., acted, induced or natural, the number of subjects, language, type of speech i.e., spontaneous, scripted, or nonsense. Ververidis et al. [101] reviewed 64 corpora by focusing on emotion elicitation approach, the content, and modality. They also report most frequent features and classification algorithms commonly used for designing automatic system.

Here we report a few of the corpora focusing on the emotion elicitation such as *acted*, *induced* and *natural* as the performance of an automatic system also varies a lot depending on the emotion elicitation approaches.

Corpora of Acted Emotion

The most widely used public available corpora include Danish Emotional Speech (DES) [105], Berlin Database of Emotional Speech (EMO-DB) [106], and Speech Under Simulated and Actual Stress (SUSAS) [107] in which actors manifested emotional episodes. In DES corpus, emotion categories include anger, happy, neutral sadness and surprise. Four professional Danish speakers (two male and two female) acted on two single words, nine sentences and two passages of fluent speech. EMO-DB contains 500 utterances spoken by actors in which different categories of emotions include anger, boredom, disgust, fear, happiness, neutral and sadness. Ten emotionally undefined German sentences have been acted in these emotions by ten professional actors, in which five of them are female. The whole set comprises around 800 utterances and out of them only 494 phrases are commonly used for the experiment. The database is recorded in 16 bit and 16 kHz under studio noise conditions. The SUSAS corpus is a spontaneous corpus recorded by a total of 32 speakers (13 female and 19 male) with recordings of 16000 utterances. The corpus includes five stress domains, in which the content include 35 air-commands manifested in different speaker's states such as high stress, medium stress, neutral and scream.

Corpora of Induced Emotion

The corpora of induced emotion include eNTERFACE [108], Airplane Behavior Corpus (ABC) [109], FAU-Aibo robot corpus [110]. The eNTERFACE corpus consists of 42 subjects (81% male and 19% female), from 14 different nationalities. In order to induce emotion, participants were asked to listen to six successive short stories, each of them inducing a particular emotion. Participants were then reacted to each situation, and two human experts judged whether the reaction contains an emotional expression or not. Finally, the annotated corpus contains a total of 1166 video sequences with six emotion categories such as anger, disgust, fear, happiness, sadness and surprise. The ABC is a mood annotated corpus consist of thirteen and ten scenes as a start, serving of wrong food, turbulences, falling asleep, conversation with a neighbor, or touch-down. The annotated emotion categories include aggressive, cheerful, intoxicated, nervous, neutral and tired. We report FAU-Aibo robot corpus in Section 3.2, which is non-prototypical and available with a license agreement.

Corpora of Natural Emotion

Compared to the corpora, which are acted and induced there are very few corpora with natural emotional expressions. To name a few are Belfast naturalistic corpus [111], Geneva airport lost luggage corpus [112] and two other corpora reported in [113]. In Belfast naturalistic corpus, clips are taken from television chat shows, current affairs programs, and interviews. It consists of 239 clips (209 from TV recordings and 30 from interview recordings) from 125 subjects (31 male and 94 female). The corpus includes both dimensional and categorical emotional labels. The Geneva airport lost luggage corpus consists of one hundred twelve airline passengers' video recordings, which has been recorded after their reporting to the baggage retrieval service about their lost luggage in a major international airport. The study mainly focused on a particular focus of appraisal theory and emotion categories include anger, good humor, indifference, stress and sadness. The corpora reported in [113] include real agent-client recordings obtained from a Web-based Stock Exchange Customer Service Center, and from a conversation between a medical emergency call center and the LIMSI-CNRS. The former corpus contains about 6200 speaker turns, and annotation includes 5000 non-overlapping speaker turns by listening to an audio signal in the dialog context. Emotion categories include anger, fear, satisfaction, excuse and neutral. The later corpus has been annotated with 21 fine-grained emotion labels plus the three labels (positive, negative, unknown). The fine-grained emotion labels were grouped into seven coarse-grained labels such as fear, anger, sadness, hurt, positive, surprise and neutral.

Scenarios of Emotion Elicitation

For different application scenarios, typically corpora have been collected by focusing on various interaction type such as human-human and human-machine. For both interaction types, scenarios can be designed in a laboratory or in a real-life setting. Corpora of human-human interaction include Audiovisual Interest Corpus (AVIC) [114], and the corpora reported in [113]. Most of the corpora are mainly designed by focusing on human-machine interaction scenarios. In the case human-machine settings, the scenarios are designed using Wizard-of-Oz, in which a human operator play the role of the machine. Examples of human-machine corpora include, The Vera-Am-Mittag Corpus [115], the ICSI meeting corpus [116], and SmartKom [117].

2.3.2 Approach to Annotation/Labeling

After the data collection, annotation/labeling[2] has been done by self-report, expert or naive crowd annotators by listening/observing the emotional data in the form of audio and/or visual information. Since this approach leads to uncertainty in the ground truth, therefore multiple annotators are commonly employed. An estimate is then derived by majority voting for categorical annotation, and average (mean) for continuous valued annotation. In order to measure the reliability of such estimate inter-annotator agreement[3] has been calculated. For describing emotions either dimensional and/or categorical approach, annotations have been done based the specification of the task/scenarios, each of which has an underlying theory of

[2] annotation or labeling is used interchangeably throughout this thesis, which means assigning a label to an emotional event, for example.

[3] also called inter-rater or inter-labeller agreement

emotion.

Dimensional labeling The dimensional approach to annotations includes a continuous range of values, in which annotator assign a value to an utterance or speech segment for a particular dimension of emotion based on their perceptual judgment. A two-dimensional model is mainly popular in which dimensions are 'evaluation' such as positive/negative or pleasant/unpleasant and 'activation' such as high/low or active/passive. Several tools have been developed for this kind of labeling such as FeelTrace [118] and Geneva Emotion Wheel [119]. The automatic system is usually designed to predict the continuous value using regression based techniques.

Categorical labeling In the categorical approach, a set of emotion categories is predefined based on the specification of the task and the annotators assign a category to the spoken utterance from the predefined emotion list. The theory of categorical model of emotion is defined based on the use of everyday language, which theorists believe is the easiest way [120], even though there are a lot of disagreement in defining a number of categories. It ranges from three to twenty one. For designing the automatic system, either generative or discriminative approach of classification has been used.

Time dependent labeling In either categorical or dimensional labeling, the annotators are usually given an utterance or speech segment to annotate by assuming that emotional state is consistent throughout the segment. This assumption may be true for acted data as the actor or actress are enacting the emotion from a predefined set for an utterance. This is different for induced or naturally occurring emotion in a conversational interaction, in which boundary of the emotional segment is not known. This is an ideal situation in any real application where subject's emotional state may change over time, leads to generate a sequence of emotional states. Finding the emotional segment is a challenging task for the annotators as they need to decide the boundary of the segment as well as the emotional label. For solving this problem, in FAU Aibo Emotion Corpus [28], audio is broken down into word level by assuming that the emotional state would be constant at this level.

For the annotation of emotional manifestations, in our work, the annotator defined the boundary of the emotional segment based on the annotation guideline and their perceptual judgment (see Section 3.1.1).

Measures of inter-annotator agreement The inter-annotator agreement is a measure of the degree of agreement among the annotators or how well they can agree on the same annotation. It provides information about how difficult the annotation task was and also the appropriateness of the labels. To calculate the reliability of the inter-annotator, the common methods include Cohen's Kappa [121] for categorical annotations and weighted Kappa or Spearman's rank correlation for continuous valued annotations. Some other methods for reliability measures include Cronbach's alpha or Krippendroff's alpha, which can deal with cases when there is an annotation missing by an annotator, for an example.

2.3.3 Automatic Classification

A typical emotion recognition consist of three important parts such as preprocessing, feature extraction and classification as shown in Figure 2.3 [16]. An optional part include feature selection and generation as presented with dotted box in Figure 2.3. The preprocessing part includes filtering, and segmenting or chunking data into a meaningful unit. Feature extraction part includes extracting relevant features. Followed by feature selection module, which is an optional part. Finally, training the classification/regression model, which maps the training examples associated with class labels/scores. More details of each part are presented below.

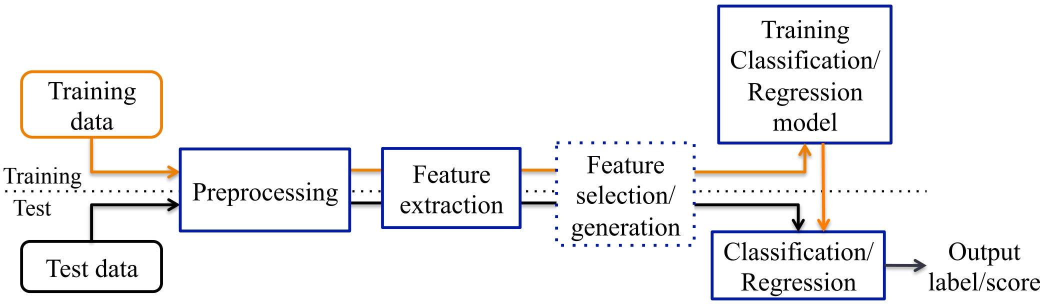

Training data
Training
Test
Test data
Preprocessing
Feature extraction
Feature selection/ generation
Training Classification/ Regression model
Classification/ Regression
Output label/score

Figure 2.3: Typical emotion recognition system. Dotted box represents optional part.

Preprocessing

For spoken conversation, the preprocessing part includes speech enhancement, source separation removing noise and a long silence, de-reverberation, and segmentation. For most of them, one can rely on current state-of-art speech processing techniques. The important issue is that emotion classification system needs to deal with segmenting the conversation into an emotionally meaningful unit. Finding the meaningful emotional unit has several important advantages such as it improves the classification performance and it is also important for incremental processing in real-time application [122]. For acted emotion, the current approach is utterance or word as an emotional unit as utterances or words are predefined. For natural or spontaneous speech it is much more difficult. Typically turn is considered as an emotional unit, which is obtained using voice-activity detector or speech *vs* non-speech segmenter. In [123] Vogt et al. compared the classification performance with different units such as utterances, words, words in context, and fixed time interval. Their findings suggest that larger and linguistically motivated units tend to perform better in automatic classification. In [122], Batliner et al. compared words, chunk and they report that *syntactic chunk* and *emotion/ememe chunk* could be the most promising unit. From the perspective of Vogt et al. [16], an emotional unit should fulfill two requirements such as

1) long enough to reliably calculate features regarding statistical functionals, 2) short enough that represent stable acoustic properties. However, it is yet to be discovered how long and short the unit should be.

Researchers have also been investigated several approaches to segment emotional episodes such as 1) phoneme as a segment [93, 124], 2) segment based on voiced speech [125], 3) forced aligned word as segment, 4) syntactic chunks, 5) ememe chunks [122], 6) utterances based and 7) regions-of-interest. In [125] Shami et al. worked on segmenting emotions using global statistical features, where they used KISMET emotional corpus and compared the performance of SVM and K-NN by designing the speaker dependent classifiers. They show that classification accuracies increase 5% when segment-level decisions are integrated compared to the utterance-level feature generation approach. In [126], they used a segmentation method to extract a sequence of voice segments and then recognize each segment individually, using acted emotional dataset. Batliner at el. in [122], put a considerable effort to understand and define the emotion units based on speech and coined the term “ememe" to define the smallest unit of emotion, where a “word” is considered as a smallest meaningful emotional unit.

Feature Extraction

The second part of the automatic classification system is to extract relevant features from spoken conversation in term of acoustic and linguistic information. The aim is to represent them in n-dimensional feature vector for classification. Depending on the availability of the resources both acoustic and linguistic features has been investigated in the current state-of-the-art as detailed below.

Acoustic features Most common approach is to extract low-level acoustic features at the first step, then project them onto statistical functionals to have an equal sized feature vector for each training examples. Other approaches include extracting features frame-by-frame followed by HMM based modeling.

Typically, acoustic features are extracted at approximately 100 frames per second and a window size varies from 10 to 30 ms. The windowing functions include Hamming or Hanning window for frequency domain and rectangular for time domain. The low-level acoustic features include pitch, intensity, Mel Frequency Cepstral Coefficient (MFCC), Linear Prediction Cepstral Coefficients (LPCCs), Perceptual Linear Prediction (PLP), formants, spectrum, harmonicity, noise-to-harmonics ratio, duration based features and perturbation such as jitter and shimmer [7]. Most often derivatives, smoothing, and normalization are also applied on the low-level features.

Mostly, the length of the segment or conversation is not fixed, therefore, the size of the feature vector varies. Hence, the common approach is to use statistical functionals, and its purpose is to map the variable-length input sequences to a fixed-size vector. The statistical functionals include mean, standard deviation, extremes, percentiles, peaks, higher order moments, segment level statistics, and temporal information. Commonly used tools for extracting these features include praat [127], openSMILE [128], Hidden Markov Model Toolkit (HTK) [129].

Linguistic features To extract linguistic features typical approach is to transcribe the audio into text, which has been done either manually or automatically. To obtain automatic transcription, one needs to rely on Automatic Speech Recognition (ASR). For extracting feature from transcriptions most common approach is Bag-of-Words, also known as vector space model. It is a numeric representation of text that has been mainly used in text categorization [130]. For this model, first, vocabulary is designed using the word in the training dataset. Then, each word in the vocabulary represents an element in the feature vector in the form of frequency count for a training example.

Since this approach results in a large feature vector, therefore, stop word removal and/or low-frequency word removal are also applied. Other than term frequency, a most common technique is called TF-IDF, Term Frequency (TF) multiplied by Inverse Document Frequency (IDF), is also commonly used.

Feature Selection

Feature selection or dimensionality reduction is one of the steps in order to improve the classification performance and reduce the computational cost. Higher dimensional feature vector introduces a curse of dimensionality. Therefore, we need to select relevant feature and remove irrelevant and redundant features. Relevant features are the features that have an influence on the correct recognition and have a higher correlation with the classes. Irrelevant features are those that do not have any influence on the correct output and the redundant features carry the same information as of the other features. Two main approaches to the feature reduction are 1) feature selection, 2) feature transformation. Feature selection techniques can be divided into feature ranking and subset selection. In feature ranking approach, features are ranked based on some criteria and the features above a certain threshold are selected. Feature ranking algorithms that use in this domain are information gain [131], gain ratio [132]. Subset selection approach is divided into two parts: 1) *embedded approach:* the feature selection has been done as a part of the classification algorithm, 2) *wrapper approach:* it utilizes a classifier as a black box to score the subsets of features based on their predictive power. Feature subset selection algorithms that are commonly used in this domain are wrapper-based approach with sequential forward search (SFS) [133], sequential floating forward search (SFFS) [1] or genetic algorithms [134] and correlation based feature selection (CFS) with subset feature selection algorithm.

The second main approach to reducing dimension is the feature transformation or feature extraction, where higher dimensional feature spaces are mapped into lower dimensional spaces by keeping as much information as possible. Principal component analysis (PCA) and linear discriminant analysis (LDA), Heteroscedastic LDA are the two well-known techniques [38]. Other algorithms include Independent Component Analysis (ICA), which maximizes the statistical independence of the features, and Non-negative Matrix Factorization (NMF), which assumes non-negative data and additive features. A recent trend is to integrate feature engineering and classification in the deep neural network framework [8].

Classification

There are many issues that need to be considered while choosing diverse machine learning algorithms, for example, small and sparse data, high dimensionality, non-linear problems and higher generalization. There are many supervised classification algorithms widely used for emotion recognition [38, 135]. The algorithms that performed best in the Interspeech 2009 emotion challenge are described here. The discriminative learning algorithm - Support Vector Machines (SVMs) [136] has the ability to solve the non-linear problem by kernelized transformation of the feature space and has a better generalization capability. Decision trees and Artificial Neural Networks (ANNs) [133] are other two non-linear discriminative algorithms used in these domains. Random Forests (RFs) [137] becomes popular and it is one of the algorithms. RFs are a combination of tree predictors and build a series of classification trees, and each tree is generated with random input features (subspace of features), and random samples are taken by sampling with replacement. Each tree on its own makes a prediction. RFs use these predictions by majority voting to make the final prediction. Adaboost [138] is another ensemble learning algorithm that performed best in the Interspeech 2012 speaker traits challenge. It constructs a strong classifier as a linear combination of weak classifiers by overweighting the examples that are misclassified by each classifier. The predictions from all of the classifiers are then combined through a weighted majority vote to produce the final prediction. Other popular algorithms are Hidden Markov Models, Gaussian Mixture Models (GMMs), K-Nearest neighbors, Dynamic Bayesian Networks, Deep Neural Network (DNN) and fusion of different heterogeneous classifiers [139, 140] by majority-vote or stacking.

The state-of-art performance of the acted emotion corpora is very high as also reported in [141], where a benchmark comparison of performances has been done. This study reports that best performance on acted or prototypical emotions, e.g., EMO-DB and eNTERFACE, whereas the performance of corpora with non-acted or induced emotions are lower.

Evaluation Methods

Two different kinds of evaluations are used in these domains [133]. The first method is on the basis of accuracy, precision, recall and F1 measure of the system, which are widely used evaluation matrix in information extraction. The other approach is to calculate the weighted average (WA), i.e., accuracy and un-weighted average (UA), i.e., average of class label recall. Receiver Operating Characteristic (ROC) measure is also used in some studies.

2.4 Research on Affective Scene

In the past few decades, there have been efforts to design and develop methodologies for affective interaction. Methodologies based on behavior observation protocols can identify recurrent categories of emotional exchange by utilizing data that is generated by listening to recorded conversations and then manually coding the relevant emotional transitions. However, it is well known that such methods are widely acceptable but highly time-consuming. In the affective computing literature, there is evidence of behavior analysis methods in domains where traditional observational protocols can be applied. In [142], the authors proposed an approach to automate a manual human behavior coding technique for couple therapy. They use acoustic features in order to classify the occurrences of basic and complex emotional attitudes of the subjects such as sadness, blame, and acceptance.

Focusing on a sequential organization is important when studying emotion in a real setting, as reported by Goodwin & Goodwin [143] in the field of conversation analysis. Their results show that powerful emotional statements can be built by the use of sequential positions, resources provided by the environment where the action occurs, and *artful orchestration* of a range of embodied actions such as intonation, gesture, and timing. In [9], Lee et al. reported that modeling the conditional dependency between two interlocutors' emotional states in sequence improves the automatic classification performance where they used a corpus in which actors manifested emotional episodes.

2.5 Summary

In this chapter, we presented the studies of *affective behavior* both in psychological and affective computing perspectives. Psychological studies are mainly concerned with theorizing and modeling emotion and there are different tradition such as discrete and dimensional approach. Among many theories of emotion, the appraisal theory is becoming popular, which defines how organisms evaluate events/situations based on the appraisal process. Following the appraisal theory, the modal model of emotion is defined, which considers three core features of emotion. We also presented the research on empathy and emotion regulation. Since we are mainly interested in understanding how vocal cues are good predictors for the recognition of emotion, therefore, we presented the reviews of literature related to the vocal expressions of emotion. From the affective computing perspective, we reviewed the literature that is related to data collection, annotation, feature extraction, classification, and evaluation. Finally, we discussed the studies of the affective scene.

Chapter 3

Datasets for Affective Behavior

The aim of our study is to design an automatic computational models for the classification of affective behavior from spoken conversations recorded in real-life situations. In particular we analyzed call-center applications, where both agent and customer engage in interaction to solve problems or to seek information. In the current state-of-the-art, there are only a very few corpora, which have been collected in real-life situations. Our Signals and Interactive Systems Laboratory (SISL) affective corpus has been collected from Italian call-centers with real-users that were engaged in real conversation with the agents. This corpus has been annotated with respects to the empathy of the agent, and to basic and complex emotional states of the customer. The corpus includes many unique characteristics such as annotation of empathy and representation of emotional sequence throughout the conversation. In addition to the SISL affective corpus, we have also exploited FAU Aibo robot corpus for the segmentation and classification of the emotional states.

3.1 SISL Affective Behavior Corpus

The SISL affective behavior corpus has been collected from real call centers, where customers are calling to solve a problem or for seeking information. Since the goal of the experiment was to analyze real-life emotional manifestations, therefore no prior knowledge has been given to the subjects during the data collection. The inbound Italian phone conversations between call center agents and customers are recorded on two separate audio channels with a quality of 16 bits, 8kHz sample rate. The length of the conversation, in seconds, is average ± std is 396.6 ± 197.9 (all 10K conversations)[1].

[1] The original dataset contains 10063 conversations where average ± std is 395.9 ± 198.2 and later some of them has been discarded

3.1.1 Annotation of Affective Behavior

For develpoing the annotation scheme used in our research task, we followed the *modal* model [4] of emotion and choose to annotate discrete (categorical) *affective* states. We use the term *affective* and *emotional* synonymously to represent the same meaning throughout the thesis. During the annotation process annotators assign an emotional state to an speech segment. Annotators choose a label from a label set $C = \{c_1, c_2, ..., c_n\}$. The term *label*, *class label*, or *tag* refers to the emotional state that assigns to the speech segment and the process is interchangeably called as *labeling*, *tagging* or *annotating* for the study.

Since our data has been collected from natural settings, we have done a preliminary analysis before choosing the affective states. From our analysis, we observed that in the call center scenario, the customer can evoke emotional arousal that can vary from disappointment to satisfaction. The customers usually inquire to the call center for some of their urgent problems to be solved, and most of the time the problems are related to costs of the service or questions related to contractual change or activations. We manually investigated some sample data, which are task oriented or knowledge transfer conversation.

For capturing those nuances of the emotional attitude we initially considered four emotional labels such as anger, frustration, satisfaction and dissatisfaction. From our initial investigation, we observed that for the satisfaction and dissatisfaction in many cases the manifestation of customer emotional expressions are partial. For these two cases, we also included two more affective labels such as not-complete satisfaction and not-complete dissatisfaction.

Our preliminary analysis of the agent's side conversations reveals that agent was trying to compassionately soothe customer's emotional state, which leads to a satisfactory conversation. There are also the cases that agent was neutral where the customer was manifesting negative emotions. Hence, we choose to annotate *empathy* label only on the agent side.

As a results, in our study, the annotations include *empathy*, basic emotion such as *anger* and complex emotions such as frustration, satisfaction, non-complete satisfaction, dissatisfaction and not-complete dissatisfaction. In our annotations, we also have an implicit label, such as neutral. Neutral is considered by assuming that there are portions of the conversation that are not characterized by a particular emotional attitude. For example, it is likely that the dialogue turns where the agent collects the identification details of the customer, which might not be relevant from the point of view of emotional attitude.

Annotation Guidelines

We designed the annotation guidelines by following the *modal* model [4] of emotion. Gross's modal model is based on appraisal theory, which has been studied by many psychologists for the investigation of emotional episodes from the point of view of appraisal dimensions. Gross [4] has provided evidence that concepts such as *emergence* — derivation from the expectations of relationships — and *unfolding* — sequences that persist over time — may help in explaining emotional events. It has been shown that temporal unfolding of emotions can be conceptualized and experimentally tested [144]. The modal model of emotions developed by Gross [4, 45] emphasizes the attentional and appraisal acts underlying the emotion-arousing process. In Figure 3.1, we provide the original schema of Gross model. The individuals' core *Attention-Appraisal* processes (included in the box) are affected by the *Situation* that is defined objectively in terms of physical or virtual spaces and objects. The **Situation** compels the **Attention** of the individual; it triggers an **Appraisal** process and gives rise to coordinated and malleable **Responses**. It is important to note that this model is dynamic, and the situation may be modified (directed arc from the **Response** to the **Situation**) by the actual value of the **Response** generated by the *Attention-Appraisal* process.

The annotator needed to identify the emergence, appraisal, and unfolding of emotions felt by the call center agent in the ongoing (sub-)dialogs. In doing so, the annotator also needed to be able to perceive if, and to what extent, an emotional, e.g., empathic, response may modify a situation where other emotions such as frustration or anger are expressed by the customer.

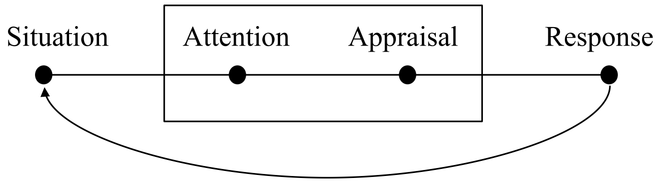


Figure 3.1: The modal model of emotion [4].

Our annotation guidelines aim at a continuous perception of the *variation* in the speech characteristics for a limited context support. The goals that motivate the annotation guidelines are:

- The ability to annotate and enable cause-type interpretations of emotion manifestations. We aim at augmenting the traditional label-type annotation with a context that may contain the necessary information to interpret or anticipate the emotion-filled event [94].
- The efficiency of the selection task of annotators. Rather than having a continuous tagging task over complex human conversations, we aim at focusing the attention of observers on variations and detections.

Such goals motivated the following annotation guidelines:

1. Annotating the onset of the signal variations that supports the perception of the manifestation of emotions.
2. Identifying the speech segments preceding and following the onset position.

3. Evaluating the communicative situation in terms of appraisal of the transition from a neutral emotional state to an emotionally connoted one.
4. Annotating the context (left of the onset) and target (right of the onset) segment with an emotion label (e.g., empathy, anger).

For the annotation task, we recommended the annotators to focus on their perception of speech variations, in particular on the acoustic and prosodic quality of the pairs of speech segments, while minimizing any effort to pay attention to the semantic content of the utterances. They were asked to judge the relevance of the perceived variations on the expression of emotions. Gross' model provides a useful framework for describing the dynamics of emotions within an *affective scene*([145]), because not only does it focus on appraisal, but also on the process feeding back to the initial situation.

Annotation Unit

The annotation unit is the stimulus presented to the observer (annotator) to perform a selection task over a decision space such as the set of emotional labels. In general, the annotator may be presented with images or speech segments (stimuli), and a set of emotional labels. The stimulus is defined in terms of the medium the emotion is being *transmitted* through and its content and context. The medium may be speech [146], image [147] or multimodal [148]. The content refers to the information encoded in the stimulus signal such as facial expression of anger or a speech utterance. The context of the stimulus is represented by the spatial or temporal signals neighboring the stimulus. Knowledge of the context may be crucial in interpreting the cause of emotion manifestations. For a speech stimulus, the context is represented by the preceding dialog turns [146].

Most research in affective computing has been limited to stimuli that are designed in advance and are artificially generated. Respective examples of such stimuli are sentences to be read and actors enacting affective scenes [94].

Another limitation of previous annotation tasks is that stimuli are *context-free*, and speech utterances or images are annotated in isolation.

The limitations of determining the temporal boundaries of the annotation units have motivated researchers to investigate the process of continuous annotation [149, 150]. Yet, state-of-the-art complete continuous affective scene annotation techniques are highly demanding for observers and are not very effective in terms of inter-annotator agreement.

In our work, we address the complex task of defining and searching the annotation unit in real-life spoken conversations. Annotation unit is defined by the segment, in which identification of the emotional segment boundaries is set by the annotator's perceptual judgment. This manual annotation of the emotional segment may consists of one or more turn(s) in a conversation. It has later been mapped into automatic segment during the evaluation of the classification experiments as discussed in Chapter 5.

Operational Definition of Affective Labels

Following the modal model of emotions and the annotation guidelines discussed above, we first describe the context of the situation, the attention, the appraisal and the response components of the emotional manifestations.

The context of the situation: In call center conversations, customers may call to ask for information or for help to resolve technical or accounting issues. Agents are supposed to be cooperative and empathic, and they are trained for the task. However, variabilities in the agents' or customers' personalities, behavior, and random events lead to statistical variations in the emotional unfolding of affective scenes. The operational definition of empathy requires that annotators should be informed of the social context and task. They are trained to focus over sub-dialogues where the agent anticipates solutions and clarifications (**attention**), based on the understanding of the customer's problem (**appraisal**). As a consequence, the acts of the agent may relieve or prevent customer's unpleasant feelings (**response**).

The selection of the stimuli to be assigned include a continuous search of the speech segments preceding and following the perception of the *onset* of the emotional event. The task of the annotator is to identify the context (left of the onset) and target (right of the onset) emotional segment. The context of the onset is defined to be *neutral* with respect to the target emotional segment. We have required to annotate the neutral segment to support annotators in their perception process while identifying the segment of the emotional manifestation. Instructions were given to the annotators to achieve the most confident decision in identifying and tagging the neutral and emotional segment pairs, based on the perceived paralinguistic and/or linguistic cues. The set of emotional categories that we have chosen for our study are operationally defined below.

Empathy (Emp): We operationally defined *empathy* as *"a situation where an agent anticipates or views solutions and clarifications, based on the understanding of a customer's problem or issue, that can relieve or prevent the customer's unpleasant feelings"*. It relates to the ability to recognize the affective or cognitive states of mind of the others, and to react accordingly. The affective kind of empathy implies the capacity of recognizing the emotions that other human or fictional beings are experiences in virtue of our ability to being affected by that emotional state. The cognitive kind of empathy is related to the capacity to understand the other's perspective on a given issue. At present, we will only use one label for the two kinds of empathy.

Since empathy is both a personal attitude of the speaker and a cognitive strategy adopted for achieving good levels of cooperation in the conversational interaction, it is important that the annotator can be able to identify the moments in the conversation when an empathic reaction is occurring due to something that is happening in the conversation. This can often anticipate some solutions, proposals, and clarifications that can relief or prevent other speaker's unpleasant feelings. A few examples with a manual transcription of the empathic annotation are given in Table 3.1.

Table 3.1: Few examples of the **empathic** manifestations with manual transcription. C - Customer, A - Agent

Dialog excerpt	Notes
C: Ascolti ... io ho una fattura scaduta di 833 euro vorrei sapere ... tempo in cui posso pagarla. *(Listen... I have an 833 euros expired bill... I would like to know... the time left to pay it.)* **A:** Ma perché non ha chiesto il rateizzo di questa fattura? Proviamo a far il rateizzo, ok? Così gliela blocco e lei ha più tempo per effettuare il pagamento. *(Why did not you ask to pay it in installments? We try to divide it into installments, is it ok for you? So I stop the overdue notices and you will have more time to pay.)*	The tone of voice and the hesitations of the customer show that she is not angry, she is ashamed for not being able to pay immediately the bill. This causes an empathic reply in the Operator's attitude. The choice of the speech act (question instead of authoritative declarative), the rhetorical structure of the second question, the lexical choice of "proviamo", instead of - for instance, "adesso provo a vedere...", all these contribute to prevent the customer's feeling of being inadequate.
A: Vede, questo è un passaggio: la bolletta del vecchio gestore e la nuova sono molto vicine ... si sono trovate accavallate ... vediamo ... *(See, this is a change of provider: the final bill of the last provider and the first of the new one are very close.. almost gathered... let's see...)*	The customer shows delusion: he changed energy provider because he thought the new one was more convenient, but now he has to pay two bills in a short period. The Agent understands the emotional state of the customer.

Anger (Ang): Anger is usually described as *the emotion related to one's psychological interpretation of having been offended, defamed, or denied and a tendency to react through retaliation.* We annotate the emergence of anger when the perception of the speech signal shows typical anger voice traits, such as tension, hyper-articulation (cold anger), and raised voice. Some examples of events with the manifestations of anger are when customer threats the agent of taking some legal actions with respect to the company, when s/he uses offensive words. In Table 3.2 we present a few examples of utterance with the manifestations of anger.

Table 3.2: Few examples of the manifestation of **anger** with manual transcription. C - Customer

Dialog excerpt	Notes
C: a parte che che che chiederò un risarcimento danni perché non è possibile perché io ho fatto i fax come mai questi fax non arrivavano io qui voglio una spiegazione metterò l avvocato e mi farò risarcire tutti i danni. *(except that... that ...that I will ask refund for damage because is impossible, cause I sent fax, why didn't arrive? I pretend an explanation I 'put' the lawyer and get compensate for damage)*	The customer accuses the agent and the company to be responsible for his damage, he pretends to get an explanation and be compensated.
C: no ma devono ripristinare il servizio oggi se no io chiamo il servizio consumatori e vi denuncio cioè il punto è quello il punto allora non me ne frega niente *(No, they must reactivate the service today, if not I call customer service and denounce you that is, the point is that, that's the point than, I don't give a damn)*	Although the agent explains to the customer it will take time to reactivate power, however, the customer is alarming to denounce the agent and the company if they don't solve the problem immediately. The customer shows hot anger, with raised voice, and reapeating the same words many times.

Frustration (Fru): Frustration is operationally defined as *'a complex emotional state that arises from the perceived resistance to the fulfillment of individual will or needs'*. For example, in our application domain it is likely that the customer is experiencing frustration in scenarios such as when s/he has to call back many times to solve the same problem, when s/he needs to call back for the issues that have been misunderstood before in the conversation, and when the agent cannot solve the issues.

Table 3.3: Few examples of the manifestation of **frustration** with manual transcription. C - Customer. C - Customer, A - Agent

Dialog excerpt	Notes
C: Noi ... è che qui ci manca la corrente... *(We... here electricity runs out of power)* **A:** Deve contattare la segnalazione guasti. *(You must call damage signal number)* **C:** Noi non sappiamo ... noi non sappiamo come fare[..] è tutta la mattina che proviamo. *(We don't know... we do not know how to do it, we have been trying to call all this morning.)*	Nevertheless the agent is able to offer a response to the customer's problem, the customer is frustrated because he probably had the hope that in this (repeated) call he could have the opportunity to solve the problem immediately.
C: Dopo tante telefonate che non vi siete fatti sentire ... due bollette nello stesso giorno! *(After so many calls you do not answer... two bills in the same day!)*	Customer's frustration is showed by the intonation and sense of delusion expressed in the last part of the turn.

Satisfaction (Sat): We operationally defined *satisfaction* as *'a state of mind deriving from the fulfilment of individual will or needs'*. It is likely that the issues dealt within the task-oriented dialogs of our application domain are solved when the customer is satisfied. It is a state of mind deriving from the fulfillment of individual will or needs. This implies that customer received detailed information, and agent understood her/his problem very well. An example is given in Table 3.4.

Table 3.4: An example of the manifestation of **satisfaction** with manual transcription. C - Customer

Dialog excerpt	Notes
C: ah va benissimo okay questa qua la posso già buttare ah okay grazie mille gentilissima arrivederci arrivederci. *(Very well, okay, this (bill) I can throw it away, okay many thanks, really kind of you, goodbye goodbye)*	The customer obtain final confirmation from the agent to his question, he appreciates the agent's kindness.
C: ah allora vabbè ancora il tempo c è va bene okay allora per il momento è tutto tranquillo okay va bene grazie buongiorno *(Alright, still have time, alright, okay at the moment everything is 'quiet', alright thank you have a nice day)*	The agent relieved the customer of his worries, giving him the information he needed – and expected.

Table 3.5: An example of the manifestation of **not-complete satisfaction** with manual transcription. C - Customer

Dialog excerpt	Notes
C: grazie, arrivederci buonasera. *(Thank you, see you again. Have a good evening)*	The customer put emphasis on his words.
C: Va benissimo, okay, la ringrazio, le auguro una buonagiornata *(Al right, okay, thank you. Have a goodday)*	The customer's kindness could be not directly related to the satisfaction about the interaction.

Not complete satisfaction (Ncs): Sometimes the criteria stated for labeling the speech segment with Sat may be only partially met. While the annotator is confident that no other emotion label can be perceived other than Sat, however, the annotator may still be uncertain. Uncertainty can be due to the fact that, for example, the speech segment seems to be too short, or because Sat is expressed as an appreciation toward the company or the service or other general motivations, and not specifically with respect the ongoing interaction. In these cases we label the segment as not-complete satisfaction as shown an example in Table 3.5.

Dissatisfaction (Dis): The dissatisfaction state of mind does not show a cognitive attitude towards understanding the validity of other's reasons, which differs from the frustrated subject. This appraisal is usually present when the subject is frustrated. While frustration suggests the intention to solve the problem within the dialog, dissatisfaction signals a general closure and hopeless attitude in the conversation, and more in general toward the chance to be understood, and helped, by the agent. The emotional attitude arises from the perceived impossibility to be understood or helped by the agent. The emotion is persistent. We define dissatisfaction as the emotional attitude opposed to satisfaction. It is defined as *'implying some disappointment of individual expectancies'*. A clear idea can be found the examples give in Table 3.6.

Not complete dissatisfaction (Ncd): When the criteria stated for annotating speech segment with Dis is partially met, and no other label is perceived other than Dis, then we define that as not-complete dissatisfaction. It is a similar case of not-complete satisfaction. An example is presented in Table 3.7.

Neutral (Neu): The neutral emotional attitude was not annotated. However, it was assumed that the speech segment before any given annotated emotion is neutral with respect to that emotion. The context of the onset was defined to be neutral with respect to the target emotional label. We have introduced the neutrality as a relative concept to support annotators in their perception process of the target emotional state while identifying the situation context support.

Table 3.6: A few examples of the manifestation of **dissatisfaction** with manual transcription. C - Customer, A - Agent

Dialog excerpt	Notes
C: è una cosa cioè che noi non riusciamo a parlare con un amministrazione è assurdo va bene grazie lo stesso. *(It's a matter.. that we can't talk to an administration office, is an absurd! Alright thanks anyway)*	The customer tried many times to get more information about his service but once again he can't obtain clarification from the company administration.
A: non serve a niente non è possibile fare quello che dice lei è già in corso (la procedura di verifica) bisogna attendere signora *(it doesn't worth, it is not possible to do what you say.. is already processing the check . Is necessary to wait madame?)* **C:** cioè ma veramente avrei dovuto rispondere cosi quando dovevo pagare settemila euro, va bene la ringrazio, arrivederci salve *(that is.. really! I should answer like you when I had to pay 7000 euros, alright thank you, goodbye)*	The customer is offended by the agent mood.

Table 3.7: An example of the manifestation of **not-complete dissatisfaction** with manual transcription. C - Customer

Dialog excerpt	Notes
C: Va bene, a posto, arrivederci. *(Al right, it's okay, goodbay.)*	Customer close the conversation without solving her problem, with an ironic sentence.

Other (O) It is an implicit annotation, a segment that appears after any emotional label of the conversation, which has not been defined as any of the emotional episode mentioned above.

Annotation Procedure

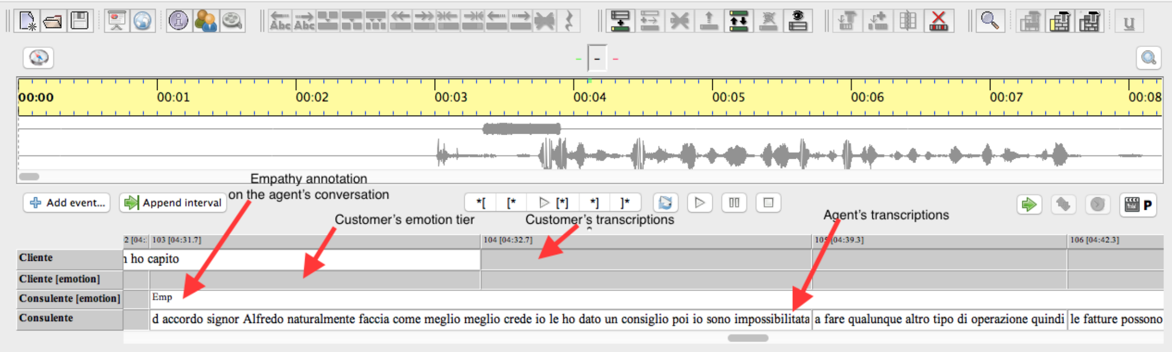
By following the guidelines and the recommendations, we trained two expert annotators with the psycholinguistic background. For this annotation task, only the emotional categories were annotated. They were instructed to mark the speech segments where they perceive a transition in the emotional state of the speaker with the relevant emotional label.

In the annotation process, we set the goal of maximizing the number of annotated conversations. For this reason, we annotated the first instance of neutral-empathy segment pairs within each conversation. Annotators manually refined the boundaries of the segments generated by an off-the-shelf Speech/Non-Speech segmenter.

The annotators tagged the `empathy` segments on the agent channel and the basic emotion such as *anger* and complex emotions such as *frustation*, *satisfaction*, *not-complete satisfaction*, *dissatisfaction* and *not-complete dissatisfaction* on the customer channel.

Annotators were instructed to select the candidate segment pairs with a decision refinement process. Once the relevant speech region was identified, the annotators could listen to the speech as many times as they needed to judge if the selected segment(s) could be tagged with any of the target labels. The average per-conversation annotation time was 18 minutes for an average duration of a conversation of 6 minutes.

The annotation of emotional attitude is carried out using the Partitur-Editor of EXAMRaLDA [151]. EXMARaLDA is an acronym of "Extensible Markup Language for Discourse Annotation". It is a tool of concepts, data formats, and tools for the computer assisted transcriptions and annotation of spoken language, and for the construction and analysis of spoken language corpora. The annotation of the emotional attitudes requires that a specific tier of annotation is added to the file. It is not necessary to work on the transcriptions of the audio file. An example of annotation using Partitur-Editor is shown in Figure 3.2, which contains emotion and transcription tiers. Annotation of empathy is started at 271.7 seconds and finished at 291.97 seconds in a conversation.



Cliente: ı ho capito

Consulente [emotion]: Emp

Consulente: d accordo signor Alfredo naturalmente faccia come meglio meglio crede io le ho dato un consiglio poi io sono impossibilitata | a fare qualunque altro tipo di operazione quindi | le fatture possono

Figure 3.2: Annotation example using Partitur-Editor, containing emotion and transcription tiers. Annotation of empathy is started at 271.7 seconds and finished at 291.97 seconds.

Annotation Evaluation

To assess the reliability of the annotation model, we designed the following evaluation task. Two annotators, with the psychology background, worked independently over a set of 64 spoken conversations randomly selected from the corpus. The annotators were of similar age, same ethnicity, and opposite gender. We intended to assess if the annotators could perceive a change in the emotional state at the same onset positio, e.g., `neutral` leading into `empathy`, and their inter-annotator agreement with the assignment of the labels.

In 53.1% of the annotated segments, the two annotators perceived the empathic attitude of the agent in adjacent segments, while 31.2% of the speech segments were tagged with empathy labels on the same onset position.

To measure the reliability of the annotations we calculated inter-annotator agreement by using the kappa statistics [121, 152]. It was originally designed to measure the agreement between two annotators for categorical labels. Later, it has been extended for multiple annotators [153]. It is frequently used to assess the degree of agreement among any number of annotators by excluding the probability that they agree by chance. The kappa coefficient ranges between 0 (agreement is due to chance) and 1 (perfect agreement). Values above 0.6 suggest acceptable agreement. Our annotation task was challenging because it combined categorical annotation with the continuous perception of the slowly varying emotion expression from speech-only stimuli. Thus, we evaluated the agreement between annotators based on a partial match: two annotators agreed on the selection of the onset time stamps within a tolerance window of 5 sec. We found reliable results with kappa value 0.74. Most categorical emotion annotation research in speech deals with the lower human agreement (greater than 0.50) maybe due to a variety of factors, including short audio clips or utterance ([146, 154]), multi-label annotation tasks, and annotator agreement when the annotation task is based on continuous and discrete label annotations [149]. In our case, the positive evaluation results may be motivated by the operationalized definition of emotional states, by the observability of the complete paralinguistic and linguistic contexts.

3.1.2 Transcriptions

Manual Transcriptions

A subset of the corpus has also been manually transcribed for a deeper understanding of *what* has been said by the agent and customer during their conversational interaction. The manual transcriptions in our corpus contain *955* conversations. It has been used to design ASR, as detailed in the following section. Moreover, in Section 3.1.4.4, an emotion related specific linguistic analysis has also been presented based on the manual transcriptions.

Automatic Transcriptions

The automatic transcriptions were generated using a large vocabulary ASR system [155]. It was designed using a subset (see Section 3.1.2) of 1894 conversations containing approximately 100 hours of spoken content and a lexicon of ∼18000 words. Mel Frequency Cepstral Coefficient (MFCC) features have extracted from the conversations and then spliced by taking three frames from each side of the current frame. It was followed by Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature-space transformations to reduce the feature space. Then, the acoustic model was trained using speaker adaptive training (SAT). In order to achieve a better accuracy Maximum Mutual Information (MMI) was also used. The Word Error Rate (WER) of the ASR system was 31.78% on the test set and 20.87% on the training set, using a trigram language model of perplexity 87.69. For the training and decoding process, an open-source implementation system, Kaldi [156], was used.

3.1.3 Speech *vs* Non-Speech Segmentation

An in-house speech *vs* non-speech segmenter has been designed using a set of 150 conversations, 300 wave files, containing approximately 100 hours of spoken content and used Kaldi [156] for the training and decoding process. Training data has been prepared using force-aligned transcriptions. Mel Frequency Cepstral Coefficient (MFCC) and their derivatives have been used as features. Number of Gaussian and beam width have been optimized using a development set of 50 conversations, 100 wave files. The final model has been designed using 64 Gaussians and a beam width of 50, which has been tested using a test of 50 conversations, 100 wave files. As a part of the post-processing, three rules has been applied: 1) removed non-speech segments, which are between speech segments and are less than 1 second, 2) added an non-speech segment between speech segments if there is a gap greater than 3 seconds, 3) concatenated the consecutive speech and non-speech segments, respectively. The F-measure of the system was 66.0% on the test set. We use the term *SISL speech segmenter* to refer to the segmenter mentioned here.

3.1.4 Corpus Analysis

3.1.4.1 Corpus Summary

Based on the annotation guidelines, 1894 conversations were annotated with empathy (Emp) on the agent channel, and other basic and complex emotions on the customer channel. Also, agent and customer are also specified in the annotation. The corpus consists of 210 hours and 23 minutes of conversations with average ± std is 406.3 ± 196.3. Out of 1894 conversations, 515 conversations are neutral in respect of these emotional episodes and in 1379 conversations at least one of the channels (i.e., either agent or customer) contains these emotional episodes, as shown in Table 3.8. Several interesting characteristics have been observed that are as follows. In about 27% of the conversations neither customer nor agent manifested any emotions as shown in Figure 3.3, where in 73% conversations at least one of the channel, either agent or customer, manifested emotions. In the subsequent analysis, we ignored the 27% conversations from the computation of corpus statistics. However, these 27% conversations have been used for conversation level classification experiments. We also grouped satisfaction and not-complete satisfaction into satisfaction, and dissatisfaction and not-complete dissatisfaction into dissatisfaction.

The next observations are that, in the subset of the corpus where at least one of the channel contains emotions, in about 29% cases both customer and agent manifested emotions as shown in Figure 3.4. These conversations represent ideal situations where *affective scene* can be explained. In about 62% cases customer manifested emotions where the agent was neutral, even if the customer were angry, frustrated, or dissatisfied. This is the case where agent needed to be active and positively responsive about the customer's emotions. Withing this 62% cases, in 27% cases customer was manifesting

Table 3.8: Statistics for the behavioral corpus.

Description	Count
Total no. of conversations	**1894**
Conversations containing emotions at least one of the channels	**1379**
Conversations that don't have any of the emotional tags	**515**
Agent channels containing empathy	**525**
Agent channels do not contain empathy	**1369**
Agent channels contain empathy and customer neutral	**121**
Customer channels containing emotional tags	**1258**
Customer channels do not have any emotional tags	**636**
Customer channels contain emotions and agent neutral	**854**
Agent and customer containing emotional tags	**404**
Out of 1379 conv. Agent's conv. are neutral	**854**
Out of 1379 conv. Customer's conv. are neutral	**121**

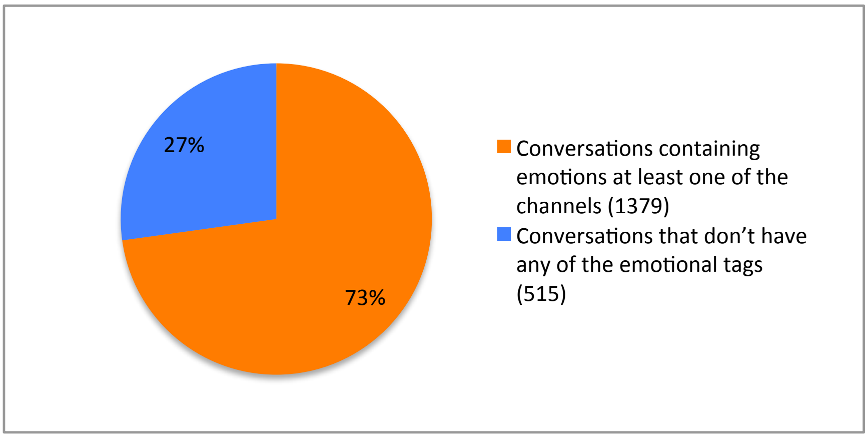


Figure 3.3: Distribution of conversations with and without containing emotional episodes.

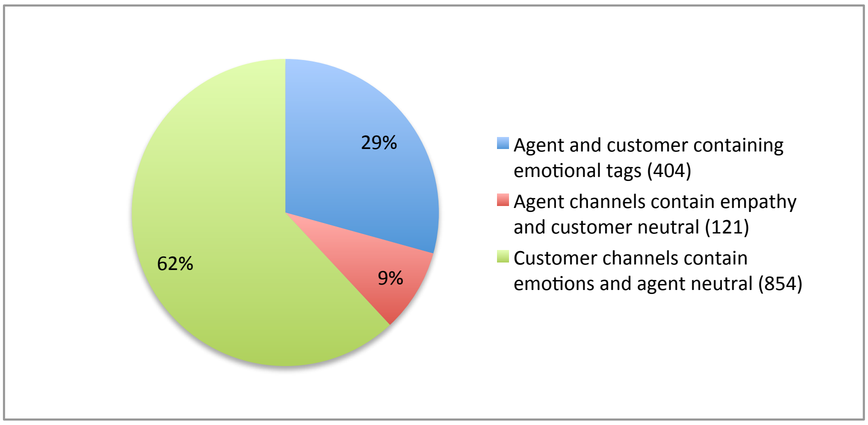


Figure 3.4: Distribution of emotional episodes regarding agent's and customer's emotional manifestations from the set of the corpus where at least one of the channel contains emotion.

negative emotions, and in 35% cases customer expressed satisfaction even in agent was not empathic. The rest 9% is also a perfect condition for call center scenario in an assumption that even if the customer does not show any negative emotional manifestation, the agent still can be empathic.

Note that, annotation has been done at the segment level, therefore, in the whole conversation, there might be more than one emotional segment. This is true for both agent and customer channels. Moreover, on the customer channel, the same conversation might have more than one emotional episodes. In Figure 3.5, we present a complete breakdown of the 62% conversations with the emotional sequence. As we see in the Figure, in many cases we have a single emotional segment in customer channel emotion. Only $\sim 20\%$ cases we have more than one emotional segments. However, in those $\sim 20\%$ some interesting phenomena exists such as anger followed by frustration, frustration leads to anger, dissatisfaction or in some cases satisfaction.

In Figure 3.6, we present another break-down analysis of empathic conversations. We observed that even if the agent was empathic, the customer was not convinced and manifested negative emotions such as anger, frustration and dissatisfaction, which has happened in 30% cases.

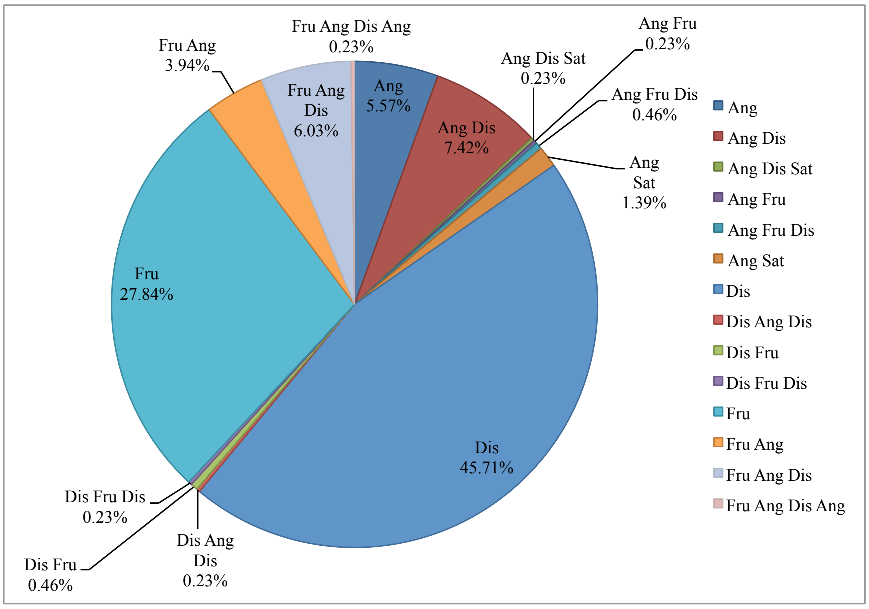


Figure 3.5: Emotional sequence in conversations on the customer channel, which is a break down of 62% conversations from the Figure 3.4.

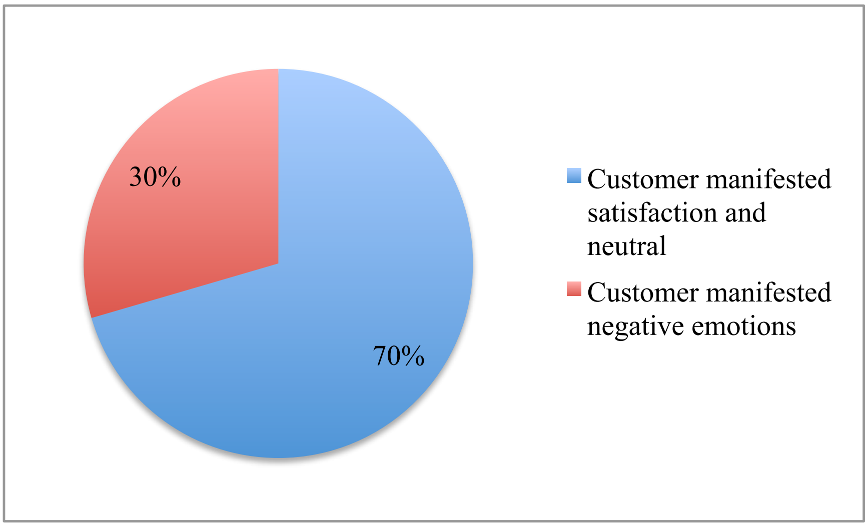


Figure 3.6: Conversations with empathic manifestations on the agent side and basic and complex emotions on the customer side with the positive and negative outcome.

Some other intersting observations and scenarios has also found from our analysis pointed out below:

- Customer was showing anger and frustration and agent was neutral. This is the case in which agent was needed to be empathic but failed to be so.
- Agent was empathic while customer was either anger, frustrated or both. It shows that agent attempted but not able to sooth customers negative emotions.

Such evidence is important for a call center quality assurance manager. There might also be the cases that customer shows grievance during the call not because of the expected services from the company but because of his/her previous personal reasons. It might be difficult for the agent to tackle such as situation and we did not provide analysis on such evidence as this is beyond the scope of this study.

3.1.4.2 Segment/Duration Distribution

A duration distribution of emotional segment is given in Table 3.9, which shows that the observed emotional manifestations are very low for anger and dissatisfaction. In the table, we present segment's duration distribution, where average duration of satisfaction, not-complete dissatisfaction, and not-complete satisfaction categories are lower compared to other categories.

3.1.4.3 Speaker/Gender Distribution on the Agent Side

On the agent side, the corpus contains speaker meta-data where we have 1403 speakers on the whole 10K dataset, however, in the SISL affective corpus the conversations are from 763 speakers. In Figure 3.7, we present the distribution of calls received per operator, which presents a skewed distribution of the call received per speaker. Very few speakers received more than two calls. Due to the lack of information about the customer identity and the call we do not know whether the same customer called more than once

Table 3.9: Duration and segment statistics of emotional segments in the corpus.

Channel	Class	Avg. (sec)	Std. (sec)	Total (H:M:S)	# of Seg
Agent	**Emp**	18.56	12.75	2H 44M 15.38S	531
	Neu	219.87	148.36	32H 25M 50.41S	531
	O	195.51	187.05	28H 24M 11.66S	523
Customer	**Ang**	17.94	10.65	35M 34.38S	119
	Fru	15.70	9.94	1H 28M 12.44S	337
	Dis	15.83	8.88	30M 20.634S	115
	Ncd	8.17	5.97	37M 19.10S	274
	Sat	8.25	4.66	41M 22.44S	301
	Ncs	6.94	4.64	50M 31.65S	437
	Neu	282.28	191.41	123H 34M 31.26S	1576
	O	46.63	111.41	12H 43M 10.56S	982

or not. Therefore, we do not present such information. In Table 3.10, we present the number of conversations and the distribution of the unique male and female speaker for both SISL affective corpus and the whole dataset. The reason for presenting the information of the whole dataset is to show the number of conversations that we have with male and female tags. Given this distribution of male and female conversations in the corpus, a general automatic gender identification model can be designed. For the customer side conversations, we do not have male and female tags for the whole dataset as can be seen in Table 3.10.

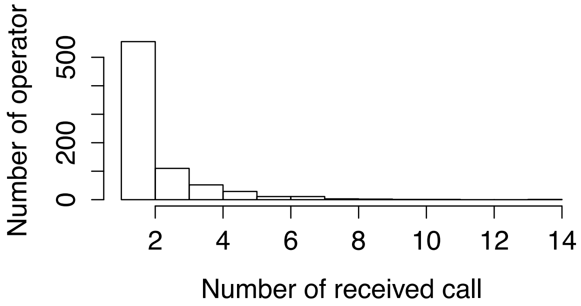


Figure 3.7: Distribution of the number of call received per agent in the behavioral corpus.

Table 3.10: Gender distribution on the agent side of the datasets.

Channel	Gender	SISL affective corpus		Whole dataset (10K)	
		Unique Speaker(%)	No. of Conv	Unique Speaker (%)	No. of Conv
Agent	**F**	519 (65%)	1097	947 (65%)	6254
	M	282(35%)	671	456 (35%)	3746
Customer	**F**	619 (51%)	619	No gender info	
	M	606 (49%)	606		

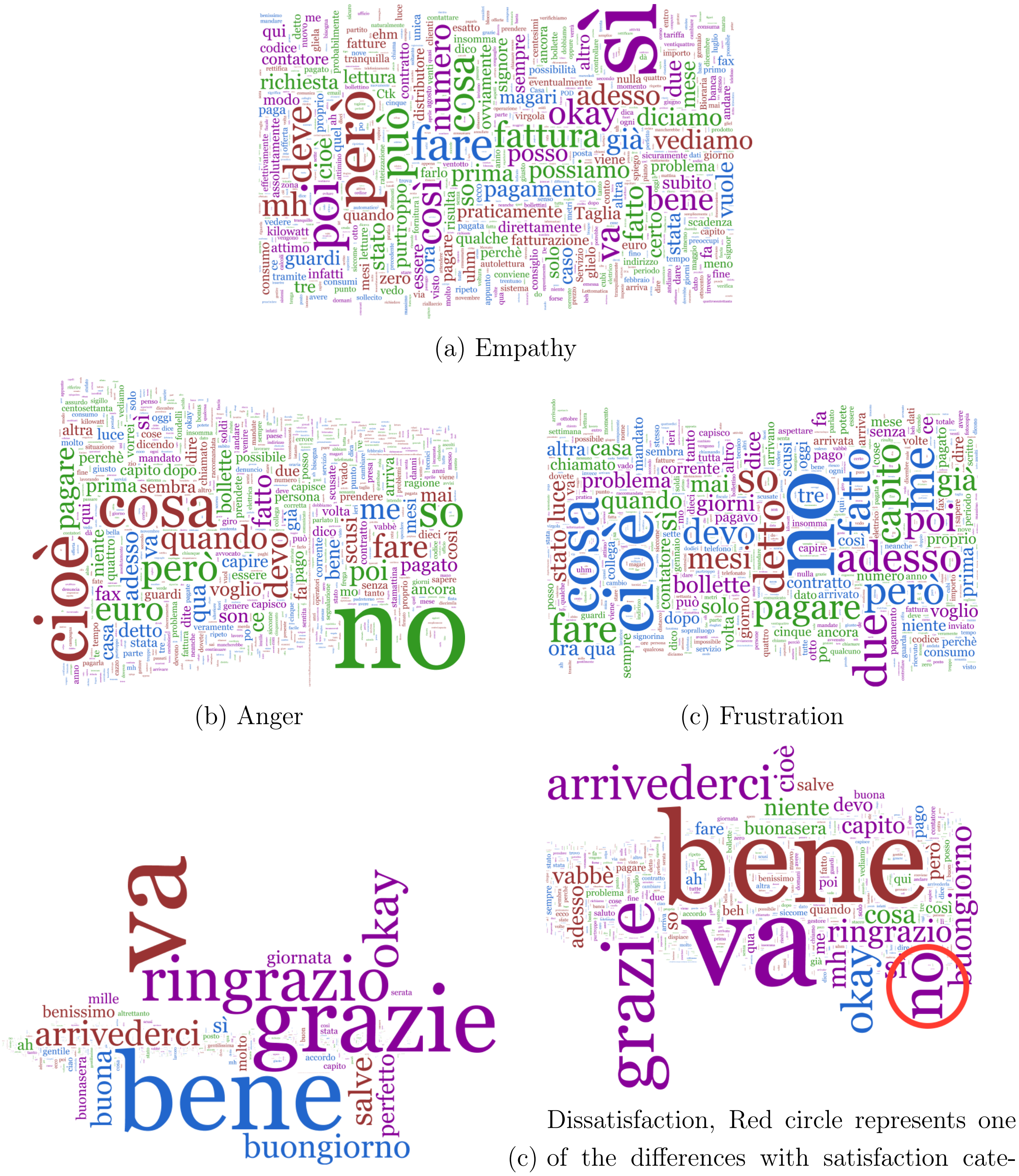
3.1.4.4 Linguistic Analysis

An analysis has been conducted using manual transcriptions to understand *what* has been said while manifesting emotion in conversations. It includes finding most frequent as well as syntactic (part-of-speech) categories. Since a subset of emotional conversations has manual transcriptions, therefore, all linguistic analysis has been done on that subset.

Token Analysis For token based analysis, our investigation includes n-gram and the word-cloud. Both approaches are based frequency analysis and for both cases, we removed stop words. In Table 3.11, we present a few top ranked tri-grams for each emotional category and in Figure 3.8 we present word cloud. The empathic manifestations represent the words that show the compassionate response such as “vediamo” (let’s see) or “posso aiutarla” (can I help you). The negativity is mainly represented by anger manifestations. As we see in the Table, there are also slang words, which shows the extreme form of anger. Frustration also shows the negativity, however, it mostly shows different kinds of problems. The manifestations of satisfaction show the positivity, which is clearly distinguishable from other categories. In the case of dissatisfaction both the table and the word-cloud shows that it is overlapped with satisfaction and frustration. The reason of overlap between satisfaction and dissatisfaction might be because they are appearing at the end of the conversation.

Table 3.11: Top ranked tri-grams for each emotional states. English translations are inside parenthesis.

Emotional state	Top ranked 3-grams
Empathy	vediamo (let's see), posso aiutarla (can I help you), possiamo, facciamo, abbiamo, vediamo un attimo (we see a moment), quindi stia tranquilla (then do not worry), possiamo gestire (we can handle), assolutamente infatti (absolutely indeed)
Anger	non (no), io non (I do not), non è possibile (it is not possible), problema (problem), non prendete (do not take), problemi (problems), non ho ancora (I have not), perché io (because I), pago carissimo eh c***o (I pay dear eh f ***), non si fanno (do not make)
Dissatisfaction	non è possibile (it is not possible), non è (is not), io non (I do not), problema (the problem), purtroppo (unfortunately), è il problema (is the problem), perché non è (because it is not), l aspettavo perché (because the expected), non è aumentato (has not increased), vi denuncio lo (will sue him)
Frustration	impossibile (impossible), devo aspettare (I have to wait), purtroppo (unfortunately), non ho capito (I did not understand), vorrei capire (I would understand), perché guardi (because they look at it), capire quando (figure out when), già partiamo male (already we start wrong), non è mai (it's never), non ha capito (did not understand), problema perché (because the problem), può essere che (can be that)
Satisfaction	bene (well), va bene (well), perfetto (perfect), bene grazie (well thank you), va bene grazie (well thank you), okay va bene (okay good), molto gentile (very kind), gentilissima (gentle), perfetto la ringrazio (perfect thank you), capito va bene (got it all right), perfetto grazie mille (well thank you so much), la ringrazio tanto (thank you so much)



(a) Empathy

(b) Anger

(c) Frustration

(e) Satisfaction

(c) Dissatisfaction, Red circle represents one of the differences with satisfaction category.

Figure 3.8: Word cloud for each emotional category.

Table 3.12: Top-ranked POS tags for the agent's channel emotion. POS tags are in ranked order.

Empathy	
POS	**Description**
SS	Singular noun
B	Adverb
E	Simple preprosition
RS	Singular article
C	Conjunction
PS	Singular pronoun
AS	Singular qual. Adj.
VF	Main Verb

Part-of-speech (POS) Analysis In Table 3.12 and 3.13, we present the top-ranked parts of speech categories for agent and customer's emotional states, respectively. For extracting the POS tags, we used TextPro [157], a suite of NLP tools for analysis of Italian and English text. For empathy, adverb appears as second most, whereas for other emotional categories it appears in the top.

Feature Selection and Ranking Since frequency based analysis does not entail that top ranked tokens or ngrams are important. Therefore, we also investigated feature selection followed ranking based approach before the classification. For this analysis, we extracted trigrams from manual transcriptions, in order to understand whether there are any linguistically relevant contextual manifestations while expressing any emotional manifestations. For the analysis of the lexical features, we used Relief feature selection algorithm [158]. Prior to the feature selection, we have transformed the raw lexical features into bag-of-words (vector space model), and then transformed into logarithmic term frequency (tf) multiplied with inverse document frequency (idf). Then, we applied Relief feature selection algorithm and ranked the features, based on the score computed by the algorithm. More details

Table 3.13: Top-ranked POS tags for each emotion category of the customer's channel emotion. POS tags are in ranked order.

Anger		**Dissatisfaction**	
POS	**Description**	**POS**	**Description**
B	Adverb	B	Adverb
SS	Singular noun	SS	Singular noun
PS	Singular pronoun	PS	Singular pronoun
E	Simple preprosition	E	Simple preprosition
C	Conjunction	AS	Singular qual. Adj.
RS	Singular article	C	Conjunction
AS	Singular qual. Adj.	RS	Singular article
VF	Main Verb	I	Interjection
CCHE	Che	VF	Main verb infinitive
SP	Plural noun	SP	Plural noun
Frustration		**Satisfaction**	
POS	**Description**	**POS**	**Description**
B	Adverb	B	Adverb
SS	Singular noun	SS	Singular noun
PS	Singular pronoun	AS	Singular qual. Adj.
E	Simple preprosition	I	Interjection
RS	Singular article	RS	Singular article
C	Conjunction	C	Conjunction
VSP	Main verb. Past part. singular	PS	Singular pronoun
VIY	Aux. verb	E	Simple preprosition
AS	Singular qual. Adj.	VSP	Main verb. Past part. singular
SP	Plural noun	SP	Plural noun

of this approach can be found in Chapter 4. For example, using such an approach we found that agent commonly used “posso aiutarla” (can I help you) while manifesting empathy.

3.1.4.5 Emotional Onset Analysis

We have done an analysis to understand whether different acoustic and lexical patterns exist before and after the onset point, from the neutral to the emotional segment as can be seen in Figure 3.9. The purpose was to understand whether such difference can help in designing automatic emotion segmenter.

The whole pipeline for the feature pattern analysis is shown in Figure 3.10. For the analysis, we extracted low-level features such as fundamental frequency, mfcc and others from both segments, i.e., neutral and empathy. The feature extraction has been done for each conversation and for both segments separately. In the next phase, we computed the average for each feature and for each segment. In this phase, we align both segments for each conversation. It resulted in two vectors for each feature. Then, we applied the statistical test, such as two-tailed two-sample t-test in order to find the difference. The findings of this analysis are presented in Chapter 5.

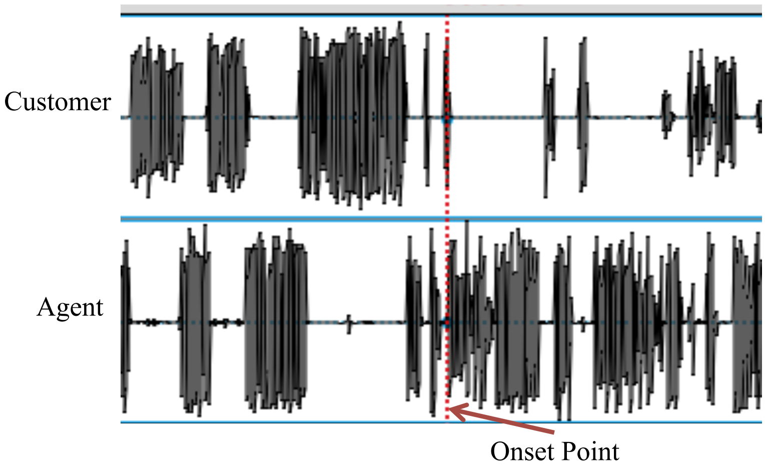


Figure 3.9: Onset point of an emotional segment, in this case neutral to empathy.

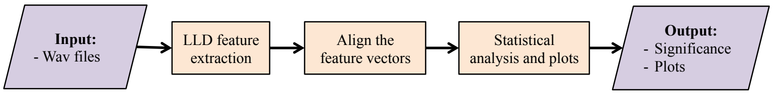


Figure 3.10: Onset point of an emotional segment, in this case neutral to empathy.

3.1.5 Data Preparation for Classification Experiments

As can be seen in Table 3.9 and 3.14 for some emotional states the number of annotations are very low, which may results in skewed class distribution. Hence, we grouped some of the categories into a particular category, whenever possible in order to facilitate the designing of the computational models. For example, satisfaction and not complete satisfaction is grouped into satisfaction. This type of grouping has been done both for conversation and segment level classification experiments.

3.1.5.1 Conversation Level Classification

Corpus was prepared with the conversations that are listened by the annotator. For empathy only agent channel is considered. For frustration, satisfaction and anger only customer channel was considered. Therefore, for empathy, positive are those that contain empathic emotional marker and negative are those that are neutral. For frustration, satisfaction and anger, positive examples are those that contain respective emotional marker and negative are those that include neutral and other emotional markers as shown in figure 3.11. For the classification experiments at the conversation level, we grouped satisfaction and not complete satisfaction into satisfaction and dissatisfaction and not complete dissatisfaction into dissatisfaction. As a result, the class labels for our conversation level experiment consist of empathy, frustration, satisfaction, anger and dissatisfaction as shown in Table 3.15. Even after grouping, class distribution is still skewed. In Chapter 5, we will see how we upsampled the minority class and downsampled the majority class in order to solve the unbalanced class distribution. For the classification, we designed a binary classifier for each emotional category and used

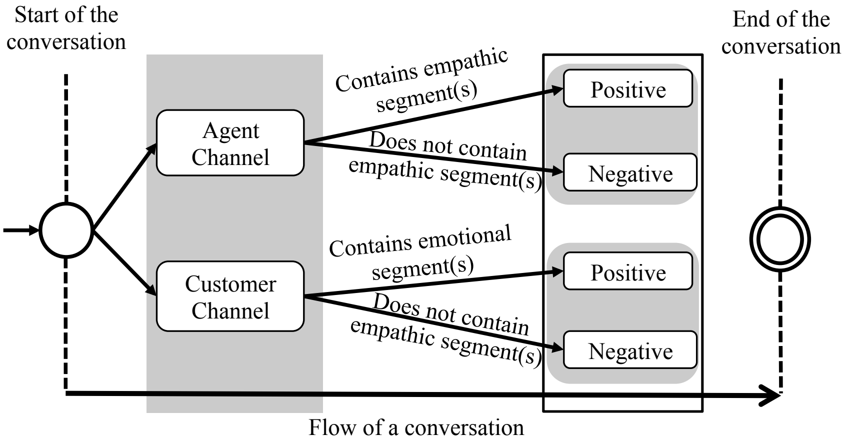

Start of the conversation
Agent Channel
Customer Channel
Contains empathic segment(s)
Does not contain empathic segment(s)
Contains emotional segment(s)
Does not contain empathic segment(s)
Positive
Negative
Positive
Negative
End of the conversation
Flow of a conversation

Figure 3.11: Data preparation for the conversation level classification experiments.

Table 3.14: Distribution of the class labels at the conversation level **before grouping**.

Class	Y	N	Total	Y (%)	N (%)
Emp-agent	525	1250	1775	0.30	0.70
Ang-customer	118	1776	1894	0.06	0.94
Fru-customer	338	1556	1894	0.18	0.82
Dis-customer	108	1786	1894	0.06	0.94
Ncd-customer	274	1620	1894	0.14	0.86
Sat-customer	301	1593	1894	0.16	0.84
Ncs-customer	434	1460	1894	0.23	0.77

different cross-validation methods for the evaluations. More details of the classification and evaluation approaches can be found in Chapter 4, 5 and 6.

Table 3.15: Distribution of the class labels at the conversation level **after grouping**.

Class	Y	N	Total	Y (%)	N (%)
Emp-agent	525	1250	1775	0.30	0.70
Ang-customer	118	1776	1894	0.06	0.94
Fru-customer	338	1556	1894	0.18	0.82
Dis-customer	382	1512	1894	0.20	0.80
Sat-customer	735	1159	1894	0.39	0.61

3.1.5.2 Segment Level Classification

For the segment level classification, our study include binary and multi-class classification experiments for the agent and customer channel's emotional states, respectively. In order to train and evaluate the classification system, we split the data into train, development and test set with a proportion of 70%, 15% and 15% repectively as shown in Table 3.16. For the agent channel, we have speaker information, therefore, we separated data in speaker independent way. Whereas for the customer channel, we do not have speaker information, therefore, we assumed that each conversation is independent and data has been separated at the conversation level. As mentioned earlier, we grouped some emotion categories to reduce the class imbalance problem for the classification experiments, which we present in Table 3.17. For example, class label 'Neg', was obtained by grouping anger and frustration. We also removed the segments associated with class label O from the classification experiments.

Table 3.16: Data-split and the distribution of the associated manual segments.

Channel	Class	Train	Dev	Test	Total
Agent	**Emp**	372	78	80	530
	Neu	371	79	80	530
	O	361	77	78	516
	Total	1104	234	238	1576
Customer	**Ang**	89	18	11	118
	Dis	85	14	16	115
	Fru	227	56	51	334
	Ncd	188	48	35	271
	Ncs	303	60	67	430
	Neu	1090	240	230	1560
	O	397	79	105	581
	Sat	206	44	49	299
	Total	2585	559	564	3708

Table 3.17: Data-split and the distribution of the associated manual segments after grouping.

Channel	Class	Train	Dev	Test	Total
Agent	**Emp**	373	80	78	531
	Neu	373	80	78	531
	Total	1113	239	233	1585
Customer	**Neg**	597	119	129	845
	Dis	1099	232	245	1576
	Neu	1099	232	245	1576
	Sat	507	115	116	738
	Total	2896	605	640	4141

3.2 FAU-Aibo Robot Corpus

FAU-Aibo Robot Corpus [28], is one of the publicly available human-machine real-life, spontaneous, corpus containing recordings where children are interacting with Sony's pet robot Aibo. Even if Aibo was controlled by a human operator, however, the interactions between children and Aibo was set up in such a manner that children believed Aibo was responding to their commands. Aibo's actions were very predetermined, which caused children to manifest emotions. The data consists of recording from 51 children, which has been collected from two different schools such as MONT and OHM. The recording contains 9.2 hours of speech with 16 bit, 16 kHz. The recordings were segmented automatically into turns using a pause threshold of 1 second. The annotation has been done at the word level and then combined them into chunk level. It has been distributed into different set such as Interspeech emotion challenge set [102] and CEICES [28]. The distribution of different split is presented in Table 3.18, 3.19, and 3.20. For the purpose of this study, we utilized IS09 dataset with 2-classes and 5-classes for cross-corpus study and CEICES dataset for emotion segmentation task.

Table 3.18: Data split with 2-classes problem of the IS2009 Emotion challenge.

Class	**Neg**	**IDL**	**Total**
Train	1964	4178	6142
Dev	1394	2423	3817
Test	2465	5792	8257
Total	5823	12393	18216

Table 3.19: Data split with 5-classes problem IS2009 Emotion challenge.

Class	**Ang**	**Emp**	**Neu**	**Pos**	**Rest**	**Total**
Train	525	1225	3435	475	482	6142
Dev	356	868	2155	199	239	3817
Test	611	1508	5377	215	546	8257
Total	1492	3601	10967	889	1492	18216

Table 3.20: CEICES data split with 4-classes at the word level

Class	**Ang**	**Emp**	**Motherese**	**Neu**	**Total**
Train	530	629	758	483	2400
Dev	390	350	228	354	1322
Test	637	666	237	808	2348
Total	1557	1645	1223	1645	6070

3.3 Summary

In this chapter, we described the SISL affective behavior corpus, which is collected from naturally occurring conversations in call centers. Since the goal of the experiment was to collect ecologically valid data, therefore, no knowledge has been given to the subject regarding the experiment. For the annotation of affective behavior, we designed an annotation model by following the Gross's modal model of emotion. The affective behavior of the corpus include empathy of the agent channel and basic and complex emotion of the customer channel. The basic emotion includes anger, whereas the complex emotion includes frustration, satisfaction, and dissatisfaction. We described how data collection and annotation has been done in detail and also provided detail corpus summary and statistics. In addition to the SISL affective behavior corpus, we also discussed FAU-Aibo Robot corpus, which we used generative based segmentation and classification task in Chapter 7.

Chapter 4

Features, Classification and Evaluation

In this chapter, we present the type of features, different classification methods and evaluation metrics that we investigated for the experiments of affective behavior. As classification methods, we employed two different approaches. One approach is to use a discriminative based static classifier, such as SVM, to classify emotion at the conversation and segment level. The other approach is to use a dynamic classifier as opposed to say generative classifier, such as Hidden Markov Models (HMMs), for segmenting the conversations and labeling the each segment with an emotional label. For the two different classification experiments, we also employed two different approaches to measure the performance of the systems as detailed in the following subsections. For the former approach of classification, in addition to acoustic features, we also utilized lexical, and psycholinguistic features. For the later approach, we only relied on acoustic features because of the complexity of the task. In the following sections, we discuss the detail of the feature extraction, selection, and fusion, followed by classification and evaluation methods.

4.1 Feature Extraction

In this section, we report the approaches used to extract the acoustic, linguistic and psycholinguistic features. We discuss differences and similarities of the features used in our experiments to those used in other emotion and personality recognition tasks [133].

4.1.1 Acoustic Features

We extracted acoustic features using openSMILE [128], a feature extraction tool capable of extracting a large set of audio features. The low-level acoustic features were extracted with approximately 100 frames per second, with $25-60$ milliseconds per frame. These low-level descriptors (LLDs) were then projected onto single scalar values by descriptive statistical functionals. The details of the low-level features and statistical functionals are given in Table 4.1. For the generative based approach, we directly used the low-level descriptors (LLDs).

Table 4.1: Low-level acoustic features and statistical functionals

Low-level acoustic features
Raw-Signal: Zero crossing rate
Energy: Root-mean-square signal frame energy
Pitch: F0 final, Voicing final unclipped, F0 final - nonzero
Voice quality: jitter-local, jitter-DDP, shimmer-local, log harmonics-to-noise ratio (HNR)
Spectral: Energy in bands 250-650Hz, 1-4kHz, roll-off-points (0.25, 0.50, 0.75, 0.90), flux, centroid, entropy, variance, skewness, kurtosis, slope band (0-500, 500-1500), harmonicity, psychoacoustic spectral sharpness, alpha-ratio, hammarberg-index
Auditory-spectrum: band 1-10, auditory spectra and rasta
Cepstral: Mel-frequency cepstral coefficitnts (mfcc 0-3)
Formant First 3 formants and first formant bandwidth
Statistical functionals
Relative position of max, min
Quartile (1-3) and inter-quartile (1-2, 2-3, 3-1) ranges
Percentile 1%, 99%
Std. deviation, skewness, kurtosis, centroid, range
Mean, max, min and Std. deviation of segment length
Uplevel time 25 and rise time
Linear predictive coding lpc-gain, lpc0-1
Arithmatic mean, flatness, quadratic mean
Mean dist. between peaks, peak dist. Std. deviation, absolute and relative range, mean and min of peaks, arithmatic mean of peaks, mean and Std. of rising and falling slope

4.1.2 Lexical Features

The transcriptions of the conversations or segment were converted to bag-of-n-grams vectors weighted with logarithmic term frequencies (tf) multiplied with inverse document frequencies (idf) as presented in the equation 4.1.

$$tf \times idf = log(1 + f_{ij}) \times log\left(\frac{\text{number of conversations}}{\text{number of conversations that include word i}}\right) \tag{4.1}$$

where f_{ij} is the frequency for word i in conversation j.

For the conversation level experiment, we considered the conversation as a document. Where for the segment level experiment we considered the segment as a document. In order to take advantage of the contextual benefits, we extracted n-grams with $3 \geq n \geq 1$. While doing so, we also removed stop words. Because this resulted in an unreasonably large dictionary, we filtered out lower frequency features by preserving $10K$ most frequent n-grams.

4.1.3 Psycholinguistic Features

Similar to the lexical features we extracted the so-called psycholinguistic features from the transcriptions. Over the past few decades, Pennebaker et al. have designed psycholinguistic word categories using high frequency words and developed the Linguistic Inquiry Word Count (LIWC) [159]. These word categories are mostly used to study gender, age, personality, and health to estimate the correlation between these attributes and word usage (see [160] and the references therein). It is a knowledge-based system containing dictionaries for several languages. We used the dictionary that is available within LIWC for Italian [161]. The Italian dictionary contains eighty five word categories. Using the LIWC system we extracted five general descriptors and twelve punctuation categories constituting a total of one hundred two features. We then removed LIWC features, which were not observed in our dataset and obtained a final set of 89 psycholinguistic features. The LIWC feature processing differs according to types of features. Some features are counts and others are relative frequencies (see [162]). The types of LIWC features that we extracted are presented in Table 4.2.

Table 4.2: Psycholinguistic features extracted using the LIWC system

LIWC features
General features
Word count, words/sentence, percentage of words exist in dictionary, percentage of words greater than 6 letters, and numerals
Linguistic features
Pronouns, articles, verbs, adverbs, tense, prepositions, conjunctions, negations, quantifiers, and swear words
Psychological features
Social processes: family, friends and humans **Affective processes:** positive, negative, anxiety, anger, and sadness **Cognitive processes:** insight, causation, discrepancy, tentative, certainty, inhibition, inclusive, exclusive, perceptual, see, hear, and feel **Biological processes:** body, health, sexual, and ingestion **Relativity:** motion, space, and time
Personal concern
Work, achievement, leisure, home, money, religion, and death
Paralinguistic features
Assent, nonfluencies, and fillers
Punctuation features
Period, comma, colon, semi-colon, question mark, exclamatory mark, dash, quote, apostrophe, parenthesis, other punctuations, and percentage of all punctuations

4.2 Feature Selection

For feature selection, we used the *Relief* [158] feature selection technique. In [163], we comparatively evaluated the Relief method against other algorithms, and it outperformed them in the classification performance and computational cost. We ranked the feature set according to Relief scores and generated the learning curves by incrementally adding batches of ranked features. We then selected the optimal set of features by stopping when the performance saturated or started decreasing [163].

The goal of Relief is to estimate weights to find relevant attributes with the ability to differentiate between instances of different classes under the assumption that nearest instances of the same class have the same feature values, and different class has different values. Relief estimates weight of a feature, F, using Equation 4.2.

$$W[F] = P(x|nearest\ miss) - P(x|nearest\ hit) \tag{4.2}$$

where x is a value of F, *nearest miss* and *nearest hit* are the nearest instances of the same and a different class, respectively.

Since some feature selection algorithms do not support numeric feature values such as information gain and suffer from data sparseness such as Relief, we discretized feature values into ten equal-frequency bins [164] as a pre-processing step of the feature selection. The equal-frequency binning approach divides data into $k = 10$ groups, where each group contains an approximately equal number of values. The equal-frequency binning approach and the size of the bin, $k = 10$, were empirically found optimal in other paralinguistic task [163]. It is needed to mention that we applied feature selection on acoustic, linguistic and their combination, but not on the psycholinguistic features due to the limited size of the feature set.

4.3 Feature Fusion

We merged acoustic and lexical features into a single vector to represent each instance in a high-dimensional feature space. Let $A = \{a_1, a_2, ..., a_m\}$ and $L = \{l_1, l_2, ..., l_n\}$ denote the acoustic and lexical feature vectors respectively. The feature-combined vector was $Z = \{a_1, a_2, ..., a_m, l_1, l_2, ..., l_n\}$ with $Z \in R^{m+n}$. Given the high-dimension of the feature vector and the *curse of dimensionality*, we applied feature selection on the Z space to achieve an optimal feature dimension of size k, $k < (m + n)$. Another objective of the feature selection process was to find the best compromise between the dimension of the input and the performance of a target classifier.

4.4 Classification Approaches

For the static classification approach, we trained our classification models using a different implementation of Support Vector Machines (SVMs), Sequential Minimal Optimization (SMO) [165], which is a technique for solving the quadratic optimization problem of SVMs training. We utilized an open-source implementation Weka machine learning toolkit [164] for the experiments. As a part of the training, we chose to use *linear kernel* of the SVM in order to alleviate the problem of higher dimensions such as over-fitting. We used different kernels depending on the feature set, for example linear kernel for acoustc, lexical and their combination and used a gaussian kernel with the psycholinguistic feature set. The reason of using the SVM is that in [163], we found its success compared to two other state-of-the-art machine learning algorithms such as Random Forest and Adaboost.

For the segmentation and classification approach, we employed Hidden Markov Models (HMMs) using *kaldi*, which is an open-source implementation toolkit for ASR [156]. While doing that, we used ASR pipeline by building a lexicon of emotion labels, a grammar using unigrams. During training the system, we optimized the following parameters using the development set:

- number of states in HMM topology,
- total number of gaussians,
- beam width,
- acoustic model score, and
- language model score.

4.5 Decision Fusion

Regarding the classifiers trained on different feature sets, we combined classifiers' decisions by applying *majority voting*, as shown in Equation 4.3.

$$H(s) = c_{\hat{j}}; \quad where\ \hat{j} = argmax_j \sum_{i=1}^{T} h_i^j(s) \tag{4.3}$$

where $H(s)$ is the combined classification decision for an input instance s; $h_i^j(s)$ is the output of the classifier h_i for the class label c_j; $i = 1...T$ is the number of classifiers; $j = 1...C$ is the number of classes.

4.6 Evaluation

In our study, we have different experimental settings such as 1) classification at the conversation level, 2) segment level, and 3) segmentation and classification using a generative approach. Each such setting required different performance metrics and different error estimation methods. In the following subsections, we discuss the evaluation approaches for the former two. In Chapter 7, we will discuss the evaluation of the segmentation and classification.

Conversation level: To measure the performance of the system at the conversation level experiments, we used and Un-weighted Average (UA), as shown in the equation 4.4.

$$UA = \frac{1}{2}\left(\frac{tp}{tp + fn} + \frac{tn}{tn + fp}\right) \tag{4.4}$$

where tp, tn, fp, fn are the number of true positives, true negatives, false positives and false negatives, respectively. It has been widely used for the evaluation of the paralinguistic task [102].

In order to estimate the performance, for the conversation level classification, we used 10-fold cross-validation method for the customer side emotion. Whereas for the agent side emotional state we used Leave-One-Speaker-Group-Out (LOSGO) cross-validation method since we had speaker information. The reason to use LOSGO is due to the limited size of the conversational dataset and the skewed distribution of the agents. In LOSGO, for each fold, we included a) agent's spoken conversation-side features, b) a random selection of conversations, and c) a class label distribution close to the corpus empirical distribution.

Segment level: For the segment level experiments, the evaluation procedure requires the alignment of the manual segment and labels with the output of the automatic segmentation and classification system. In Figure 4.1, we show a sample alignment of the reference (manual) and automatically generated segments and their labels. The reference segmentation spans from $t = 0$ to $t = t_e$ and labels the *Neutral* (N) segment spanning from $t = 0$ to $t = t_i$ and the *Empathy* (E) segment from $t = t_i$ to $t = t_e$. Automatic segments inherit the reference label that falls inside its boundaries (e.g., the segment spanning the interval $[0, t_1]$ or $[t_3, t_4]$). For the evaluation purpose, automatic segments that span across the onset, $t = t_i$, and end, $t = t_e$, (e.g., the segment spanning the interval $[t_2, t_3]$) are split in two segments with two distinct reference labels. For instance the segment spanning $[t_2, t_3]$ will be evaluated with the segment $[t_2, t_i]$ (reference label N) and the segment $[t_i, t_3]$ (reference label E). The alignment process will generate the correct decision statistics for all segments as shown on the last row of Figure 4.1.

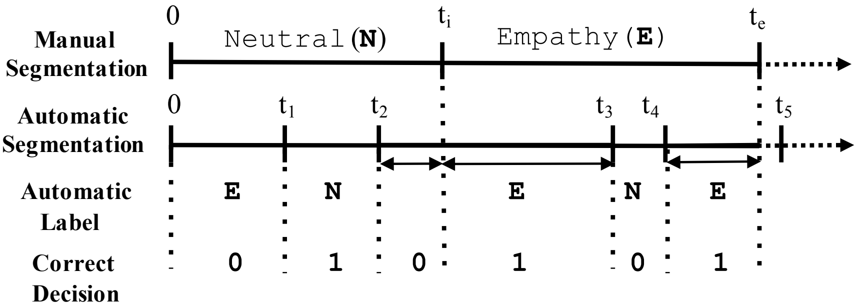


Figure 4.1: Sample alignment of the manual and automatic segments and their labels. The evaluation spans from $t = 0$ to $t = t_e$, the end of the *Empathy* segment. Automatic labels are evaluated against the reference labels and error statistic are computed (last row).

For the segment level experiment, we also used UA as a performance metric. We have extended such measure to account for the segmentation errors as evaluated in similar cases by NIST in diarization tasks [166,167]. However, here, the UA from a weighted confusion matrix, as shown in Equation 4.5, where the weight for each instance is the corresponding segment length:

$$C(f) = \left\{ c_{i,j}(f) = \sum_{s \in S_T} [((y = i) \wedge (f(s) = j)) \times length(s)] \right\} \tag{4.5}$$

In Equation 4.5, $C(f)$ is the $n \times n$ confusion matrix of the classifier f, s are the automatic segments in the test set S_T, including the segments with boundary t_i and t_e (see Figure 4.1), $length(s)$ is the duration of s, y is the reference label of s, $f(s)$ is the automatic label for s. The indices i and j represent the reference and automatic class label of the confusion matrix.

For the performance estimation of the segment level classification, we have maintained training, development and test split (see Section 3.1.5.2). Using this setting, we optimized the parameters on the development set and obtained the final results using the test set.

4.7 Summary

In this chapter, we presented the feature extraction, selection, combination, classification and evaluation methods, that are used for designing and evaluating the classification models for affective behavior for both conversation and segment levels. The feature extraction process include, extracting acoustic, lexical, and psycholinguistic features. The acoustic feature set include spectral, voice quality and prosodic features. For the lexical features, we extracted tri-grams and used Bag-of-ngram model, in which we removed stop words and transformed the feature values with tf-idf. Psycholinguistic features are knowledge based word categories, which are extracted using the dictionary of LIWC. The extracted feature vectors for acoustic and lexical became very large, which increased the computational complexity. Moreover, higher number of features are not useful in many cases as most of the classification algorithms are not able to deal with redundent or irrelavant features. Therefore, to reduce the dimension, we applied feature selection. We also investigated different, feature and decision level combination in order to obtain a combined decision. As a classification algorithm we employed the SVM with its different kernels for different feature sets. For the evaluation of the classification systems, we used different evaluation methods, which we discussed in detail.

Chapter 5

Empathy Classification

The importance of automatic classification of empathy has been highlighted in [40, 168] where behavioral analysis experiment has been conducted by human experts in workplaces such as the call centers to evaluate the agent's empathic behavior during the phone conversations. In literature, there are only a few studies, which focused on the automatic classification of therapist's empathy, however, it remains unexplored in the other areas such as agent's empathy in call center.

In this chapter, we present our study of automatic classification of agent's empathy using real call-center dyadic conversations. For the classification experiment, we investigated both conversation level as well as segment level performance of the automatic system.

5.1 Background

Over the past decades there have been significant efforts in investigating empathy in the fields of psychology and neuroscience. The complexity of the neural and psychological phenomena to be accounted for is huge and, in part, that complexity explains the existence of several psychological definitions of empathy. For example, the work in [169] accounts for different empathic phenomena occurring in the literature on empathy, and [170] examines eight distinct phenomena commonly labeled as empathy including emotional contagion, sympathy, projection, and affective inferential processes. Decety and Lamm [171] observe that some of the different definitions of empathy may share the underlying concept of "[...] an emotional experience that is more congruent with another's situation than with one's own". The authors also state that empathic emotional experiences imply self-other differentiation, as opposed to emotional contagion. Actually, if we abstract from the differences in the theoretical perspectives, we may find some common features. Most of the definitions describe empathy as a type of emotional experience and/or emotional state. Moreover, the different definitions can be divided into two main classes. One encompasses the cognitive aspects of empathic behavior, such as one's ability to accurately understand the feelings of another person. The other class entails sharing or the subject's internal mimic of such feelings such as sympathy, emotional contagion, and compassion.

Computational models of emotional states are needed to design machines that can understand and interact *affectively* with humans. Different signal components have been considered for analyzing the manifestation of emotional experience in speech. Both verbal and vocal non-verbal levels, such as paralinguistic features, of spoken communication [172] have been considered since both are suggested to embody the expressive potential of language. Major focus has been devoted to the paralinguistic features of emotional speech, on the basis of the experimental evidence that emotional information is mostly conveyed by those levels (see [173] for a state-of-the-art review).

Providing explicative models for annotating the emotional process itself in naturally occurring conversations is still an open challenge. Efforts in this direction are currently being made in the affective computing research, where awareness about the need for continuous annotation is increasing. The models that foster this approach, such as those discussed in [150] and [174], require annotators to continuously assign values to emotional dimensions and sets of emotional descriptors. Metallinou and Narayanan [149] emphasize that continuous annotation has several benefits, as well as some open challenges. One interesting finding is that continuous annotation may show regions which are characterized by perceived transitions of the emotional state. In [149], authors report a high number of factors that may affect inter-annotator agreement, such as person-specific annotation delays and confidence in understanding emotional attributes. There are very few studies in terms of empathy classification and most of them are carried out within controlled scenarios. Kumano et al. [175] studied four-party meeting conversations to estimate and classify empathy, antipathy and unconcerned emotional interactions utilizing facial expression, gaze and speech-silence features. In [176] and [177], Xiao et al. analyzed therapists' conversation to automatically classify empathic and non-empathic utterances using linguistic and acoustic information.

5.2 Conversation Level Classification

The motivation of conversation level experiment was to understand whether segment level information can be used to detect the presence of an emotional state in a conversation.

5.2.1 Experimental Methodology

Classification Task

For the conversation level binary classification experiments, we used a subset of the corpus, which contained a total of 905 conversations. We have chosen this subset because we have full manual transcriptions for this set, and also we have performed complete acoustic and lexical performance analyses. For the experiments, we designed binary classifiers. In order to define class labels, the conversations containing at least one empathic segment were considered as positive and rest of the conversations were considered as negative as can be seen in Figure 5.1. We labeled 302 empathic conversations (33.30%) containing empathic segment(s) as positive examples and 603 non-empathic conversations (66.60%) as negative examples.

Classification System

In Figure 5.2, we present a computational architecture of the automatic empathy classification system, which takes the agent's speech channel as input and generates a binary decision regarding the presence (absence) of

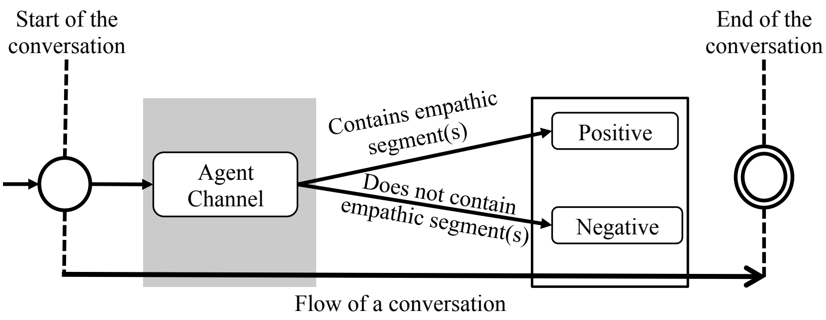

Start of the conversation
End of the conversation
Agent Channel
Contains empathic segment(s)
Does not contain empathic segment(s)
Positive
Negative
Flow of a conversation

Figure 5.1: Data preparation for the conversation level classification.

empathy in the agent's behavior. The system evaluates the cues present throughout the spoken conversation and then commits to the binary decision. To evaluate the relative impact of lexical features, we considered the case of clean transcriptions of the conversation (right branch in Figure 5.2) as well as the case of noisy transcriptions (left branch in Figure 5.2) provided by an automatic speech recognizer (ASR). We extracted, combined, and selected acoustic features directly from the speech signal and generated the classification system's training set. We implemented both feature and decision fusion algorithms (bottom part of Figure 5.2) to investigate the performance of different system's configurations. The details of the ASR system is discussed in Section 3.1.2.

Feature Extraction, Selection and Combination

In Section 4.1, we presented how feature extraction has been done for different classification studies in this thesis, where we report how we extracted large-scale acoustic features. The utilization of such acoustic features was inspired by previous studies in emotion and personality recognition tasks, in which low-level features were extracted and then projected onto statistical functionals [133, 160]. For this study, we extracted features using openSMILE [128]. Before extracting features, we automatically pre-processed speech signals of the conversations to remove silence at the beginning and end of the recordings. We also removed silences longer than one second.

We extracted and categorized features into four different groups, voice-

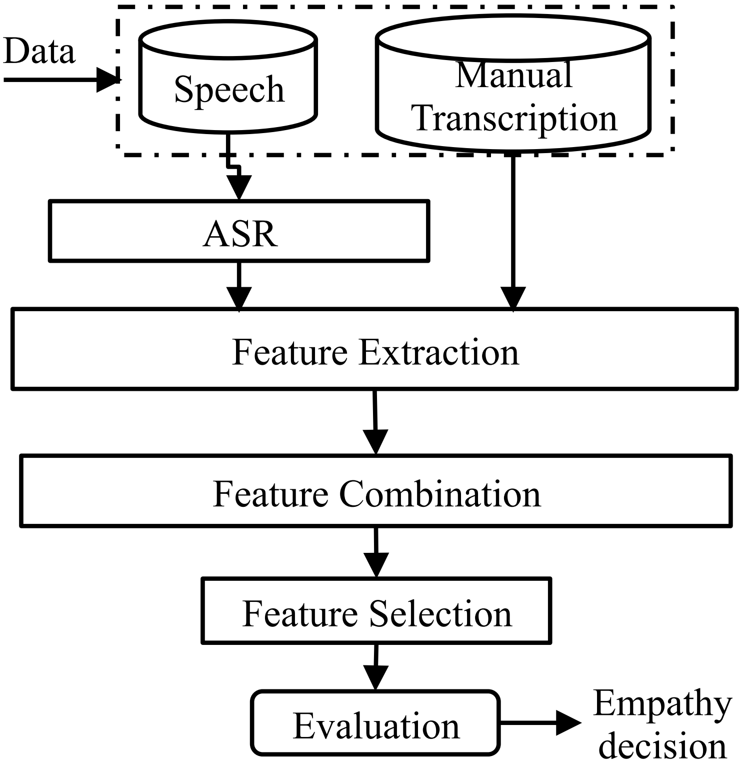


Figure 5.2: System architecture of the conversation level classification.

quality, cepstral, spectral, and prosody together with the list of statistical functionals as presented in Table 5.1. It was recently shown that grouping the acoustic features followed by feature selection improves the performance of the classification [178]. We thus grouped a large set of acoustic features. In addition to the feature set defined in [179], which has 130 low-level features including first-order derivatives, we also used formants features, constituting 150 low-level features in total. We extracted low-level acoustic features at approximately 100 frames per second. For the voice-quality features the frame size was 60 milliseconds with a gaussian window function and $\sigma = 0.4$. A frame size of 25 milliseconds with a hamming window function was used for the other low-level features.

The lexical features were extracted from both manual and automatic transcriptions. To utilize the contextual benefits, we extracted trigram features, which eventually results in a very large dictionary, removed the stop-words

Table 5.1: Extracted group-wise low-level acoustic features and statistical functionals

Low-level acoustic features
Voice Quality
Probability of voicing, jitter-local, jitter-DDP, shimmer-local, log harmonics-to-noise ratio (HNR)
Cepstral
MFCC 1-14
Spectral
Auditory spectram (RASTA-style) bands 0-25 (0-8kHz),
Spectral energy 250-650Hz, 1-4kHz,
Spectral roll-off points (0.25, 0.50, 0.75, 0.90),
Spectral flux, centroid, entropy, variance, skewness, kurtosis,
Spectral slope,
Psychoacoustic spectral sharpness, harmonicity
Prosody
F0 final, F0 envelope, F0final with non-zero frames,
Root-mean-square signal frame energy,
Sum of RASTA-style auditory spectra,
Loudness, Zero crossing rate,
Formant frequencies [1-4], bandwidths [1-4]
Statistical functionals
Percentile 1%, 99% and percentile range 1%-99%
Quartile (1-3) and inter-quartile (1-2, 2-3, 3-1) ranges
Relative position of max, min, mean and range
Arithmatic mean, root quadratic mean
Mean of non-zero values (nnz)
Contour centroid, flatness Std. deviation, skewness, kurtosis
Uplevel time 25, 50, 75, 90,
Rise time, fall time, left curvature time, duration
Mean, max, min and Std. deviation of segment length
Linear prediction coefficients (lpc0-5), lpc-gain
Linear regression coefficients (1-2) and error
Quadratic regression coefficients (1-3) and error

and selected top-ranked features, discussed in Section 4.1.2.

Psycholinguistic features were extracted using Linguistic Inquiry Word Count (LIWC), which comprised of 102 features. More details can be found in Section 4.1.3.

In out feature sets we have a large number of features for both acoustic and lexical sets. In order to reduce the computational cost and avoid overfitting we have chosen Relief [158] as a feature selection technique as discussed in Section 4.2. For the feature fusion, we merged acoustic and lexical features into a single vector to represent each instance (see Section 4.3).

Classification and Evaluation

In this study, we designed binary classification models using Support Vector Machines (SVM) [136]. We chose the linear kernel in order to alleviate the problem of higher dimensions of lexical and combination of *acoustic+lexical* features. We used a gasussian kernel with different groups of acoustic feature sets and psycholinguistic features as it performed better with the small-sized feature set. We optimized the penalty parameter C of the error term by tuning it in the range $C \in [10^{-5}, ..., 10]$ and the gaussian kernel parameter G in the same range as well, using cross-validation.

At the feature fusion level, we applied feature selection on the combined acoustic and lexical features. For the decision fusion, we combined decisions from the best classifiers of three different feature sets by applying *majority voting*. In the experiment with acoustic features, we first applied feature selection for each group, then merged the feature vectors into one single vector. We then re-applied the feature selection process to the merged feature vector to obtain an optimal subset from all groups.

We measured the performance of the system using the Un-weighted Average (UA) and also used the Leave-One-Speaker-Group-Out (LOSGO) cross-validation method, as discussed in Section 4.6.

Table 5.2: Empathy classification results at the conversation level using acoustic, lexical, and psycholinguistic (LIWC) features together with feature and decision level fusion. Ac - Acoustic Features; Lex (M) - Lexical features from manual transcriptions; Lex (A) - Lexical features from automatic transcriptions; Ac+Lex - Linear Combination of Acoustic and Lexical Feature; Maj - Majority voting; LIWC (M) - Psycholinguistic features extracted from manual transcriptions; LIWC (A) - Psycholinguistic features extracted from automatic transcriptions. Dim. - Feature dimension.

Experiments	**Dim.**	**UA-Avg**	**UA-Std**
Random baseline		49.7	2.2
Ac	200	61.1	4.3
Lex (M)	5000	63.5	5.5
Lex (A)	3800	62.3	5.3
LIWC (M)	89	63.4	4.8
LIWC (A)	89	62.9	4.1
Ac+Lex (M)	6800	62.3	5.9
Ac+Lex (A)	6600	60.0	4.4
Maj: {Ac,Lex(M),LIWC(M)}		**65.1**	6.2
Maj: {Ac,Lex(A),LIWC(A)}		**63.9**	4.5

5.2.2 Results and Discussion

In Table 5.2, we report the performances of the classification system for a single feature type, feature combination, and classifier combination. We report them in terms of average UA of the LOSGO cross-validation and its standard deviation. We computed the baseline by randomly selecting the class labels, such as empathy and non-empathy, based on the prior class distribution of the training set.

In Table 5.2, we present that the system trained on lexical features extracted from manual transcriptions outperformed any other system trained on single feature type. The features from the ASR transcriptions outperformed all automatically extracted features, including the acoustic-only system, *Ac*.

The results of the acoustic feature are better than random baseline, which were statistically highly significant with $p - value < 0.001$. The value 0.001 refers to the significance level with statistically highly significant. It provides a useful label prediction, when no transcriptions are available. We obtained better results with *majority voting*. The statistical significance test between Lex, and $Maj(A)$ revealed that the performance improvement of $Maj(A)$ were statistically significant with $p - value < 0.05$. We performed significance test using paired t-test over the set, where each set contains 10 LOSGO cross-validated estimates. Compared to the baseline, the best model for automatic classification provides a relative improvement over the baseline of 31%. In addition, all systems' results are higher and statistically highly significant with $p - value < 0.001$ compared to the baseline results. Linear combination of lexical with acoustic features in the $Ac + Lex(M)$ and $Ac + Lex(A)$ systems did not provide statistically significant change in performance. Despite its success in other paralinguistic tasks [160], the linear combination of the feature space does not necessarily provide improved performance even when combined with feature selection.

The results of the psycholinguistic feature set indicate its usefulness using which we obtained a comparable performance compared to other feature sets. Some of the distinguishing features of this feature set are perceptual e.g., feel and cognitive e.g., certainty, which are ranked using relief feature selection technique. From the investigation of acoustic features, our findings suggest low-level spectral, F0-envelope and MFCC features contribute most to the classification decision, whereas the higher-level statistical functionals are peak and regression (linear and quadratic) coefficients.

5.3 Segment Level Classification

In many real applications, we need to find the answer of *which* emotion manifested *when*. In other terms, finding the emotional segments of interlocuators during the time span of dyadic conversations, which unfolds over time. This problem leads to a proper understanding the smallest segmental unit or *time course* of discrete emotional meanings [180]. In [173], Schuller et al. presented three important reasons of finding appropriate segmental unit for emotion recognition, such as 1) it will help for optimal classification, 2) for incremental processing and 3) for multi-modal processing. From the computational perspective, it is evident in the literature that there are various challenges to automatically segment emotional episodes. It includes identification of a smallest unit of segment for emotional episodes, static *vs* dynamic classifier, verbal and/or non-verbal information and data from a real-life scenario [122, 181, 182]. Other computational challenges include acoustic variability introduced by the existence of different, speakers, speaking styles, speaking rates, environment and channel distortions. Even though segmenting emotion units is one of the important aspects in all modalities [183], however, it has been remained unexplored.

Compared to the research on the general problem of emotion recognition at the utterance level, there are only very few contributions on segmenting or spotting emotion in a conversation, in continuous time space, and it remains an open research issue [122, 125]. It is evident in the literatures that researchers have been investigated several approaches to segment emotional episodes such as 1) phoneme as a segment [93, 124], 2) segment based on voiced speech [125], 3) forced aligned word as segment, 4) syntactic chunks, 5) ememe chunks [122], 6) utterances based and 7) regions-of-interest. In [125] Shami et al. worked on segmenting emotions using global statistical features, where they used KISMET emotional corpus and compared the performance of SVM and K-NN by designing the speaker dependent classifiers. They present that classification accuracies increases 5% when segment-level decisions are integrated compared to the utterance-level feature generation approach. Batliner at el. in [122], attempted to find and define the emotion units based on speech and coined the term “ememe" to define smallest unit of emotion, where a “word” is considered as smallest meaningful emotional unit. In [126], they used a segmentation method to extract a sequence of voice segments and then recognize each segment individually, using acted emotional dataset. They also compared different segmentation, feature selection and classification methods. Jeon et al. proposed an two-step approach where they utilizes sub sentence segments' decision to obtain sentence level global decision [184], however, segmentation process is not yet automatic. Kim at el. in [185], proposed a real-time emotion detection system where they fused decision from intra- and supra- frame level systems, and argued that their multi-modal fusion system outperform individual system. In [186], Mansoorizadeh at el. investigated frame vs voised based segmentation approach, and report that recognition accuracy is better when speech segments are longer or there are 10-12 voiced speech segments.

There are many application areas where it would be useful. These include emotional-monitoring, emotional-mirror, emotional-chat and behavioral-agent (see [17, 133]). To be very specific *in call center application*, it is necessary to provide customer a real time feedback, to guide the call center agent or give daily summary to the customer care manager. It is also necessary for incremental processing in real applications. In addition, another uses of this task would be spoting emotional segments in the whole conversation, which unfolds over time.

Hence, the importance and wide applicability of segment level emotion classifcation leaded us to analyze data and the design of automatic computational model to takle such as issue.

5.3.1 Qualitative Analysis

We have performed qualitative analysis of the spoken conversation corpus to gain insights into the manifestation of basic, complex emotions and empathic phenomena in affective scenes. The analysis was carried out over a corpus of human-human dyadic Italian call center conversations that will be discussed in Section 5.3.2. We analyzed one hundred conversations (more than 11 hours), and selected dialog turns where the speech signal showed the emergence of the set of basic and complex emotions (e.g. frustration, anger) and empathy. We evaluated the communicative situation in terms of appraisal of the transition from a neutral to an emotionally-connoted state.

In Table 5.3, we present a dialog excerpt with annotations to further illustrate the paralinguistic, lexical and discourse cues. The dialog excerpt is reported in the first column of the table, where **C** is the customer, and **A** is the agent. The situation is the following: **C** is calling because a payment is actually overdue: he is ashamed for not being able to pay immediately and his speech has plenty of hesitations. This causes an empathic response by **A**: that emerges from the intonation profile of **A**'s reply and from her lexical choices. In the second question of **A**'s turn, she uses the hortatory first person plural instead of the singular one. Also, the rhetorical structure of **A**'s turn, i.e., the use of questions instead of assertions, conveys her empathic attitude. The annotator perceived the intonation variation and marked the speech segment corresponding to the intonation unit starting with the word *proviamo* (*let us try*) as onset of the emotional process.

The outcome of the qualitative analysis has supported the view that emotionally relevant conversational segments are often characterized by significant transitions in the paralinguistic patterns or the occurrence of lexical cues. As expected, such variations may co-occur not only with emotionally-connoted words but also with functional parts of speech (POS) such as Adverbs and Interjections. Phrases and Verbs, as shown in Table 5.3, could also lexically support the expression of emotional states.

Table 5.3: An excerpt from a telephone conversation where the agent (**A**) is empathic towards a customer (**C**). The agent perceives the customer's feeling and proactively takes actions to cope with customer's emotional discomfort. English translations are in italics.

Dialog excerpt	Notes
C: Ascolti ... io ho una fattura scaduta di 833 euro vorrei sapere ... tempo in cui posso pagarla. *(Listen... I have an 833 euros overdue bill... I would like to know... the time left to pay it.)*	The tone of voice and the hesitations of the customer show that she is not angry, she is ashamed for not being able to pay immediately the bill. This causes an empathic reply in the Operator's attitude.
A: Ma perché non ha chiesto il rateizzo di questa fattura? Proviamo a far il rateizzo, ok? Così gliela blocco e lei ha più tempo per effettuare il pagamento. *(Why did not you ask to pay it in installments? We try to divide it into installments, is it ok for you? So I stop the overdue notices and you will have more time to pay)*	The selection of the speech act (question instead of authoritative declarative), the rhetorical structure of the second question, the lexical choice of "proviamo", instead of - for instance, "adesso provo a vedere...", all these contribute to prevent the customer's feeling of being inadequate or ashamed.

5.3.2 Data Analysis

Acoustic Feature Analysis

We investigated and compared the pattern sequences of low-level acoustic features before and after the onset point, from the neutral to the empathy segment. An example of the speech segment annotation is shown in Figure 5.3, where we plot the spectral centroid feature values across the neutral and empathy connoted segment. Each segment is 15 seconds long and the onset is marked by a vertical bold line, which separates the left (context) and right segment annotated with empathy. From the signal trend of this feature we see that there is a distinctive profile change, which is corroborated by its high statistical significance (p-value=4.61E-51, using *two-tailed two-sample t-test*).

The low-level features were extracted from both left and right segment

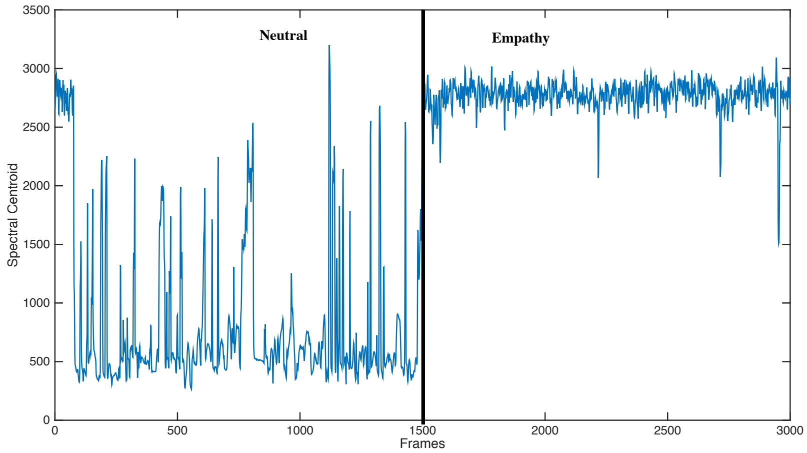


Figure 5.3: Spectral centroid (in Hz) of an annotation unit. The onset is marked in bold. The neutral (left) segment is the context support preceding the right segment annotated with the empathy label (in this case a female speaker). Both segments are 15 seconds long.

with 100 overlapping frames per second, pre-emphasis with k=0.97, and hamming-windowing, using openSMILE [128]. For voice quality features we used gaussian windowing function. Then, we computed averages for each segment of the corresponding conversation. In order to evaluate the relevance of each feature we applied a statistical significance test, the *two-tailed two-sample t-test* at p-value = 0.01. We analyzed 45 low-level acoustic features from five categories: pitch (4), loudness (1), zero-crossings (1), spectral (13), and auditory-spectrum bands (26). In Table 5.4, we list the acoustic features that passed the significance test for male (40 features) and female (34 features) speakers. We also report effect sizes, d, which are computed using Cohen's d, as in Equation 5.1. The number of samples for this analysis is 302 conversations, $n = 302$, sample size.

$$Cohen's\ d = \frac{m1 - m2}{\sigma^p} \ \ ; \ \sigma^p = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}} \tag{5.1}$$

where $m1$ and $m2$ are the means of two samples, and σ_1 and σ_2 are the standard deviations of the two samples.

Table 5.4: Statistical significance (two-tailed two-sample t-test) of acoustic features is reported for each category (rows). In the second and third column we report largest *p-value* and the range of effect size (Cohen's d) for female and male speakers, respectively. The value of d, 0.2, 0.5 and 0.8 denotes small, medium and large effect sizes, respectively.

Feature type	**Female**	**Male**
Pitch (F0, voice-probability, voice-quality)	p<=6.04E-06 d=[0.3-1.1]	p<=6.73E-03 d=[0.2-1.5]

Loudness	p=3.08E-25
	d=0.7
		p=1.86E-29
		d=1.1

Zcr	p=4.81E-08
	d=0.3
		p=2.01E-05
		d=0.4

Spectral (Energy in bands: 0-650 Hz, 250-650Hz, 1-4kHz; Roll-off points 25%, 50%, 75%, 90%; Position of spectral maximum and minimum, Centroid, Flux)	p<=5.66E-03 d=[0.2-1.4]	p<=3.64E-04 d=[0.2-1.8]
Auditory-spectrum bands 0-25 for male and 0-21 for female	p<=6.25E-04 d=[0.2-0.7]	p<=1.26E-04 d=[0.3-1.0]

From our analysis we observe that the pitch patterns are higher in non-empathic segments. The spectral features such as centroid and flux are more stable and smooth when the agent is empathic compared to abrupt changes in non-empatic segments. Spectral patterns captures the perceptual impression of sharpness of sounds. The vocal pattern of loudness is higher in non-empathic situations while it is low when the agent is empathic. We also observed that the auditory-spectrum bands of non-empathic segments were comparatively higher. Our analysis on the relevance of pitch and loudness for empathy signal realization is consistent with the findings of [187] and [177]. There are no significant differences in the relevance of features in Table 5.4 for male and female speakers.

Linguistic Feature Analysis

Several categories of personnel who interact with customers or patients, including call center agents and physicians, are trained to improve their communication skills and develop empathy in their interactions through careful choice of words [188,189]. For example, they are recommended to use phrases such as *"I understand"* when they are listening to the customer who is explaining their problem, or to use *"I would"*. For example, after hearing a customer's story, agent may respond by saying *"I would be upset as well if I were in a similar situation"*, before proceeding to propose possible solutions or provide advice.

In our call center corpus, we analyzed the lexical realization occuring in empathic and neutral segments by comparing the different word frequencies and POS distributions of unigrams, bigrams and trigrams respectively.

We tested the statistical significance over the observed differences with a *two-tailed two-sample t-test* and a p-value of 0.01 with the same number of conversations ($n = 302$) we used for the analysis of acoustic features.

The comparison between neutral and empathy word trigrams showed that agents' phrases such as *vediamo un po'* (*let's see a bit*), *vediamo un attimo* (*let's see a bit*), *vediamo subito allora* (*let's see now, then*) are statistically significant cues for the agent while interacting and manifesting their empathy. It is worth noticing that the Italian verb *vedere* (*see*) is more frequently used in the first person plural; the same holds true for other frequent verbs such as *facciamo* (*let's do*) and *controlliamo* (*let's check*). Those lexical choices are usual when the speaker cares about the problem of the other person. Similarly, significantly different rankings in the empathic distribution affect unigrams, bigrams and trigrams such as *non si preoccupi* (*do not worry*), *allora vediamo* (*so, let's see*) that are often used in Italian to grab the floor of the conversation and reassure the other person. For the above lexical features p-value was < 0.01 and the range of effect sizes, d, was $[0.7 - 3.8]$, as shown in Table 5.5.

Table 5.5: Statistical significance (two-tailed two-sample t-test) of lexical features.

Lexical Features	p-value	d (effect size)
vediamo un po' *(let's see a bit)*	1.6E-03	1.0
vediamo un attimino *(let's see a bit)*	2.7E-03	1.7
vediamo subito allora *(let's see now, then)*	9.7E-03	3.8
vedere *(see)*	3.3E-05	1.1
facciamo *(let's do)*	6.1E-07	1.3
controlliamo *(let's check)*	1.2E-03	0.7
non si preoccupi *(do not worry)*	5.0E-03	1.1
allora vediamo *(so, let's see)*	4.3E-04	0.9
assolutamente *(absolutely)*	4.2E-06	1.8

Regarding the POS distributions, the Adverbs that occur frequently in the empathy distribution, such as *assolutamente* (*absolutely*) and *perfettamente* (*perfectly*), may have a kind of evocative potential for showing understanding of the other person's point of view, in particular when they are uttered with a tone of voice appropriate to the context.

5.3.3 Experimental Methodology

In this Section, we describe the training of an automatic classification system for the recognition of empathy from spoken conversations. We report experimental details of the feature extraction, fusion and classification task and discuss the results.

Classification Task

In the automatic classification experiment, our goal was to investigate the segment level operator's empathic manifestations. We have selected a subset of the corpus, that includes a total of 526 conversations annotated with automatic speech transcriptions as well as `Neutral, Empathy` segment labels. This subset of the corpus allows us to perform a complete computational model training and evaluation in noisy and clean input signal conditions. We partitioned the data-set into train, development and test with 70%, 15%

and 15% partitions and no-speaker overlap amongst them. In order to train and evaluate the system we extracted neutral-empathic segment-pair from these conversations, which has been obtained from the manual annotation.

Table 5.6: Segment duration statistics (seconds) of the `Neutral` and `Empathy` segment-pairs and the total amount of speech for each category (hours).

Class	**Avg. (s)**	**Std. (s)**	**Total (h)**	**# of Seg**
`Empathy`	19	13	3	526
`Neutral`	220	148	32	526

Duration-based descriptive statistics of these segment pairs are provided in Table 5.6 along with averages and standard deviations on the natural distribution of the data. The segment length of neutral is comparatively longer than the empathic segment as we see in the Table 5.6, since it spans from the start of the conversation until the onset of the first empathic event. The total net duration of these segment-pairs is approximately 35 hours.

Classification System

In Figure 5.4, we present a computational architecture of the automatic classification system, which takes the agent's speech channel as input, then pass it to the automatic speech *vs* non-speech segmenter (see Section 3.1.3). After that it generates a binary decision for each speech segment of the agent's behavior in terms of neutral *vs* empathy. In order to evaluate the relative impact of lexical features we considered the case of noisy transcriptions (left branch in Figure 5.4) provided by an automatic speech recognizer (ASR). We extracted, combined, and selected acoustic features directly from the speech signal and generated the classifier's training set. We implemented both feature and decision fusion algorithms (bottom part of Figure 5.4) to investigate the performance of different classifier configurations. This architectural design can be used in real time application, which combines all automatic processes.

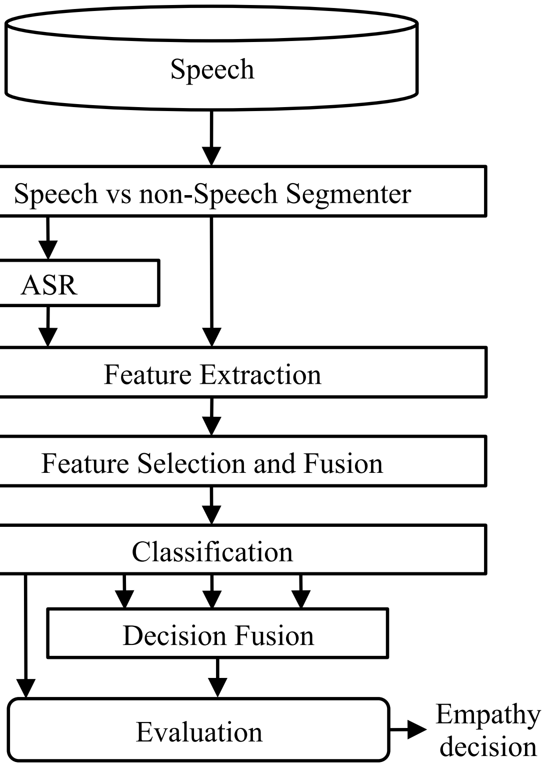


Figure 5.4: The segment level classification system.

Undersampling and Oversampling

The statistics in Table 5.6 manifest the data imbalance problem for the two classes, `Empathy` and `Neutral`. Once the manual segments are processed through the automatic segmenter, the ratio of `Empathy`/`Neutral` labelled segments is 6% *vs* 94%. The mapping between manual and automatic segments leads to three different type of mismatch in the segment boundaries as presented in Figure 5.5. We solved the mismatched segment boudaries using a rule-based approach presented in Algorithm 1.

A common approach to cope with imbalance class distribution is via oversampling or undersampling in the data or feature space. We have undersampled the instance of the majority class at the data level and oversampled the minority class at the feature level as presented in Figure 5.6. In the literature it is reported that the combination of oversampling and undersampling often leads to a better performance [190].

Algorithm 1 Pseudocode to align manual and automatic segment for each segment-pair with emotion label. $threshold = 0.5$

```
Input: $aSegmentList = automatic\ speech\ non\ speech\ segments$
Input: $mSegmentList = manual\ segments$
Output: $alignedSegmentList$
  procedure $alignment(mSegmentList, aSegmentList)$
    $i \leftarrow 0$
    for all $mSeg \in mSegmentList$ do
      for all $aSeg \in aSegmentList$ do
        if $aSeg.startTime <= mSeg.startTime$ & $aSeg.endTime >= mSeg.endTime$ then
          $newSeg \leftarrow createSeg(aSeg, mSeg)$
          $alignedSegmentList[i++] \leftarrow newSeg$
        else if $(aSeg.startTime <= mSeg.startTime$ &
            $aSeg.endTime <= mSeg.endTime$ &
            $aSeg.endTime >= mSeg.startTime$ &
            $(aSeg.startTime - mSeg.startTime) >= (aSeg.endTime - aSeg.startTime) * threshold)$ then
          $newSeg \leftarrow createSeg(aSeg, mSeg)$
          $alignedSegmentList[i++] \leftarrow newSeg$
        else if $(aSeg.startTime >= mSeg.startTime$ &
            $aSeg.startTime <= mSeg.endTime$ &
            $aSeg.endTime >= mSeg.endTime$ &
            $(mSeg.endTime - aSeg.startTime) >= ((aSeg.endTime - aSeg.startTime) * threshold)))$ then
          $newSeg \leftarrow createSeg(aSeg, mSeg)$
          $alignedSegmentList[i++] \leftarrow newSeg$
        else if $(aSeg.startTime >= mSeg.startTime$ & $aSeg.endTime <= mSeg.endTime)$ then
          $newSeg \leftarrow createSeg(aSeg, mSeg)$
          $alignedSegmentList[i++] \leftarrow newSeg$
        end if
      end for
    end for
  end procedure
  procedure $createSeg(aSeg, mSeg)$
    $newSeg.setContent(mSeg.label)$
    $newSeg.setStartTime(aSeg.startTime)$
    $newSeg.setEndTime(aSeg..endTime)$
    return $newSeg$
  end procedure
```

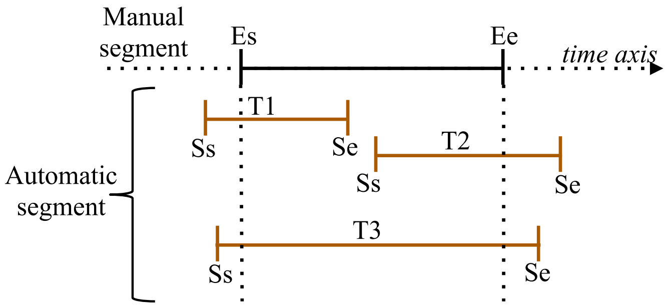


Figure 5.5: Type of mismatch between manual and automatic segment boundaries. E_s and E_e refer to start and end of manual emotion segment boundaries, respectively. S_s and S_e refer to start and end of automatic segment boundaries, respectively.

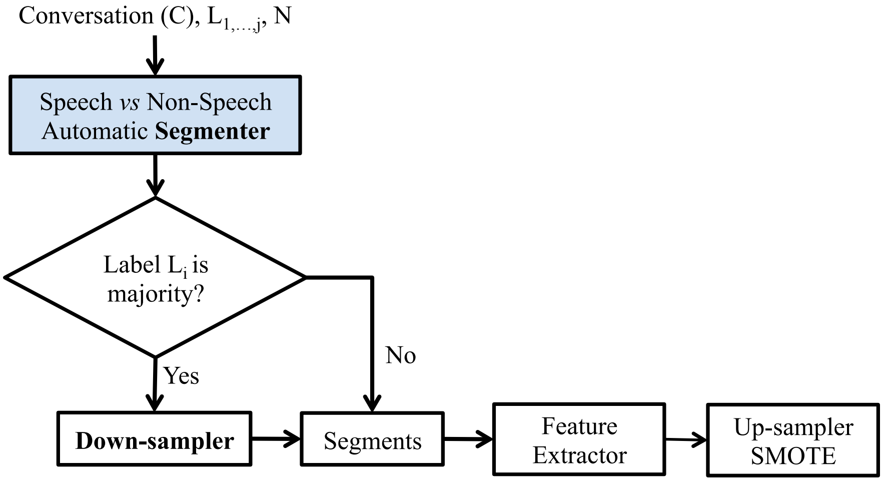


Figure 5.6: System architecture for undersampling and oversampling approaches. Undersample the instances of majority class and oversample the instances of minority class.

For undersampling, we defined a set of bins with different segment lengths, and then randomly selected K segments from each bin. We used $K = 1$ for this study. The number of bin and size of K has been optimized empirically on the development set and by investigating the descriptive statistics such as percentiles, mean and standard deviation. The undersampling stage generated a 18% *vs* 82% ratio of `Empathy` *vs* `Neutral` segments up from the initial 6% *vs* 94%.

For oversampling we used Synthetic Minority Oversampling Technique (SMOTE) [190] and its open-source implementation in weka [164]. In SMOTE, the oversampled examples are generated based on the K nearest neighbors of the minority class. Nearest neighbors have been chosen randomly based on the percentage of target oversampling. It computes the difference between the feature vector and its nearest neighbor. Then, multiply this difference by a random number between 0 and 1 and add it to the feature vector. More details of this approach can be found in [190]. The oversampling was tuned on the development set and we achieved a further improvement on the imbalance problem. Before oversampling the class distribution was 18% *vs* 82%, and after the oversampling it became 30% *vs* 70%.

Feature Extraction, Selection and Fusion

For the segment level classification experiment, we extracted the similar set of acoustic, lexical and psycholinguistic features discussed in Chapter 4.

Classification and Evaluation

We designed our classification models using an open-source implementation of SMO [165] by the Weka machine learning toolkit [164], with its linear kernel for lexical and acoustic features, and its gaussian kernel for the psycholinguistic features. We chose SMO for its high generalization performance and used the linear kernel in order to alleviate the problem of higher dimensions of acoustic, lexical and combination of $acoustic + lexical$ features. We used gaussian kernel with psycholinguistic features as it performed better with small-sized feature set. We optimized the penalty parameter C of the error term by tuning it in the range $C \in [10^{-5}, ..., 10]$ and the gaussian kernel parameter G in the same range as well, using the development set. To obtain the results on the test set we combined the training and development set and trained the models using the optimized parameters.

5.3.4 Results and Discussion

In Table 5.7, we report the performances of the automatic classification system trained on different feature types: lexical (automatic transcriptions), acoustic and psycholinguistic. We report test set results for feature combination-based system as well as classifier combination. In the latter system we applied majority voting. In order to compute the baseline we have randomly selected the class labels based on the prior class distribution. For the statistical significance, we have computed McNemar's significant test over the test set [191].

Table 5.7: Empathy classification results at the segment level using acoustic, lexical, and psycholinguistic (LIWC) features together with feature and decision level fusion. Maj - Majority voting. Dim. - Feature dimension.

Experiments	**Dim.**	**Test-Set**
Random baseline		49.3
Acoustic	2400	68.1
Lexical	8000	65.6
LIWC	89	67.3
Acoustic+Lexical	2600	68.3
Maj(Acoustic+Lexical+LIWC)		70.1

For single feature-type systems, acoustic-based models provided the best performance compared to lexical and psycholinguistic alone. The results of acoustic-based system are significantly better than random baseline with $p < 2.2E - 16$. The acoustic-based system provides a useful and low-computation classification model, when no automatic transcriptions are available. LIWC's system performance improve over the lexical-only system with very few psycholinguistic features (89). In addition, all system's UAs are higher and statistically significant with $p < 2.2E - 16$ compared to the baseline results.

In terms of feature and system combination, we obtained the best results with *majority voting*. The statistical significance test showed that the results of the majority voting are statistically significant with $p <= 0.0004$ compared to any other system's results. Compared to the baseline, the best model for automatic classification provides a relative improvement over the baseline of 35.7%. Linear combination of lexical with acoustic features has not improved performance, despite its success in other paralinguistic tasks [160], linear combination in the feature space has not improved performance even when combined with feature selection.

5.4 Summary

Empathy refers to an emotional state triggered by a shared emotional experience. Being empathic is crucial for humans and their prosocial behavior as well as to facilitate human-machine interactions. In this chapter, we discussed our experimental study for both conversation and segment levels. We designed the automatic empathy classification system based on an operational annotation model, which has been designed by following Gross' modal model. The annotation process describes the scene through the Situation→Attention→Appraisal→Response processes. We have operationalized the definition of empathy and designed an annotation process by analyzing the human-human dialogues in call centers. We designed binary classifiers for each task and investigated acoustic, lexical and psycholinguistic features, and their decision and feature level fusion. The results of the automatic classification system on call center conversations are very promising compared to the baseline for both conversation and segment level. The segment level classification study will lead us to the design of affective scene, i.e., emotional sequence, for the whole conversation. The investigation of feature sets suggests that lexical and psycholinguistic features extracted from automatic transcription can be useful for the automatic classification task. Clearly, this study shades the light towards designing natural human-machine interaction system, speech, behavioral analytics systems and summarizing large-scale call center conversations in terms of emotional manifestations.

Chapter 6

Basic and Complex Emotion Classification

The SISL affective behavior corpus contains basic and complex emotions, which include anger, and frustration, satisfaction, and dissatisfaction, respectively. In this chapter, we present the classification experiments for basic and complex emotions of both conversation and segment levels using the feature sets and classification methods presented in Chapter 4. In Chapter 3, we presented the data preparation process for both conversation and segment level classification experiments. For the conversation level classification experiments, we investigated acoustic, lexical and their feature level linear combination. We evaluated them using 10-folds cross validation method. Whereas for the segment level classification experiments, we used training, development and test set data split as presented in Chapter 3. In addition to the acoustic and lexical features, we also investigated psycholinguistic features for the segment level classification task. In both classification tasks, there exists a data imbalance problem, which we solved using different sampling techniques.

6.1 Conversation Level Classification

The experimental settings of the conversation level classification for basic and complex emotion are different in a few respects compared to what we discussed in Section 5.2 for the classification of empathy. For the classification of basic and complex emotions, we only evaluated acoustic and lexical features. In addition, we only used ASR transcriptions to extract lexical features whereas for the classification of empathy we investigated both manual and automatic transcriptions as presented in Section 5.2.

Table 6.1: Class distribution for the conversation level classifier.

Class	**Y**	**N**	**Total**	**Y (%)**	**N (%)**
Emp	530	636	1166	0.45	0.55
Ang	118	141	259	0.46	0.54
Dis	367	403	770	0.48	0.52
Fru	338	405	743	0.45	0.55
Sat	736	883	1619	0.45	0.55

6.1.1 Data Sampling

For conversation level classification problem we designed binary classifiers for each emotion category by considering whether particular emotion category exists in a conversant channel or not. In Table 3.15, we presented the original distribution of the dataset, which we prepared for the classification experiments. However, such a skewed distribution results in lower classification performance. Therefore, we down-sampled examples of majority classes by randomly removing them to make a balanced class distribution for each category as shown in Table 6.1,. For each binary classifier, we prepared dataset using the following approach. If a conversation contains at least one emotional event we labeled that conversation as a positive example, otherwise we labeled it as negative, as also shown in Figure 3.11.

6.1.2 Experimental Methodology

In Figure 6.1, we present the architecture of the automatic classification system, which takes a spoken conversation as input and generates a binary decision regarding the presence *or* absence of an emotional state. The recognition system evaluates the cues present throughout the spoken conversation and then commits to a binary decision. In order to evaluate the relative impact of lexical features, we used transcriptions obtained from an Automatic Speech Recogniser (ASR). The details of the ASR system is provided in Section 3.1.2.

In the feature extraction phase, we extracted a very large scale acoustic and lexical features. For the acoustic features, we first extracted low-level acoustic features, such as fundamental frequency, intensity, mfcc, and spectral, then projected them onto statistical functionals. For the lexical features, we extracted trigram features and then represented them as bag-of-word model with a transformation of tf-idf. The details of these feature extraction process is presented in Section 4. In the feature combination, we linearly combined acoustic and lexical features. Hence, we obtained three set of features. For each feature set, we applied Relief feature selection technique, in which we generated feature learning curve by incrementally evaluating top ranked features. We evaluated each feature set. We extracted, combined, and selected acoustic features directly from the speech signal and generated the classifier's training set. We used SVM to design the classification model, and evaluated each system using 10-fold cross validation. For each classification model, we also tuned SVM complexity parameter, C, in the range of $[0.0001 - 10]$. To measure the performance, we used UA metric.

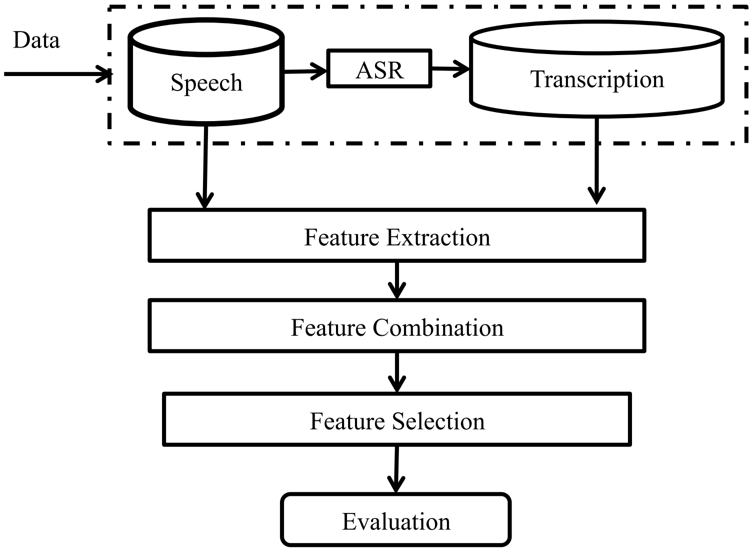


Figure 6.1: System for conversation level classification.

6.1.3 Results and Discussion

We present the classification results in Table 6.2 for all experimental settings. In the classification experiments, we obtained better results using lexical features compared to the acoustic features. The linear combination of acoustic and lexical features did not perform well due to the complexity of the large feature space. The other reason could be that the feature representation of these two sets is different, i.e., dense *vs* sparse, which may increase the complexity of the task. We computed random baseline results by randomly selecting class labels based on prior class distribution. For each emotional category, performances are statistically significant with $p < 0.05$ compared to the baseline. The significant test has been computed using two-tailed paired sampled t-test. For this study, we also computed oracle performance to understand the upperbound of the results that we can obtain using lexical and acoustic combination. The reason to compute *oracle* performance is to understand the upper-bound for each classification model, which shows that a relative improvement, ranges from 15% to 21%, can be achieved for each case.

The results of anger vary a lot in each cross-validation fold, which we see from a high standard deviation. The reason is that we have a very small number of instances for this emotional class. For satisfaction and dissatisfaction performances are comparatively lower than other two categories.

6.2 Segment Level Classification

The segment level classification experiment for the basic and complex emotion is much more complex than any other experiment we did for the classification of affective behavior. The reason is that we needed to deal it with multi-class classification settings, in which we choose to use pair-wise classification method. Due the unbalanced class distribution, we first grouped the anger and frustration into negative as discussed in Section 3.1.5.2. Then

Table 6.2: Results at the conversation level using different feature set. Ac: acoustic features, Lex (A): lexical features from ASR transcription.

Experiments	**UA Avg (Std)**				
	Ang	**Dis**	**Fru**	**Sat**	**Avg**
Random baseline	48.8(10.3)	49.7(5.0)	50.2(5.7)	50.4(4.3)	49.8(6.3)
Ac	66.3(7.4)	55.2(5.7)	61.3(4.9)	53.3(2.7)	59.0(5.2)
Lex (A)	76.3(6.5)	60.8(3.1)	65.9(7.4)	62.3(3.1)	66.3(5.0)
Ac+Lex (A)	67.6(11.8)	57.6(6.3)	63.0(4.8)	54.1(3.9)	60.6(6.7)
Oracle	84.7(3.4)	71.0(2.6)	79.9(6.6)	80.9(2.1)	79.1(3.7)

we undersampled the segments of majority class such as neutral. This has been done at the data preparation phase. More detail of this undersampling process is discussed in Section 5.3.3. After that oversampling of the minority classes, such as negative, satisfaction and dissatisfaction, has been at the feature level. For the designing and evaluating the classification model, we used the data split presented in Section 3.1.5.2, which includes training, development and test set. Development set is used for feature selection and parameter tuning, and final system is evaluated on the test set.

6.2.1 Experimental Methodology

For the segment level classification, we used the same system architecture presented in Figure 5.4. The system takes a spoken conversation as input and passes it to the speech *vs* non-speech segmenter to segment the speech and non-speech part. The system then passes the speech segment to the ASR for the transcription and to the feature extraction module. In the feature extraction phase, it extracts acoustic, lexical and psycholinguistic features. We also linearly combined the acoustic and lexical features. For each feature set, we applied feature selection using the same approach discussed earlier. Classification decisions have then combined using the majority voting. For the evaluation, we used UA as a performance metric.

6.2.2 Results and Discussion

In Table 6.3, we present the classification results on test set. Similar to the conversation level experiment, we also obtained better results with lexical features for segment level classification. The performance of LIWC features is lower than lexical features. However, it is higher than acoustic features, and the number of features is very low for this set. The decision level combination has not improved the performance for this case, whereas it shows higher improvement for the classification of empathy presented in Section 5.3.4. For the linear combination of acoustic and lexical features performance is also lower compared to the lexical features. The feature dimension for the lexical feature set is comparatively higher than the acoustic and LIWC feature sets.

Table 6.3: Results on test set for the basic and complex emotions using acoustic, lexical, and psycholinguistic (LIWC) features together with feature and decision level fusion. Maj - Majority voting. Dim. - Feature dimension.

Experiments	**Dim.**	**UA**
Random baseline		24.4
Acoustic	4600	47.4
Lexical	5200	56.5
LIWC	89	51.9
Acoustic+Lexical	5800	49.2
Majority: {Acoustic+Lexical+LIWC}		56.9

We observed that the recall of dissatisfaction is comparatively lower than other emotional categories. It also confuses with satisfaction due to the fact the manifestation of both satisfaction and dissatisfaction appear at the end of the conversation. For this reason, there is an overlap of the linguistic content, which also effects the paralinguistic properties of the spoken content. Using acoustic features, we obtained better performance for negative and neutral, whereas using the lexical feature we obtained better performance for negative and satisfaction.

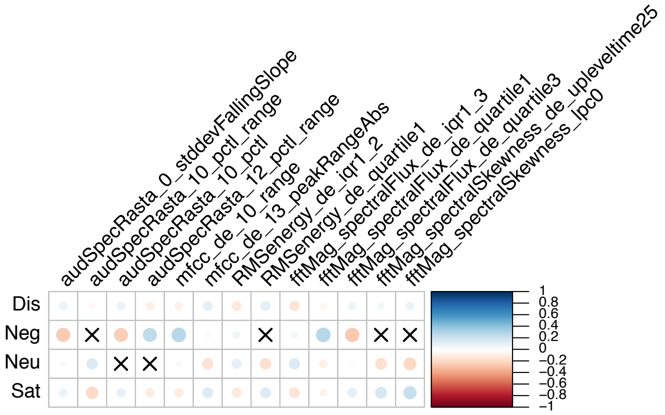


Figure 6.2: Correlation analysis of acoustic features Cell without asteric "*" are statistically significant. The color in each cell represents the correlation coefficients (r) and its magnitude is represented by the depth of the color. The "×" symbol represent the corresponding r is not significant.

To understand the upper-bound of the classification system, we designed a system by exploiting manual segments with which we obtained UA 70.9% using acoustic features.

In terms of discriminative characteristics, spectral, voice-quality, pitch energy, and mfcc features are highly important. The statistical functionals include arithmetic mean of peak, quadratic regression, gain of linear predictive coefficients, flatness, quartile and percentiles.

We have analyzed them in terms of class-wise correlation analysis as presented in Figure 6.2. The number in each cell in the figure represents the correlation value between class-label and a feature. The color in each cell represents the positive and negative association. The "×" symbol represent the association is not significant with $p = 0.05$. Even though the correlation value are very low, close to zero, however, most of them are statistically significant. Spectral and rasta style auditory spectram features are positively associated with satisfaction. For neutral spectral features are negatively associated. MFCC and rasta style auditory spectram features are positively correlated with negative emotion. Satisfaction and dissatisfaction are mostly

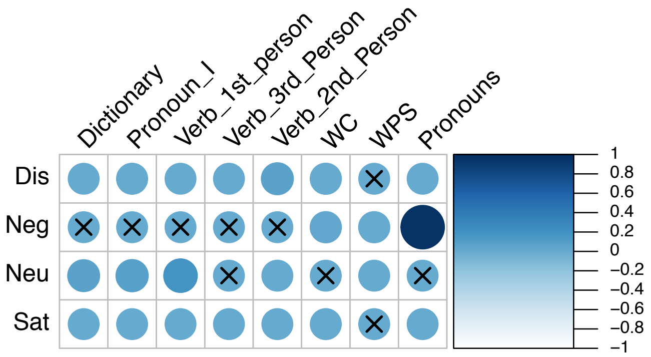


Figure 6.3: Correlation analysis of LIWC features. Cell without asteric "*" are statistically significant. The color in each cell represents the correlation coefficients (r) and its magnitude is represented by the depth of the color. The "×" symbol represent the corresponding r is not significant.

similar, only disimilar exist in the strength of positve and negative association in some features.

The correlation analysis of LIWC features is presented in Figure 6.3. The highly discrimitive LIWC features include personal pronouns, words associated with emotion and verb. Similar to the acoustic features, satisfaction and dissatisfaction are quite similar, however, there is a disassociation exist in the strength of the correlation. First three features, such as words containing in dictionary, pronoun (I), and 1st person verb, for neutral are not correlated, and these features are positively correlated with the negative emotion.

The correlation analysis of lexical features shows that negative emotion are highly associated with negative words whereas for satisfaction represents the mostly positve words such as "grazie mille/thank you so much/", "benissimo/very well/" and "perfetto/perfect/". The difference between satisfaction and dissatisfaction is that dissatisfaction represents some negativity, however, there is not much lexical difference. It might be because the annotators mostly focused on the tone of the voice, which distinguished the annotation of satisfaction and dissatisfaction. It is needed to mention that the LIWC and lexical feature analysis has been done based on ASR transcription. We do not present any figure of the correlation analysis of lexical features as it is very difficult to make a general conclusion from such graphical representation.

6.3 Summary

In this chapter, we presented our contributions to the design of a computational model for the basic and complex emotions, which include both conversation and segment level classification. The goal of the conversation level classification experiment was to detect the presence or absence of an emotional state in a conversation. We investigated acoustic and lexical features and their combination in which we obtained a better performance using lexical features. The goal of segment level classification experiment was to classify each speech segment into one of the emotional class that we predefined. For this experiment, the classification results of the lexical feature set are also perform better compared to other feature sets.

Chapter 7

Sequence Labeling of Affective Behavior

The importance of segmenting and labeling emotional state in a conversation are enormous, ranging from summarizing conversations to incremental processing in real applications. In Chapter 5 and 6, we have seen the challenges that are needed for segmentation and classification of an emotional state, in terms of system architecture, where one system's output feeds into another system. It combines automatic speech *vs* non-speech segmenter and a segment classifier. One scientific question here is that - can we design a generative sequence labeling model, which can able to perform both tasks?

In this chapter, we present our study towards answering that question. We experimented this task using SISL affective behavior corpus and FAU-Aibo robot corpus. We investigated HMM based sequence labeling approach while evaluated low-level acoustic features. Even if the research is in very early stage, however, the findings of our study highlights the future research avenues. The goal of this study is to design affective scene as presented in Figure 7.1.

7.1 Current Challenges

Emotion recognitionis not a trivial problem in real settings, whether it is from speech or other sources. As stated in Section 5.3, it is still not yet well understood what should be the smallest unit for an emotional segment, even if Batliner et al. empirically shown that "ememe" to be the smallest unit of emotion and a "word" can be considered as smallest meaningful emotional unit [122]. Their study was based on FAU-Aibo robot corpus in which the audio has been recorded in a scenario where children were interacting with a Robot. However, there is a lack of a theoretical ground regarding the smallest

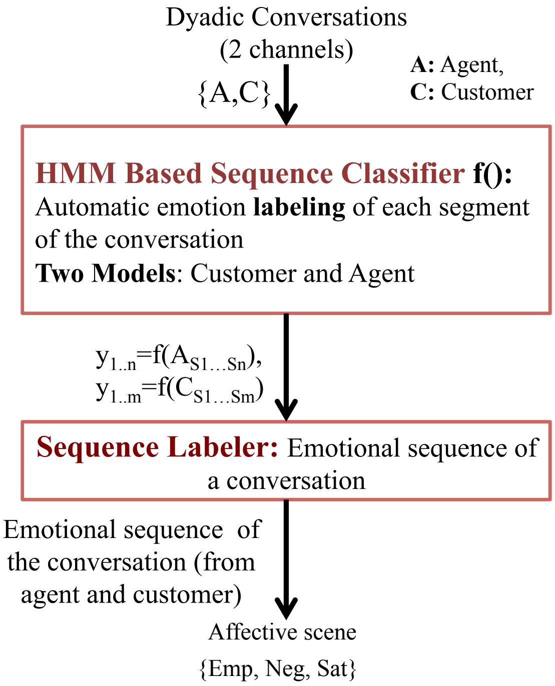

Dyadic Conversations
(2 channels)
{A,C}
A: Agent,
C: Customer
HMM Based Sequence Classifier f():
Automatic emotion labeling of each segment of the conversation
Two Models: Customer and Agent
$y_{1..n}$=f($A_{S1...Sn}$),
$y_{1..m}$=f($C_{S1...Sm}$)
Sequence Labeler: Emotional sequence of a conversation
Emotional sequence of the conversation (from agent and customer)
Affective scene
{Emp, Neg, Sat}

Figure 7.1: System architecture of the HMM based affective scene (emotion sequence) generation system.

unit of emotion compared to the studies of ASR.

In order to solve this problem in an automatic fashion a straightforward approach to one can follow is to use automatic speech recognition (ASR) pipeline. However, there are issues that made the task difficult, as mentioned below.

- For ASR, the smallest unit is a phoneme, which is theoretically well understood and well defined. Whereas for the emotion, it is not yet clear enough.
- Number of state in HMM topology is also well defined for ASR, which is based on phonetic patterns, such as different spectral characteristics. It has been studied for more than a few decades and widely accepted that 5-states (3 emitting and 2 non-emitting) HMM phone model is good enough for ASR. From the theoretical standpoint, each phoneme has a start, middle and end position, which constitutes three different

spectral characteristics for each phoneme [192]. It is also well studied in phonetics about how each phoneme differ from one another in terms of the place and manner of articulation. However, for emotion, these understandings are yet to be discovered and it poses a challenge to define how many states there should be in HMM topology.

- In terms of features, for ASR, MFCC and their derivatives with some sort of transformations are widely used to transcribe speech with a reasonable accuracy [193]. The challenge here is that - can we only use MFCC features for emotion? Or it is yet to discover whether we have to employ other acoustic features that have been studied and reported in the literature for emotion recognition task [7, 92].
- Type of topology for HMM, whether fully-connected –ergodic HMM or left-to-right–Bakis. The use of these two types of topology has been reported in the literature for utterance based emotion classification [7].
- The other thing is whether to employ hand-crafted grammar or a simple unigram based grammar of emotion sequence from the corpus.

Tackling these challenges is not a trivial task. It involves empirically optimizing every single parameter and validating the performance with reference labels. The details of the experimental procedures are reported in the following subsections.

7.2 Sequence Labeling System

In Figure 7.2, a functional diagram of the system is shown. For the experiments with SISL corpus, we used the same data-split, as reported in Section 3.1.5.2. We also used CEICES dataset of FAU-Aibo robot corpus for the study, which is discussed in Section 3.2. In terms of segmentation study, the difference between two corpora is that in Aibo corpus data has been released at the chunk level whereas the SISL corpus consists of complete recorded conversations. As presented in Figure 7.2, the system takes an audio as input, then extract acoustic features, which can optionally feed passed through the feature transformation into the training or label generation module to train or produce emotion label sequence.

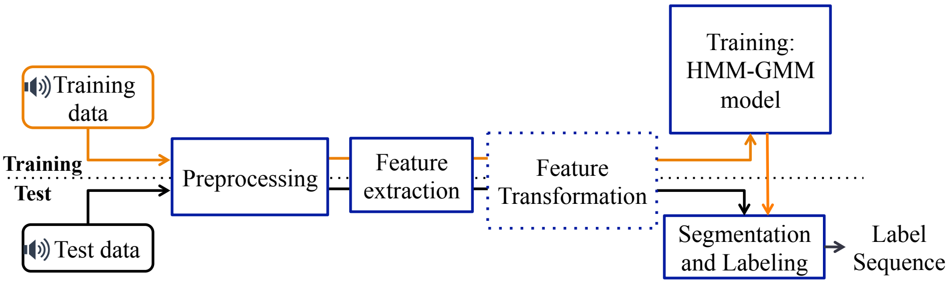


Figure 7.2: System architecture of the HMM based affective scene (emotion sequence) labeling system.

Acoustic features: As features, we exploited MFCCs, their Δ and $\Delta\Delta$ coefficients, which we extracted using a window of 25 ms and a frame shift of 10 ms. The reason for choosing smaller window is that statistical properties might remain constant within this region. We used Hamming windowing technique, which is state-of-the-art technique as a windowing function. It basically shrinks the values of the signal towards zero at the window boundaries and avoids discontinuities.

Model Training: For training the baseline model we used the following definition of HMM topology.

- 3 emitting states and 2 non-emitting states
- left-to-right transition (Bekis) with self-loop for emitting states
- initial transition probabilities are equally likely
- the use of self-loops allows a phone/state to repeat (perhaps many times) so as to cover a variable amount of time of the acoustic input.

There is an intuition to design 5 states (3 emitting and 2 non-emitting) HMM phone model for ASR. As each phone has three different spectral characteristics, at the beginning, middle, and at the end. Therefore, 5-states HMM model is the most common configuration. The idea is that each state in the model corresponds to some acoustic feature vectors and we would like the feature vectors assigned to each state to be as uniform as possible so that we can get an accurate Gaussian model.

For emotion, it is not yet clearly understood whether 5 states model is sufficient or not. Therefore, the experiment has been conducted by using a different number of states. It is empirically found that 5 states work better for emotion model. For each experiment the following parameters has been optimized:

- number of states,
- acoustic model weight,
- number of components in gaussian mixture model, and
- beam width.

While training the system, we optimized the parameter on the development set and finally evaluated the system on the test set, which has been done for each corpus.

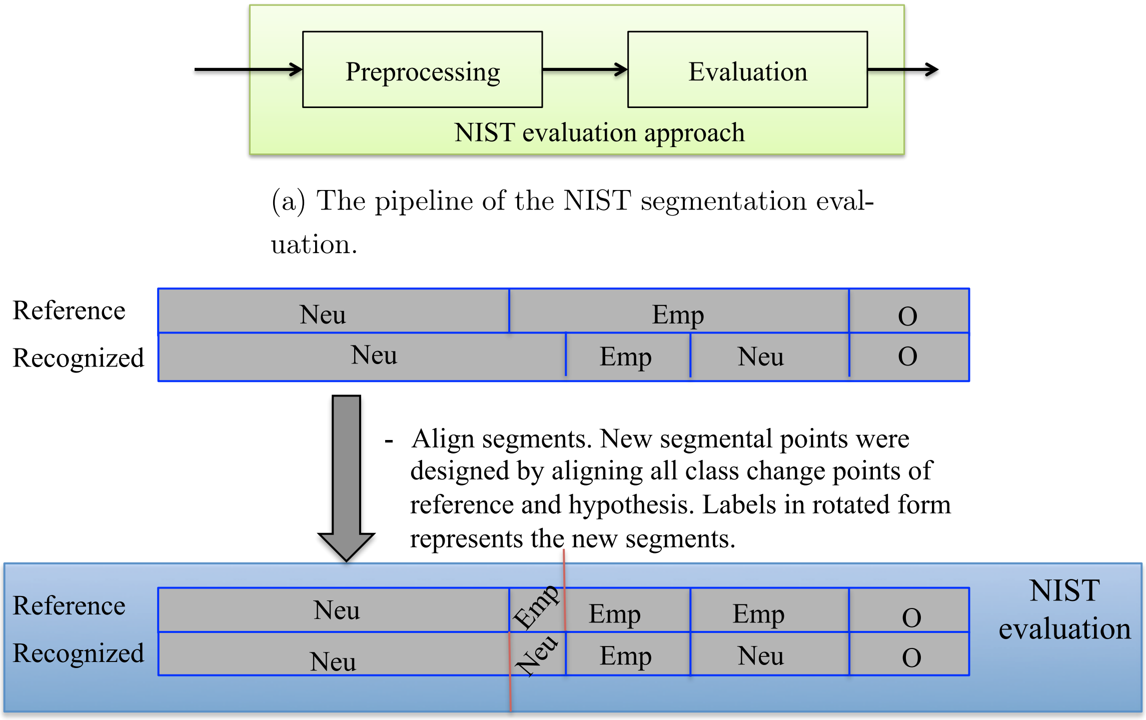
7.3 Evaluation Methods

For the evaluation of the emotional sequence labeling, we adopted NIST speaker diarization evaluation method [166]. A modified version of this approach has been used in [167, 194, 195]. As shown in Figure 7.3a, the NIST evaluation approach work as follows:

- **Preprocessing (Alignment) Step:**
  - Each conversation is divided into contiguous segments at all class change points. Then, it performs an alignment between reference and hypothesis as shown in Figure 7.3b.

  - A class change points occurs each time a reference or hypothesis class start and end.

- **Evaluation:** We used the evaluation metrics discussed below.



- Align segments. New segmental points were designed by aligning all class change points of reference and hypothesis. Labels in rotated form represents the new segments.

(a) The pipeline of the NIST segmentation evaluation.

(b) Preprocessing steps of hte NIST evaluation approach.

Figure 7.3: Preprocessing steps of the NIST evaluation approach.

The score of the segmentation error is computed as the fraction of the class time that is not correctly attributed to that class. The score is defined as the ratio of overall segmentation error time to the sum of the durations of the segments that is assigned to each class in the conversation. The segmentation error for each segment is defined as follows:

$$E\left(seg\right) = dur\left(seg\right) \times \left(max\left(N_{Ref}\left(seg\right), N_{Hyp}\left(seg\right)\right) - N_{Correct}\left(seg\right)\right) \tag{7.1}$$

where,
$dur\,(seg)$ = duration of the segment
$N_{Ref}\,(seg)$ = number of reference classes that are present in segment
$N_{Hyp}\,(seg)$ = number of hypothesis/predicted classes that are present in segment
$N_{Correct}\,(seg)$ = number of classes correctly predicted in segment

The overall segmentation error was computed using the Equation 7.2:

$$Overall\ Segmentation\ Error = \frac{\sum\limits_{All\_segs} E\,(seg)}{\sum\limits_{All\_segs} \{dur\,(seg) * N_{Ref}\,(seg)\}} \tag{7.2}$$

Class error time: The class error time, as shown in Equation 7.3, is the amount of time that has been assigned to an incorrect class, which happens in the following scenarios:

- if the number of predicted classes is greater than or less than the number of reference classes, and
- if the number of predicted classes and reference classes is greater than zero.

$$Class\,error\,time = dur\,(seg) \times (min\,(N_{Ref}\,(seg)\,, N_{Hyp}\,(seg)) - N_{Correct}\,(seg)) \tag{7.3}$$

Correctly segmented class time (Recall): The correctly segmented class time, as shown in equation 7.4, is the amount of time that has been assigned to a correct class.

$$Correctly\ segmented\ class\ time = 1 - Class\ error\ time \tag{7.4}$$

Missed class time: The missed class time is the amount of time where the segment contains more reference classes than the number of predicted classes.

$$Missed\ class\ time = dur\left(seg\right) \times \left(N_{Ref}\left(seg\right) - N_{Hyp}\left(seg\right)\right) \tag{7.5}$$

False Alarm time: The false class time is the amount of time where the segment contains more predicted classes than the number of reference classes.

$$Missed\ class\ time = dur\left(seg\right) \times \left(N_{Hyp}\left(seg\right) - N_{Ref}\left(seg\right)\right) \tag{7.6}$$

7.4 Experiments: SISL Affective Behavior Corpus

For the SISL corpus, two different system has been designed to deal agent and customer's emotional states separately. The emotional state for the agent's channel include neutral, and empathy, whereas for the customer's channel, it includes negative, satisfaction and neutral. In addition to the emotional categories for each system, we also designed a silence model so that the system can also deal with silence. To evaluate the performance of the systems we employed NIST speaker diarization based approach and computed duration weighted recall. The details of the evaluation procedures have been presented in Section 7.3. In Table 7.1, we present the results of the two systems. The agent model was designed for labeling empathy and neutral, whereas customer model was designed for multiple class labels such as negative, satisfaction, and neutral.

Table 7.1: Results of the emotion segmentation and classification of the agent and customer's emotional states of the SISL affective behavior corpus.

Experiments	Agent (Binary)	Customer (Multi-class)
Random baseline	49.4	32.2
HMM-Sequence labeling	52.1	31.9

The generative based sequence labeling task is more complex compared to the task we discussed in Chapter 5 and 6. Comparing the performance of the two different models, we observed that the customer's emotion model is more complex due the multiple class labels and also the imbalance distribution among them. From the automatically generated class labels of the two models we can design complete emotional sequence for the whole dyadic conversation representing the affective scene.

7.5 Experiments: FAU-Aibo Robot Corpus

We employed the CEICES dataset of FAU-Aibo Robot Corpus for the segmentation and classification task (see Table 3.20). The released Aibo dataset contains chunk level wav files and word level emotion segment information, which has been used for this study. It is observed that a portion of the wav file has not been annotated, therefore, for this study, it is labeled as *O*. Hence, the emotion label includes *anger, emphatic, motherse, neutral* and *O*. The content with label *O* might be non-speech and silence. The duration distribution of this set is presented in Table 7.2.

Table 7.2: Duration and frequency distribution of the Aibo word level emotion classes. H-hour, M-minute, S-Second

Class	Avg. (in Sec.)	Std.(in Sec.)	Total	Freq.
Ang	0.52	0.23	13M 32.78S	1557
Emphatic	0.48	0.21	13M 13.62S	1645
Mothersee	0.41	0.23	8M 22.58S	1223
Neu	0.30	0.17	8M 19.11S	1645
O	0.78	0.57	2H 35M 16.18S	11979

For the experiment, we applied the same procedures to optimize the parameter and evaluate the system. The un-weighted average (UA) of the segmentation and classification results is 43.4%. The results on neutral (34.4%) and mothersee (34.3%) are comparatively lower, which might be due to the skewed distribution in terms of duration. In [28], it is reported that at the word level with four class problem the best average recognition rate using

MFCC features is 57.5%. However, they also reported the with a combination of acoustic and linguistic features an average recognition rate of 67.2% can be obtained. It is needed to mention that their experiment does not include segmentation process with the classification pipeline. In comparison, this study deals with both segmentation and classification together and also deals with five emotion classes. In this study, the obtained results are with only MFCC and their derivatives.

7.6 Summary

In this chapter, we presented our contributions to the automatic emotion segmentation and labeling task by utilizing two real-life corpora. The task is challenging in different respects such as defining the smallest unit of the segment, and other HMM parameters are not defined. Our study focused on empirically define those parameters towards solving this problem. The limitation of our investigated dataset is that their distribution is very skewed among class labels, which is one of the reasons of lower performance. The obtained performance shows promising research avenue for future work. It is well worth to do more research on this area to design an architecture for labeling affective scene (emotion sequence).

Chapter 8

Affective Scene Classification

In many conversational scenarios, it is necessary to understand how a conversation started and end in terms of emotional manifestations. For example, in a dyadic conversation, a customer might start emotional manifestation with anger from a neutral state, then the agent might show empathy, followed by the customer might show satisfaction at the end of the conversation. For an analytical purpose, one might wants to understand the proportion of the conversation ends with customer's satisfaction positive emotion, or negative emotion.

In this chapter, we present our study, in which we investigated such research questions. For example, how to design such a model based on the affective scene (emotional sequence) in a conversation? We defined the term "*affective scene*", by analyzing SISL affective behavior corpus, designed the affective scene framework and classification model. Our research shows promising research directions for analyzing affective scenes in spoken conversations.

8.1 Affective Scene

The emotional states of individuals engaged in conversations are characterized by continuous variations. We hypothesize that the linguistic and contextual structures of such variations can be objectively described by exploiting the correlation between the continuous variations in speakers' emotional states and variations in the situational context of the interaction. In the psychological literature, there are some competing models aiming to capture and describe such variations, including Scherer's dynamic theory of emotion differentiation and sequential checking [15]. For our experiment we refer to a general, yet clear and flexible, psychological model that may account for the interplay between a variation of emotional state and the variation of the situational context, i.e. the *modal model* of emotions [76]. According to the modal model, the emotion-arousal process is believed to be induced by a *Situation*, a physical or virtual space that can be objectively defined. The *Situation* compels the *Attention* of the subject and triggers the subject's *Appraisal* process and the related emotional *Response*. The *Response* may generate actions that in turn modify the initial *Situation*. In this study, we focused on the *affective scenes* that ensue in such communicative situations.

The ***affective scene*** has been defined in the context of a dyadic human communication, but it can easily be generalized to multiparty communication. The affective scene is ***an emotional episode where one individual is affected by an emotion-arousing process that (a) generates a variation in their emotional state, and (b) triggers a behavioral and linguistic response. The affective scene extends from the event triggering the unfolding of emotions on both individuals, throughout the closure event when individuals disengage themselves from their communicative context.*** [52].

While this process is continuous in terms of the human response signals, we describe the unfolding as a sequence of discrete emotional episodes that have an initial state, a sequence of states, and a final state. In order to describe the framework of affective scenes we focused on *who* 'first' shows the variation of their emotional state, *how* the induced emotion affects the other speaker's emotional appraisal and response, and *which* modifications such a response may cause with respect to the state that triggered the scene.

In Figure 8.1, an example of an emotional sequence has been presented, which depicts an affective scene. In the example, 1) *who?:* customer 'first' manifested frustration at time instant 238.78 to 262.37 seconds, 2) *how:* agent appraised with customer's emotion and responded with empathic behavior at time instant 271.7 to 291.97 seconds, 3) *which:* customer manifested satisfaction from 416.77 to 423.19 seconds at the end of the conversations. The lack of empathic response from the agents may cause different patterns of emotional variations, including customer's anger, dissatisfaction, or frustration.

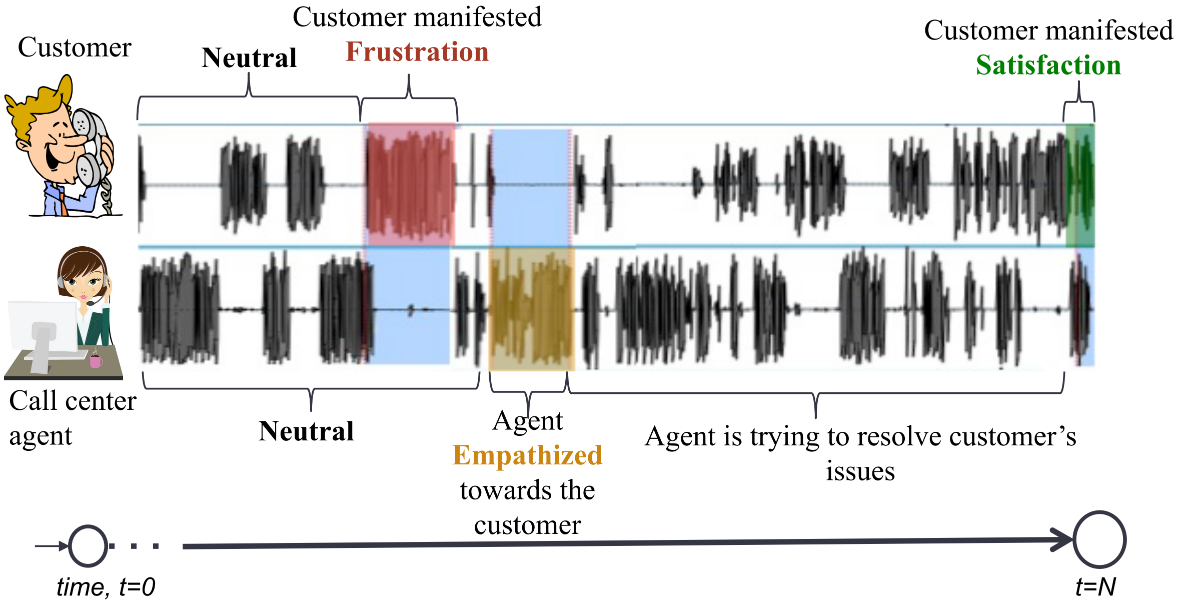


Figure 8.1: A prototypical example of an emotional interaction between call canter agent and customer. From time 238.78 to 262.37 seconds customer manifested frustration, 271.7 to 291.97 agent empathized towards customer and 416.77 to 423.19 seconds customer manifested satisfaction.

8.2 Data Analysis

The definition of the affective scene has been applied to the analysis of dyadic spoken conversations between customers and agents collected in call centers. The details of the corpus are provided in Section 3.1. The analysis of the conversations provided evidence for several occurrences of *prototypical* emotional sequences. For example, it is observed that there are situations where customers call in order to complain about an unfulfilled service request they made a few weeks before. The customers are *frustrated* due to the delay, and the manifestations of their emotional state trigger the start of affective scene instances. This scenario represents the term *who* in our definition: in this scenario, *who* initiated emotion? → customer. The agents may understand the point of view of the customers, and *empathize* with their distress. In addition, the agents may take all the required actions in order to solve the customers' problem. The appropriate response and actions by the agents may impact on the emotional states of the customers. In this case, *how* the agent shows an emotional response that reflects the second term of our definition, such as *how* agent is responding? → by empathizing.

The emotional states can vary again so that the call may end with customer's satisfaction. On the other hand, the lack of empathic response from the agents, and/or the fact that the problem cannot be solved immediately, may cause different patterns of emotional variations, including customer anger, dissatisfaction, or further frustration. The type of emotional modification we see here on the customer side reflects the third term of our definition; in this scenario, *which* type of modification? → satisfaction, anger or dissatisfaction.

In order to clarify the scenarios, we illustrate two prototypical communicative situations, as shown in Table 8.1. The first situation is characterized by a customer's initial discontent, and the second by agent's initial positive attitude towards the customer's state of mind. As it can be seen in the Table 8.1, the unfolding of the affective scenes from those initial situations may greatly vary from one scenario of communicative situations to another one. In the first example, row 1 in Table 8.1, we see that the customer 'first' manifests emotion with frustration, then the agent understands and empathizes, and 'finally' the customer changes their emotional state from frustration to satisfaction.

Table 8.1: Types of affective scenes for different communicative situations. Initial State: Initial emotional state of either agent or customer. A: Agent, C: Customer, Emp: Empathy, Ang: Anger, Fru: Frustration, Sat: Satisfaction. As an example, C: Fru means customer manifests frustration. A complete emotion sequence with → indicates the flow of emotions in a conversation.

Initial state	**Scenarios**	**Examples**
Customer initial discontent	Agent understands, and customer's discontent is attenuated	C: Fru → A: Emp → C: Sat
	Agent understands, but customer emotional state either get worse or does not evolve positively	C: Fru → A: Emp → C: Fru
		C: Fru → A: Emp → C: Ang
	Agent does not understand, and customer emotional state either get worse or does not evolve positively	C: Fru → A: Neu → C: Fru
		C: Fru → A: Neu → C: Ang
Agent preempting of possible customer discontent	Customer emotional state does not vary	A: Emp → C: Fru or Ang
	Customer emotional state evolves into a positive attitude	A: Emp → C: Sat

8.3 Affective Scenes in Spoken Conversations

We analyzed 460 annotated conversations containing empathy on the agent side and other basic and complex emotions on the customer side. In Table 8.1, we report examples of two communicative situations based on the 'initial' manifestations of emotion. In the examples, we also report the customers' 'final' emotional state during the conversation.

Based on the initial and final emotional displays we defined and categorized the conversations with different categories of *affective scenes* given in Figure 8.2, which depicts the emotional sequence examples in Table 8.1. From the figure, we see that after the start of the conversation either the agent or the customer manifest emotions. Following that, there are many emotional transitions between the customer and the agent, and there is a 'final' emotional manifestation.

Hence, considering the 'initial' emotional displays of the customer, and the agent, and 'final' emotional displays of the customer, there are three categories of affective scenes as listed below.

1. Agent or customer manifest emotions at the start of the conversation, therefore the labels - Agent First (AF) *or* Customer First (CF) has been used.
2. Agent 'first' manifests emotion after the start of the conversation and customer shows positive/negative emotion at the end of the conversation. For this scenario, the labels AF-Pos *or* AF-Neg has been used.
3. Customer 'first' manifests emotion after the start of the conversation and customer shows positive/negative emotion at the end of the conversation. To define this scenario the labels CF-Pos *or* CF-Neg has been used.

The negative emotions, in this case, are anger, frustration, dissatisfaction and not-complete dissatisfaction whereas the positive emotion is satisfaction and not-complete satisfaction.

From the analysis of emotional sequences we can see that $Emp \rightarrow Sat$ appears more frequently than others, 30.7% relative frequency distribution. Some examples of emotional sequence and their distributions are presented in Table 8.2.

8.4 Experimental Methodology

In this Section, we present the detail study of the affective scene classification by utilizing low-level acoustic and lexical features, which has been ex-

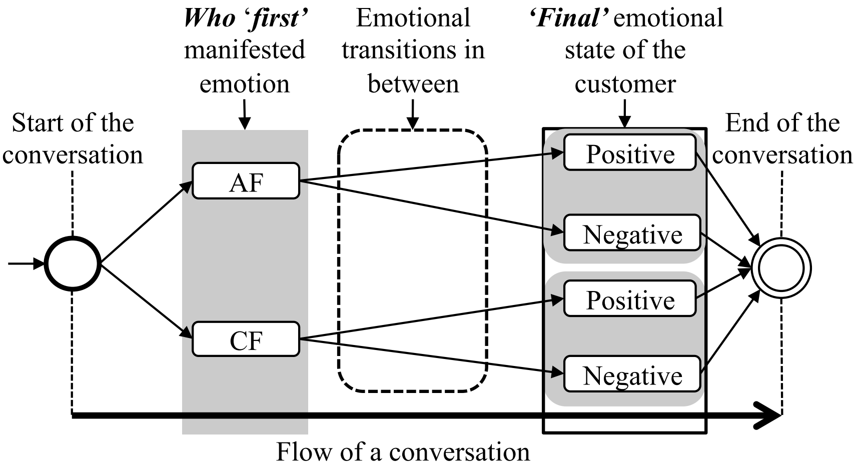

Who 'first' manifested emotion
Emotional transitions in between
'Final' emotional state of the customer
Start of the conversation
End of the conversation
AF
CF
Positive
Negative
Positive
Negative
Flow of a conversation

Figure 8.2: State traces of affective scenes. Starting from the initial state, an affective scene may reach either state **AF** (Agent first manifests emotion) or **CF** (Customer first manifests emotion). Then, following a natural unfolding of emotional states the affective scene may reach either a **Positive** final state (Customer manifests emotion with satisfaction at the end) or a **Negative** final state (Customer manifests emotion with either anger, frustration or dissatisfaction at the end).

tracted from the conversation. To automatically classify the affective scenes categories, described in Section 8.3, three binary classification tasks has been designed by utilizing two different architectural design. The three binary classification tasks are listed below. Class distribution for each of the classification task is given in Table 8.3.

1. Agent First (AF) *or* Customer First (CF);
2. Agent First, customer manifests Positive emotions at the end (AF-Pos) *or* Agent First, customer manifests Negative emotions at the end (AF-Neg);
3. Customer First, customer manifests Positive emotions at the end (CF-Pos) *or* Customer First, customer menifests Negative emotions at the end (CF-Neg).

The system architecture of the affective scene classification has been presented in Figure 8.3. Since each conversation is represented by two chan-

Table 8.2: Examples of emotional sequence (Seq.) with their relative frequency distribution (Dist.) out of 460 conversations

Initial emotional manifestation	Seq.	Dist.
Agent manifested emotion 'first'	Emp → Fru	3.5
	Emp → Dis	3.5
	Emp → Sat	30.7
Customer manifested emotion 'first'	Fru → Ang	3.3
	Ang → Dis	4.8
	Fru → Dis	10.0
	Fru → Emp	9.8
	Fru → Emp → Sat	3.5
	Fru → Sat	2.8

Table 8.3: Distribution of conversations for each classification task.

	Task 1		Task 2		Task 3	
Class	AF	CF	AF-Pos	AF-Neg	CF-Pos	CF-Neg
No. of conv.	213	247	160	53	47	200
%	46.3	53.7	75.1	24.9	19.0	81.0

nels (agent and customer), therefore, features has been extracted from both channels and then concatenated them. After that feature selection has been performed and classifier has been designed and evaluated for each task using each feature set. We extracted low-level acoustic, lexical and psycholinguistic features. These feature sets have been investigated to understand their distinctive properties for the classification of affective scenes. Another reason to investigate these feature sets was that it is needed to understand whether they can be useful for the classification without explicitly knowing the emotion sequence.

8.4.1 Feature Extraction

Acoustic feature has been extracted from each channel using the approach mentioned in Section 4.1.1. Low-level acoustic features has been extracted and then projected them onto statistical functionals, using openSMILE [128],

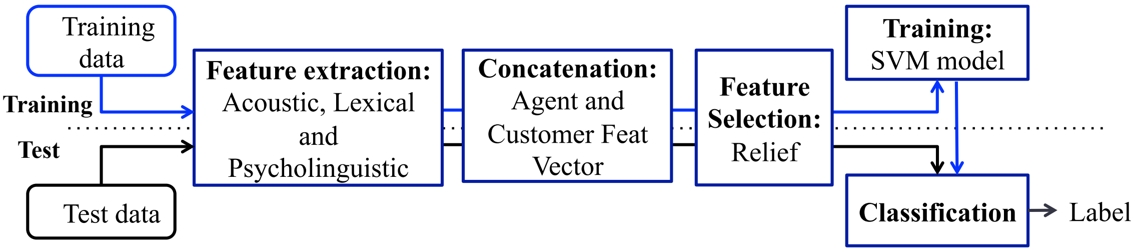

Training data
Training
Test
Test data
Feature extraction: Acoustic, Lexical and Psycholinguistic
Concatenation: Agent and Customer Feat Vector
Feature Selection: Relief
Training: SVM model
Classification
Label

Figure 8.3: System architecture of affective scene classification system using low-level features such as acoustic, lexical and psycholinguistic features.

based on the feature configuration referred in [179]. After projecting low-level features and their delta onto statistical functionals, the feature set contains 6373 features. Since each conversation is comprised of agent and customer channels, therefore, the same number of acoustic features has been extracted from the Agent, $A = \{a_1, a_2, ..., a_m\}$ and the Customer, $C = \{c_1, c_2, ..., c_m\}$. Then, merged the features from both channels to form a new feature vector, $X = \{a_1, a_2, ..., a_m, c_1, c_2, ..., c_m\}$.

Since lexical choices of the speaker provides evidences that represent emotional manifestations. Therefore, for the classification, we extracted lexical features from automatic transcriptions. Affective scene instances have been designed by concatenating the transcriptions from agent and customer channels. Then, converted them into lexical feature vector in the form of bag-of-words and used tf-idf (term frequency times inverse document frequency) weighting scheme. More details can be found in Section 4.1.2.

Following the approach presented in Section 4.1.3 we extracted 103 psycholinguistic features from automatic transcriptions using LIWC.

8.4.2 Feature Combination and Selection

In addition, to understand the performance of each feature set, such as acoustic and lexical feature sets, we also wanted to understand their combined contribution. Therefore, following the feature extraction, we merged acoustic and lexical features into a single vector to represent each instance in a high-dimensional feature space. Let $S = \{s_1, s_2, ..., s_m\}$ and $L = \{l_1, l_2, ..., l_n\}$ denote the acoustic and lexical feature vectors respectively. The feature-combined vector is $Z = \{s_1, s_2, ..., s_m, l_1, l_2, ..., l_n\}$ with $Z \in R^{m+n}$.

Since each individual feature set is higher dimensional, particularly acoustic and lexical, we applied Relief [158] feature selection technique. It has been shown in the literature [163] that this feature selection technique comparatively performs well, for the paralinguistic task, compared to other techniques such as Information gain. Relief estimates the quality of a feature based on how well its values distinguish among instances that are near to each other. For a given instance, it searches for two nearest instances, one from the same class and one from different class and estimates weight of an attribute depending on the values of the nearest instances.

As a part of feature selection process, we generated feature learning curves using ranked features from Relief and selected optimal set of features when performance starts decreasing.

8.4.3 Classification

We designed our classification models using Sequential Minimal Optimization (SMO) [165], which is a technique for solving the quadratic optimization problem of Support Vector Machines' (SVM) training. We trained the model using an open-source implementation Weka machine learning toolkit [164]. We chose to use *linear kernel* of SVM in order to alleviate the problem of higher dimensions such as overfitting. In order to measure the performance of the system we used Un-weighted Average, $UA = \frac{1}{2}\left(\frac{tp}{tp+fn} + \frac{tn}{tn+fp}\right)$, where tp, tn, fp, fn are the number of true positives, true negatives, false positives and false negatives, respectively. It has been widely used for the evaluation of the paralinguistic task [102]. Due to the limited size of the corpus we chose to use 10-folds cross-validation. In addition, we optimized the penalty parameter C in the range of $[10^{-5}, ..., 10]$.

8.5 Results and Discussion

The performance of the system for each feature set is shown in Table 8.4. We present the average UA of 10 folds cross-validation, their standard deviation, and a number of features for a particular feature set. We also present random baseline results, which we computed by randomly generating class labels based on the prior class distribution.

Table 8.4: Classification results of affective scenes categories in terms of UA, (average±standard deviation) with feature dimension (Feat.). Ac: Acoustic, Lex-ASR: Lexical features from ASR transcription, LIWC: Psycholinguistic features. {AF,CF}: Agent First, Customer First, AF:{Pos,Neg} Agent First with Positive/Negative emotion of the customer at the end, CF:{Pos,Neg} Customer First with Positive/Negative emotion of the customer at the end

	Task1 {AF,CF}		**Task2** AF: {Pos,Neg}		**Task3** CF:{Pos,Neg}	
Exp.	Avg±Std	Feat.	Avg±Std	Feat.	Avg±Std	Feat.
Random	49.3±7.0	-	49.8±10.1	-	49.0±11.2	-
Ac	58.5±6.7	1000	65.0±13.9	3000	63.9±10.0	4500
Lex-ASR	73.2±6.2	6800	67.5±12.8	5000	70.3±8.1	6800
LIWC	67.8±5.9	89	56.9±11.2	89	49.5±10.6	89

Out of the three classification tasks, we obtained better performance on task 1, {AF,CF}, compared to the other two classification tasks. The higher variation of the classification results of task 2 and 3 is due to the imbalance class distribution and a smaller number of instances compared to task 1. From the classification results, we observed that the performance of lexical features outperforms any other single or combined feature set such as acoustic, lexical, acoustic with lexical and LIWC.

The performance of acoustic feature set is better than random baseline and it might be useful when there is no transcription available. In terms of feature dimension, with feature selection, we obtained smaller-sized features for this set compared to lexical features for all three tasks. For task 2 and 3 acoustic features performs better than psycholinguistic features. For all three tasks, we found that spectral features are highly relevant for discriminating between classes.

The performance of psycholinguistic features is better than acoustic in task 1, however, in other two tasks the performances are worse. We used the linear kernel of SVM for all classification experiments, however, it might not be a better fit for this feature set. The Gaussian kernel might be a better option in this case, which we might explore in future.

Linear combination of acoustic+lexical (Z) did not perform well due to the complexity of the large feature space, which we are not presenting here. We will explore it in future by studying it with ensemble methods such as stacking.

Even though the classification performance varies across tasks and feature sets, however, from the results we can infer that automatically categorizing the affective scenes might be future research avenue to investigate. Our conceptual framework can be a good starting point towards defining affective scenes and its automatic classification.

8.6 Summary

We have proposed a conceptual framework of *affective scenes* to describe the dynamics of emotion unfolding in natural conversations. Even though we validated the *affective scenes* framework on call center conversations, nevertheless we believe that the framework is applicable to behavioral analysis of other social scenarios, for example, therapist-patient interactions. Our future research efforts will move towards this extension, as well as to the evaluation of the possible suitability of alternative psychological models, as long as we will be able to experimentally identify the still lacking features of high-level emotional flow in dyadic conversations. In this paper, we also investigated automatic classification of *affective scenes* categories by exploiting acoustic, lexical (ASR transcription) and psycholinguistic features. We obtained promising performance using lexical features, and with all other feature sets, we are still getting better than random baseline.

Chapter 9

Cross Language Emotion Classification

There are many use-cases of automatic speech emotion recognition in real-life settings such as agent-customer, therapist-patient and teacher-student interactions. To this end, most studies on emotion recognition present the performance of the system with acted data with overestimated results. For the sake of designing a generalizable a system, which is capable of working across domain and language it is necessary to evaluate it in cross-corpora. It is a great challenge to design such a system, however, the cross-corpora study can provide us an insight about its applicability.

In this chapter, we present our study of cross-corpora in three different setting such as *inter*, *intra* and *mixed*, using two real-life corpora. For the classification experiment, we utilized a large set of acoustic features and experimented with binary and multi-class settings. Our experimental results show that mixed settings perform similarly compared to intra settings, and better than inter settings. The binary classification results is a good starting point towards using the system across domain and language, however, there is much room for improvement in multi-class settings.

9.1 Research Goal and Current State-of-art

One of the important problem in any domain is to design models that are applicable across language and domain, robust enough in terms of generality and portability. Even if most of the automatic classification study has been done in speaker independent or cross validation method, however, it is difficult to tell that those systems are generalized well when used in a different domain. Designing a generalized model is complex due to the variations in recording conditions such as microphone type and position, room acoustics, observed emotions, type of interactions and language. The common approach towards understanding the generalizability of the designed system is to run evaluations in a cross-corpus setting. Another important problem is that emotional corpora in the real setting are rare and expensive to collect, however, the application domains are enormous. Therefore, it is important to evaluate the system in a cross-corpus setting. Another challenge in any real-life corpus is the distribution between emotional manifestation *vs*, neutrality, which makes the classification task more difficult. The cross-corpus study has most commonly been studied in other disciplines including ASR [196, 197], parser [198], dialog-act classification [199], and subjectivity identification [200].

The study of cross-corpus emotion classification has also been done previously as reported in [201–206]. In [201], Shami et al. utilized two different speech corpora and conducted experiments in three settings such as intra, inter and mixed. Their study suggests that mixed approach provides better results compared to inter corpus setting. In an another study [202], Shami et al. used two groups of four corpora, which consist of adult-directed-to-infant and adult-directed-to-adult. In this study, they also report that mixed approach is superior to inter-corpus settings. In [203], Eyben et al. studied four corpora for a cross-corpus emotion classification experiment, where they conducted four binary classification experiments for the three classes. Evaluation has been performed using leave-one-corpus-out cross validation. They report that using only acoustic based feature it is possible to obtain a significant improvement compared to chance level. A notable study conducted by Schuller et al. [204], where they employed six corpora for cross-corpus study, which includes 1820 different combinations of evaluations. They report the results with intra and inter corpus and with different normalization techniques. Their results suggest that with acted, prototypical or induced emotional content performs better to a certain degree compared to the spontaneous data.

The study of Lefter et al. [205] focused on exploring the generality, portability, and robustness by utilizing three acted corpora and one naturalistic corpus. They also conducted experiments with intra, inter and mixed corpus, and fusion of classifiers. Their findings suggest that mixing acted data for training, and testing on real data does not help, which might be due to the sources of variation such as language and channel. They obtained better performance with their fusion experiments and recognition rate is better for short utterances ($\sim$ 2 seconds). In [206], Zhang et al. studied an unsupervised learning approach in cross-corpus emotion recognition using six commonly used corpora, which include acted and natural emotion. They report that adding unlabeled emotional data to a mixed corpus for training improve classification performance.

Few things are common in most of the studies, which include intra, inter and mixed corpus settings for training, and a combination of acted, induced and natural emotional corpora with binary classification experiment in many cases. The reported results are better with mixed corpus setting while the performance of inter corpus settings experiments are always are lower.

In this study, we utilized two real corpora with naturalistic emotions. We conducted our experiments with intra, inter and mixed corpus settings while also explored binary and multi-class classification. These corpora are also from two different domains. One is call center human-human phone conversations and the other is human-machine (children interacted with the robot), consisting in different languages such as Italian and German. The main problem in any cross-corpus experiment is the mismatch between class labels due to the fact that each corpus has been designed by focusing on the specific problem. To cope with this problem the common approach is to map class labels or use clustering schemes [204]. For our study, we mapped customer channel's emotions of SISL corpus with the Aibo corpus.

9.2 Dataset

9.2.1 SISL Emotion Corpus

The corpus includes 1890 randomly selected customer-agent conversations, which were collected over the course of six-months, amounting to 210 hours of speech data. The corpus was annotated with respect to a set of emotions including basic emotions such as *anger*, and complex social emotions such as *satisfaction, dissatisfaction, frustration* and *empathy.* The *neutral* tag was introduced as a relative concept to support annotators in their perceptual process while identifying the situation of the context. Empathy was annotated on the agent channel and other emotions on the customer channel. The annotation has been done on the segment level, with neutral followed by an emotional manifestation, for example, Neutral→Empathy. An automatic speech *vs* non-speech segmenter has been utilized to segment the conversation and mapped manual annotation of emotion labels into automatic segments. The inter-annotator agreement of the annotation is kappa = 0.74. For this study, we only used customer channel's emotions. More details of this corpus are discussed in Section 3.1.

9.2.2 Aibo Robot Corpus

FAU-Aibo Robot Corpus [28], is one of the publicly available human-machine real-life, spontaneous, corpus containing recordings where children are interacting with Sony's pet robot Aibo. The experimental setup was designed based on Wizard-of-Oz, in which Aibo was fully remote-controlled by the experimenter. The experimental setup made the children believe that Aibo was responding to their commands. Aibo's actions were very predetermined, which caused children to manifest emotions. The data consists of recording from 51 children, which has been collected from two different schools. The recording contains 9.2 hours of speech with 16 *bit*, 16 kHz. The recordings were segmented automatically into turns using a pause threshold of $1sec$. The annotation has been done at the word level and then combined them into chunk level. It has been distributed into different set such as interspeech emotion challenge set [102] and CEICES [28]. For the purpose of this study, we utilized interspeech challenge dataset with 2-classes and 5-classes for the experiments.

9.2.3 Cross Corpus Emotion Mapping

Since annotation scheme is different for these two corpora, therefore, we mapped the emotions of both corpora to have an aligned emotion set. In SISL Emotion corpus, we have instances with negative emotions along with neutral. The negative emotion includes anger, frustration, and dissatisfaction. Whereas in FAU-Aibo corpus, there is a set containing instances with neutral and negative emotions. The negative emotions include angry, touchy, reprimanding, and emphatic and neutral. For the two-class problem, the class label mapping between two corpora is provided in Table 9.1. For the four-class problem, we mapped negative, neutral, satisfaction (positive) and other emotions of the SISL corpus with negative, neutral, positive and rest (other) emotions of the Aibo corpus as shown in Table 9.2. A train-test split with their class label distribution is presented in Table 9.4 and 9.5 for the 2-classes and 4-classes respectively.

There are significant variations between two corpora in terms of language, recording conditions, channel, data collection scenarios, type of interaction, speaker age, and annotation procedures as presented in Table 9.3. Theses variation poses a great challenge in classification experiments, for example, the speakers of the SISL corpus are adults and children in Aibo corpus. There is a significant difference in the pitch range between adults $(70 - 400Hz)$ and children $(300 - 1000Hz)$ [207]. The duration distribution (mean±standard-deviation) of the segments for SISL is 1.76 ± 0.87 and Aibo is 1.76 ± 0.82.

Table 9.1: Emotion class label mapping between SISL and FAU-Aibo emotion corpora for the 2-class classification experiment.

Class	**SISL**	**FAU-Aibo**
Neg	Negative (anger, frustration, dissatisfaction and not-completely dissatisfaction)	Negative (angry, touchy, reprimanding, and emphatic)
Neu	Neutral	Neutral

Table 9.2: Emotion class label mapping between SISL and FAU-Aibo emotion corpora for the 4-class classification experiment.

Class	SISL	FAU-Aibo
Neg	Negative (anger, frustration, dissatisfaction and not-completely dissatisfaction)	Negative (angry, touchy, reprimanding, and emphatic)
Neu	Neutral	Neutral
Pos	Satisfaction	Positive (motherese and joyful)
O	Other	Rest

Table 9.3: Characterstics of the two different corpora

Charterstics	**SISL**	**Aibo**
Language	Italian	German
Recording condition	telephone	Microphone
Sample rate, bits	8kHz, 16bit	48kHz→ 16kHz, 16bit
Channel	2 channels	1 channel
Data collection scenerios	Call center data with problem solving or information seeking	Wizard-of-Oz - children interacting with remotely controlled pet robot
Interaction type	Human-human	Human-machine
Speaker	Adult	Children
Annotation	Segment contains more than one turn	Word level annotation turned into turn/chunk level

9.3 Experimental Methodology

For the classification experiments, we used acoustic features and used SVM to design the classifiers. For all of the classification experiments, we first evaluated the system's performance using the same corpus then evaluated with cross-corpus. For the experiment, no feature selection and tuning have been performed. The motivation was not to improve or compare the performance with the previous study rather understand how the system performs in a cross-language, domain settings. For the multi-class experiment, we used pair-wise classification approach. The train and test split for both corpora in terms of 2-classes and multi-classes problem are presented in 9.4 and 9.5 respectively. For the train-test split of Aibo corpus, we maintained in the same split that used in interspeech emotion challenge [102], whereas for the SISL corpus train and test split has done at the conversation label with 85% *vs* 15% respectively. The class distributions are very skewed for SISL corpus compared to Aibo corpus and it effects the within corpus classification performance, which we will see in Section 9.4.

Table 9.4: Class distribution between SISL and FAU-Aibo emotion corpora for the two-class classification experiment.

Corpus	**Dataset**	**Neg**	**Neu**	**Total**
Aibo-2-class	Train	2974	5590	8564
	Test	2119	5277	7396
	Total	5093	10867	15960
SISL-2-class	Train	678	3477	4155
	Test	154	5888	6042
	Total	832	9365	10197

Table 9.5: Class distribution between SISL and FAU-Aibo emotion corpora for the four-class classification experiment.

Corpus	**Dataset**	**Neg**	**Neu**	**Pos**	**O**	**Total**
Aibo-4-class	**Train**	2974	5590	674	721	9959
	Test	2119	5377	215	546	8257
	Total	5093	10967	889	1267	18216
A4E-4-class	**Train**	678	3477	1836	614	6605
	Test	154	5888	510	142	6694
	Total	832	9365	2346	756	13299

9.3.1 Feature Extraction

For this study, we utilized only acoustic features. We exploited a large set of acoustic features, in which low-level features were extracted and then projected them onto statistical functionals [133, 141, 160].

The low-level acoustic features include the feature set of the computational paralinguistic challenge's feature set [179], Geneva minimalistic acoustic feature set [208] and formant features. We extracted low-level acoustic features at approximately 100 frames per second. Regarding voice-quality features the frame size was 60 milliseconds with a gaussian window function and $\sigma = 0.4$. Regarding other low-level features the frame size was 25 milliseconds with a hamming window function. The details of the low-level features and their statistical functional are provided in Table 4.1. After feature extraction, the size of the resulted feature set is 6861.

9.3.2 Classification and Evaluation

In this study, we designed classification models using SVM [136] with its linear kernel and used its default parameter. We measured the performance of the system using the Un-weighted Average (UA). It is the average recall across class labels.

9.4 Results and Discussion

In Table 9.6, we present the classification results with binary classification experiments. With SISL corpus there is only 10.32% relative difference between in-domain and cross-domain's results. However, there is a variation in the performance for both negative and neutral classes between in-domain and cross-domain case. With Aibo corpus, there is 17.32% relative performance difference between in and out domain's results.

One might compare the results of Aibo test set with the performance reported in [173] for the binary classification task, which is 70.3 (UA). However,

Table 9.6: Results (UA) of cross-corpus classification experiments with 2-class problem.

Exp type	Train	Test	UA
Intra	SISL train set	SISL test set	58.1
	Aibo train set	Aibo test set	63.5
Inter	SISL train set	Aibo test set	52.1
	Aibo train set	SISL test set	52.5
Mixed	SISL+Aibo train set	Aibo test set	63.8
	SISL+Aibo train set	SISL test set	52.7

Table 9.7: Results (UA) of cross-corpus classification experiments with 4-class problem.

Exp type	Train	Test	Avg
Intra	SISL train set	SISL test set	33.0
	Aibo train set	Aibo test set	37.3
Inter	SISL train set	Aibo test set	20.1
	Aibo train set	SISL test set	25.4
Mixed	SISL+Aibo train set	Aibo test set	38.0
	SISL+Aibo train set	SISL test set	33.9

it is not exactly comparable for a few reasons. First, the binary classification results reported in [173] are with negative and idle, where idle category includes all non-negative emotion classes including neutral. Whereas here, we used instances of the neutral class. Second, we have not done any optimization in terms of feature selection and parameter tuning, which has been done in [173].

When we move to a more complex problem, from binary to multi-class problem performance drops significantly as we can see the results in Table 9.7. For both corpora, results are lower compared to binary classification results. In this setting, results with Aibo corpus is better compared to SISL corpus. However, results on the negative class drop significantly when SISL test set was evaluated using the Aibo trained model.

9.5 Summary

In this chapter, we present our study of cross-corpus emotion classification using two real-life natural emotion corpora. We found very promising results in binary classification experiments in all settings such as intra, inter and mixed. Where as performance significantly drops for multi-class classification experiments. For this study, we only investigated acoustic features. Future work includes investigating phonetic features across corpora.

Chapter 10

Summary: Affective Behavior

In the first part of the thesis, we discussed our work on the design of computational models for detecting affective behavior, which we evaluated using real-life call-center data. Our research contributions include the investigation of conversation and segment level classification experiments for both agent and customer side's emotional states. The agent side emotional states include empathy and neutral (non-empathy) and the customer side's emotion include anger, frustration, satisfaction and neutral.

For conversation level classification experiments, we designed binary classifiers for each emotional states to detect the presence or absence of an emotional state. For the segment level classification, we have two different experimental settings for agent and customer channel's emotional manifestations - binary classification for agent channel's emotion and multi-class classification for customer channel's emotion. We obtained an average recall (UA) of 70.1% on segment level empathy classification, i.e, empathy vs neutral task. For the segment level classification model of the customer channel's emotion, the UA is 56.9%. For segmentation and labeling segment, we also explored HMM based sequence labeling technique, which presents novel directions for future research.

One of the main novelties of this thesis is modeling empathy, which has not been done on focusing call-center domain. In addition to designing computational models for both agent and customer side's emotion, we also present a pipeline for designing affective scene i.e., emotional sequence, for the whole conversation. From the affective scene, one can analyze different patterns from a large set of conversations, which presents some affective insights of the conversations. An example of such a pattern is, "who manifested emotion at the start of the conversation and emotional manifestations at the end of the conversation (e.g., positive or negative)".

For designing above computational models, we investigated acoustic, lexical and psycholinguistic features, at the feature- and decision- level combination and observed that decision level combination performs better than feature level combination. In different tasks, the performance of acoustic and lexical features varied. For the agent channel, acoustic features perform better than lexical features and for the customer channel, it is vice-versa. Due to the complexity of the tasks i.e., binary on agent channel and multi-class on the customer channel, performance also varies.

We also explored the generability of those models i.e., how such models can perform well in cross-language and domain settings. We present promising research avenues towards this direction.

Part II

Personality

In the this part of the thesis we present our study of personality. As mentioned eralier, the personality computing research is concerned about the three main problems in order to study personality such as *recognition*, *perception* and *synthesis*. Our line of research mainly focused on personality *recognition* and *perception*, in which we explored different distal cues such as verbal and (vocal) non-verbal from a varity of domains combined with mono and multi-modal channels. In chapter 11, we present the state-of-the-art of personality computing. In personality traits study, we mainly focused on social interactions such as Facebook statuses, human-human spoken conversation, broad-cast news and youtube-blog, which we discussed in Chapter 12, 13, and 14, respectively. In an another study, we explored the association of mood, personality and communication style, which we discussed in Chapter 15. We conclude this part with a brief summary in Chapter 16.

Chapter 11

State-of-the-Art: Personality

It has been a long-term goal for psychologists to understand human personality and its association with behavioral differences. In personality psychology, the goal is to find the most important ways in which individual differ from among the common nature and infinite dimensions of characteristic differences, which are measurable and remains stable over time [53]. The research on personality computing has attracted attention from several fields, the most notable of which are human-machine interaction, health diagnosis and the newly emerging field of behavioral analytics.

Researchers in psychology have been trying to understand human personality in the last century, and it has become one of the sub-fields of psychology. Later, the study of personality has become one of the central concerns in different fields such as organizational and social science, health-care and education [53]. Since the late 19th century, psychologists have been trying to define theories, rating scales, and questionnaires by analyzing lexical terms or biological phenomena [53, 209]. In personality psychology, *personality trait is defined as the coherent pattern of affect, behavior, cognition and desire over time and space, which are used to characterize a unique individual* [53, 210]. It is evident in the literature that traits are useful in predicting mental-health, marital satisfaction, job-success and mental disorder [211, 212].

Recent work on personality computing has shown how people's personality is expressed and how it can be predicted and applied in different contexts. In spoken interaction, a user's personality [213] can be predicted, which can increase the possibility of natural interaction. Job performance can be predicted by studying an applicant's personality [214] (especially, conscientiousness and neuroticism dimensions). Conversational expressions in video blogs can be analyzed to understand a blogger's personality [215]. Research findings also show that personality is closely associated with romantic relationships [216], preference of genre of music [217], and consumer choice of brands [218].

The advancement of personality traits theories and simplified inventories opened a window for its automatic processing. Hence, in the last few years, automatic personality traits recognition has become one of the mainstream topics in the field of speech and natural language processing to ease the process of interaction between human and virtual agents. Also, it adds value in different areas, such as virtual assistants, healthcare, detection of personality disorder, recommender systems such as customer profiling. Currently, it is used commercially to facilitate job recruiter in their recruiting process.

11.1 Theories of Personality Traits

Aristotle was the first to study personality, in the fourth century BC. Later, his student Theophrastus described thirty personality types, which might be considered as traits, as reported by Rusten (1993) (see [35]). In (460–377 BC) Greek physician Hippocrates also studied the traits (see [219]). Since then, psychologists have been trying to define theories from many different perspectives such as psychodynamic, humanistic, trait, behaviorist and cognitive [220].

This thesis follows the trait approach to personality. Traits theorists support to the belief that traits are stable over time. An individual's behavior varies naturally depending on the context. However, there is a consistency in those manifestations, which defines individual's true nature.

Sixteen Personality Factor(16PF) The study of trait perspective has been done by empirically analyzing lexical terms or biological phenomena. Lexically driven studies were done by Allport and Odbert [221] in which they identified 18000 personality relevant lexical terms. Cattell identified sixteen source traits by the factor-analytic approach and defined Sixteen Personality Factor (16 PF) Questionnaire [222]. This questionnaire has been used to assess the traits of an individual. Using this approach trait are represented in bipolar form, i.e., high or low, and each trait is measured using lexical terms that we use in everyday conversation. Since then verbal report in the form of questionnaire became the preferred method for personality researchers to measure the traits.

Eysenck's Three Factor Model According to the Hans J. Eysenck, the core of personality consists of three broad traits such as *introversion–extraversion*, *neuroticism–stability*, and *psychoticism* [223]. Eysenck Personality Questionaire-Revised (EPQ-R) has been defined for the assessment of these traits. Eysenck's studies mostly dealt with the biological phenomena, such as cerebral cortex, which is associated with personality.

The Big Five Among the various theories of personality traits, Big-5 is the most widely used and accepted model [35]. The Big-5 framework (Big-Five or Five-Factor-Model) describes human personality as a vector of five values corresponding to bipolar traits, as defined below. Big-Five has been defined based on lexical approach as it is believed that personality attributes are encoded in the natural language [224–227]. A high-level description of each trait is presented below.

Openness to experience (O): An appreciation for art, emotion, adventure and varying experience. It estimates the degree to which a person considers new ideas and integrates new experiences in everyday life. High scored people are presumed to be visionary and curious while low scored people are generally conservative.

Conscientiousness (C): A tendency to show the self-discipline, aim for achievement, having a planned behavior rather than having a spontaneous behavior. People with a high score are considered to be accurate, careful, reliable and effectively planned while people of low scores are presumed to be careless and not thoughtful.

Extraversion (E): Extraverted people are energetic, seeking companies of others and have an outgoing attitude while introverted personalities are presumed to be rather conservative, reserved and contemplating.

Agreeableness (A): Compassionate and cooperative, opposed to suspicion. They trust other people and are being helpful. Non-agreeable personalities are presumed to be egocentric, competitive and distrustful.

Neuroticism (N): A tendency to experience mood swings, easily influenced by negative emotions like anger, depression, etc.

Costa, McCrae, and others have done a large scale empirical research to define a measurement scale and to utilize it with other personality schemes [228]. Their studies resulted in the development of NEO-Personality Inventory Revised (NEO-PI-R) questionnaires, which include 240 questions. The response to each question is made on a five-point Likert scale, from strongly-agree to strongly-disagree. Each trait is composed of six lower-level traits, in which each of them is assessed using eight questions. Hence, it resulted in 240 questions. Answering such questions takes on average 45 minutes to complete.

In many scenarios and situations, it is difficult to use NEO-PI-R because it is too lengthy and in many research area a small number of questions are preferable. Hence, several shorter sets of the questionnaires have been developed. It includes 44-item Big-Five Inventory (BFI) [53], 60-item NEO Five-Factor Inventory (NEO-FFI) [228], 100 Trait Descriptive Adjectives (TDA) [224], which take 5, 15, and 15 minutes to complete, respectively. Five and ten items of the Big-Five traits have been developed by Gosling [217,229], which is also useful in many scenarios. Of course, a large set of questionnaires represents more psychometric properties than smaller one. However, it is important for time-limited scenario or a large scale study.

For measuring personality trait, there are two approaches: a) self-report – is used to rate oneself; b) observer-report – is used to rate others. Both approaches have advantages and disadvantages. The *self-reported* measure is easy to interpret, inexpensive and also easier to collect a lot of data. It also has potential weaknesses such as it has response biases. For example, "I get nervous easily" might be rated as disagree due to the fact that it represents negative characteristics and subject may hide it. *Observer report* is defined to provide ratings that are based on their overall conception of an individual where the observer can be a friend, an acquaintance and/or spouse. One of the main drawbacks is that it is expensive compared to the self-report. Since high correlation has been found between self and observer reported measures, therefore, the later approach is commonly used for the study of personality perception [230].

Self-reported assessment is referred to as personality recognition where as the observer reported one is referred as personality perception. For the observer reported measure it is better to rate each subject by more than one accessors and mutual agreement between judges need to computed via any agreement statistics such as kappa.

11.2 Affective Computing Research of Personality Traits

Affective computing research on personality is relatively recent [231, 232]. Their findings suggest that naturalness of interaction with a user and its efficiency increases by matching user's personality. Since then the study of personality has become one of the mainstream topics in affective computing research for several reasons: 1) to provide personalized services by utilizing the increasing amount digital content containing personal information, 2) to design human-machine interfaces with social and affective ability, and 3) facilitating domain experts such as therapist to enhance their capability of counseling. The personality of the speaker/user/blogger can be automatically recognized from different modalities such as spoken/written conversations, facial expressions, gestures, body movements and also other behavioral signals. In the following sub-sections, we highlight the relevant studies that have addressed in personality computing research and also present a review of different modalities that has been studied for automatic processing.

11.2.1 Personality Computing

Most of the focus of personality computing research was mainly how people behave in social media such as Facebook, Twitter, MySpace, blog and how they use their personal electronic devices in their daily life [215,233–236]. Studies has also been conducted to find the association between personality and how people write essays, email, diary [237]; how one interact with another in telephone conversations [238]; how speaker speak in broadcast news [10]; mobile phone uses [239]. Multi-modal information has been utilized for the automatic prediction of Big-5 personality traits by investigating different scenarios such as self-presentation, human-human and human-machine interaction [240]. An unsupervised approach has been investigated in order to solve the problem of domain adaptability where reference annotation is difficult to achieve [241]. The problem has been addressed using text-based data from social network sites. The study of Kalimeri focused on finding the association of personality and situational factors [242]. In her study, data has been collected by using wearable sensing devices and focusing on non-verbal behavioral expressions.

Personality and its association with Facebook uses has been conducted by many studies [12, 233, 243–246, 246–248]. In [243], authors analyzed how Facebook users post personal information such as name, education, religion and marital status, and how they is related to personality traits. Their study also report traits association with user's text-based profile, the number of organization/groups users belongs to and a number of characters they use to describe their favorite activities. In [244], authors studied 236 Facebook users to conduct whether user profile reflects true personality or self-idealization and the results suggest that user's profile represent their true personality. In [245], Bachrach et al. analyzed the information of user's Facebook profile, which includes the size and density of their friendship network, number of uploaded photos, events attended, group memberships and number of times user were tagged in photos. Their predictive analysis using multivariate regression shows that they obtained the best accuracy for Extraversion and Neuroticism, lowest for Agreeableness, and in between for Openness and Conscientiousness categories. The association between personality, gender and age have been studied by Schwartz et al using the data from 75K Facebook users [233]. Using a part of the same Facebook data, a large-scale comparative study has been conducted in the Workshop on Computational Personality Recognition: Shared Task [12]. Different type of features and classification methods has been investigated in order to improve the prediction accuracy.

Similar studies have been conducted using the data collected from Twitter. In [249], authors used data from 335 twitter users and using only three features such as a number of a follower, following and users in the reading list, they achieved root mean square error in the range from 0.6 to 0.9. The study of Quercia et al. in [250], suggest that popular and influentials users are extroverts and emotionally stable. Their finding also suggests that popular users are high in openness, while influentials tend to be high in conscientiousness. With a goal of identifying influential community Kafeza et al. studied twitter data while mapped personality traits with influential users [251]. The association between personality traits and how people uses the mobile phone and other wearable devices has been studied in different kinds of literature [252–254].

The study of personality computing has been addressed by focusing on three major problems such as *recognition*, *perception* and *synthesis* [255]. In this Section, we will focus on the first two problems as these were the focus of this thesis.

Personality Recognition is the task of inferring self-assessed personality while utilizing all the overt cues [255]. It is traditionally referred to as true personality of an individual [256].

Personality Perception is the task of inferring the personality that observers attribute to an individual. Since assessments are made by other individuals, therefore, it is not considered as true personality, however, it captures the traits that are assigned by multiple raters.

In personality computing research, for both recognition and perception tasks the approaches mainly adopted from affective computing domain. Focusing on a specific scenario in mind and the typical approach include data collection, annotation based on either self *or* observer, extract features focusing on different modality such as speech, text and/or visual, then design the machine learning classifier for the evaluation/prediction.

11.2.2 Modality

For both recognition and perception tasks, different modalities have been investigated in different context based on the availability of the data. For automatic processing, researchers use acoustic, lexical and audio-visual features and have very recently started to use emotional categories [257] and traits [258] as features. Personality plays a role in emotion, and this has been discussed in several kinds of literature in psychology [209]. For the automatic prediction of personality, Mohammad et al. [257] studied emotional features for personality traits prediction and showed that fine-grained emotions are more relevant predictors. Later, Farnadi et al. [248] found a correlation between emotion and personality traits using Facebook status updates and showed that users' posts of *openness* trait convey emotions more frequently than other traits. Below we highlight the studies that are specifically focused on different modality.

Speech

There has been active research since the first study done by Sapir [259] to understand the effect of speech on personality traits. Major contributions have been done in the Interspeech 2012 speaker traits challenge [1, 10, 131, 134, 139, 260–264], where one of the sub-challenges was the recognition of the speaker personality traits. The contributions in the evaluation campaign include studying different feature selection and classification techniques along with combining acoustic and linguistic features. The feature selection algorithms include Sequential Floating Forward Search (SFFS), Principal Component Analysis (PCA), Gaussian Mixture Model (GMM), Supervised/Unsupervised Set Covering Problem (SSCP/USCP), Information Gain, and Fisher information based filtering with a genetic algorithm. The classification algorithm includes Support Vector Machines (SVMs) with different kernels, GMM with Universal Background Model (UBM) and Adaboost. The extracted acoustic features include a very large set of low-level features projected onto statistical functionals, which resulted in 6125 features. A brief overview of the extracted features is presented in Table 11.1. The study by Kartik et al. [260] also transcribed the audio and analyzed the text using psycholinguistic word categories, which are obtained from LIWC [159].

The study in [265] comprised of 96 subjects where observers rated personality to the subjects and the average score is assigned to each subject. In [11], authors analyzed 119 human-human call center spoken conversations from 24 subjects to design the models for automatic personality traits predictions.

Text

Studies have been done on how the style of communication like emails, blog entries [266] and the choice of particular parts of speech [267] depend

Table 11.1: Summary of the features. sma, de(2) indicates that functionals applied after applying (i) moving average filter smoothing and (ii) first delta coefficient separately. sma(1) means, only used moving average filter smoothing before applying functionals. sma-de (1) means, moving average filter smoothing and first delta coefficient applied together, then used functionals.

LLD	**Moving average filter (sma), first delta coefficient (de)**	**Functionals applied to LLD/delta LLD and LLD only**	**Number of features**
4 energy	sma, de (2)	35	4x2x35=280
	sma (1)	23	4x23=92
	sma-de (1)	3	4x1x3=12
54 spectral	sma, de (2)	35	54x2x35=3780
	sma (1)	23	54x23=1242
	sma-de (1)	3	54x1x3=162
6 voicing	sma, de (2)	33	6x2x33=396
	sma (1)	23	6x23=138
	sma-de (1)	3	6x1x3=18
F0 voicing	sma (1)	5	1x5 = 5
		Total	6125

on the author's personality. In [268], authors studied four different types of lexical features such as function word list, conjunctive phrases, modality indicators, and appraisal adjectives and modifiers. They used an essay corpus written by students at the University of Texas at Austin collected between 1997 and 2003. Their classification study shows that function words work better for extraversion whereas appraisal use of words is better for neuroticism. In [237], Mairesse et al. have done a more details study where they investigated spoken transcriptions and written essays for automatic personality traits classifications. They utilized Linguistic Inquiry and Word Count (LIWC) based features, utterance and prosodic features for designing classifiers. Their study also includes personality perception and recognition task with an investigation of different classification algorithms.

Authors personality has been studied using a blog corpus [269], where word-ngrams is used and compared Naïve Bayes and Support Vector Machines (SVMs) classification algorithms. Their findings suggest that Naïve Bayes performs better than SVMs in most of the cases. Naïve Bayes with information gain feature selection algorithm has been investigated in [270] using a Japanese weblog corpus. Bag-of-words and the count of word categories are used as features for automatic analysis using a blog corpus [271]. Their findings suggest that neurotic people use blog to express their strong emotion, extraverted people use it to document their daily life with both positive and negative emotions, openness are more inclined in writing leisure activities, conscientiousness report their daily life, and agreeableness mostly express positive emotions.

Multi-modal

The study of multi-modal cues includes information from audio, visual and linguistic information. From the audio channel other than acoustic information most often transcription has also been extracted. The typical approach of combining different channels include either feature or decision level combination.

It is yet to be discovered which channel of information is highly correlated with traits [55]. Hence, depending on the availability of information researcher has been trying to understand their usefulness. In [215], authors used youtube blog corpus where features from different modalities have been extracted and combined for designing the classifier and evaluating the system. Using this corpus a major contribution has been done in the Workshop on Computational Personality Recognition - 2014, where the participants participated in both classification and prediction task. For the classification task the average best F1 was 0.67 and for the prediction task, the RMSE was 0.76.

The relationship between proxemics, visual attention and personality traits has been studied in [272], where a group of people interacting in a cocktail party. It consists of two 20 and 30 minutes sessions with a total of 13 subjects. The experiment has been conducted in a laboratory with four fixed cameras set at each corner of the room. Different visual and social features have been investigated. Their study suggests that personality can also be detected from a very short-term temporal behavioral sequence, particularly from one-minute information [273]. A multimodal corpus has been studied by Pianesi et al. [274], which consists of 12 multiparty meetings of 4 participants each with audio-visual recordings, for a total length of over 6 hours. It suggests that the manifestation of personality are visible enough in social interaction even from a very short temporal event. It exploited different social interactional features such as activity, emphasis, mimicry, and influence; and visual features such as head, body, and hands fidgeting.

The study of Batrinca et al. reports that personality traits can be detected even from 30 to 120 seconds self-presentations [275]. Their study consists of 89 participants, who were asked to introduce themselves and talk about either job, holiday, preferred food or sport, which has been recorded in front of the camera upto 120 seconds. Their findings suggest that conscientiousness and neuroticism traits can be easily detected during self-presentation due to the fact that the former is related to the engagement with the task activity and the later is related to the emotional reactions. They report that the low accuracy of extraversion and agreeableness are due to the situational variables and the differential activation of the behavioral dispositions. The same audio-visual corpus has been investigated to understand the behavioral cues in meetings such as speaking time and social attention are associated with extraversion. It is reported that these are important cues for the automatic detection of extraversion trait [276]. Several studies have been used user's profile picture to predict personality traits and report that it represent a significant amount of information about user's personality [277–279].

Some other studies include the use of non-verbal cues such as prosody, facial emotional expressions, appearance, salient point, and lexical features for the automatic prediction of personality traits [280, 281].

11.3 Summary

In this chapter, we discussed our review of the current state-of-the-art of personality traits in terms of different theories in psychology and the research in personality computing. Most notable personality trait theories include sixteen personality three-factor model, eysenck's three-factor model, and the Big-5 model. Among them, the Big-5 model is most widely used and studied in personality computing research. In personality computing research, the approach is to employ either self- or observer- assessed traits measure questionnaires to label the user/speaker's data for the automatic analysis. The self-assessed trait measures are termed as personality recognition task, whereas the observer-assessed trait measures are considered as personality perception task. The collected data that has been studied include different modalities such as speech, text (social media conversations or written diaries) and audio-visual. For the automatic analysis, the current state-of-the-art approach is to use machine learning algorithms to learn the patterns, i.e., verbal and non-verbal cues, from the labeled data in order to label unseen data.

Chapter 12

Personality in Social Media Corpus - Facebook

For the social communication, we interact with unknown individuals, even with machines that exhibit human-like features and behavior such as robots, embodied virtual agents and animated characters [232]. To make these automated systems more human-like, we need to understand human behavior and how it is affected by personality traits. It is also evident that there is a strong correlation between users' personality and the way they behave on the online social network (e.g., Facebook).

In this chapter, we discuss our study of automatic recognition of personality traits on the social network data. We studied different classification methods such as SMO (Sequential Minimal Optimization for Support Vector Machine), Bayesian Logistic Regression (BLR) and Multinomial Naïve Bayes (MNB) sparse modeling while used bags-of-words as features. Another contribution is to measure the performance of the systems using two different evaluation measures: (i) macro-averaged precision and recall, F1; (ii) weighted average (WA) and un-weighted average (UA).

12.1 Corpus: Facebook

In the "Workshop on Computational Personality Recognition (Shared Task) 2013" organizer released two gold standard labeled datasets: essays and myPersonality [12]. For this study, we have used myPersonality corpus. The corpus was collected from the social network (Facebook) and contains Facebook status messages as raw text, author information, gold standard labels (both classes and scores) for classification and regression tasks. Annotation of the personality traits has been done using *self-assessed* questionnaire. The data was collected from 250 different users and the number of statuses per

Table 12.1: Number of instances and their distribution (in parenthesis) of class labels of the myPersonality corpus. Y and N represent positive and negative classes respectively.

Class	**Train-set**		**Test-set**	
	Y (%)	N (%)	Y (%)	N (%)
O	4863(74.3)	1682(25.7)	2507(74.3)	865(25.7)
C	3032(46.3)	3513 (53.7)	1524(45.2)	1848(54.8)
E	2784(42.5)	3761 (57.5)	1426(42.3)	1946(57.7)
A	3506(53.6)	3039 (46.4)	1762(52.3)	1610(47.7)
N	2449(37.4)	4096 (62.6)	1268(37.6)	2104(62.4)

user ranges from 1 to 223. Based on the task organizer guidelines dataset has been split into the train (66%) and test (34%). While splitting the data into train and test set we used a stratified sampling (similar proportion of classes in two sets) technique. We used only class labels for personality traits classification. A distribution of the labels in the corpus is given in Table 12.1. Train and test set have different distributions of positive and negative cases in different personality trait categories. In total, there are $6,545$ train and $3,372$ test instances after the split. From the corpus analysis, it is observed that besides words, it contains tokens such as internet-slang (e.g., WTF-what the F***), emoticons (e.g., :-D), acronyms (e.g., BRB-be right back) and various shorthand notations that people use in their status. The maximum number of tokens per user status message is 89, minimum 1 and the average is 14.

12.2 Features

We used bag-of-words approach and used tokens (unigrams) as features, where a classification instance is a vector of tokens appearing in the Facebook status. As discussed earlier, different kinds of tokens (internet-slangs, smiles, emoticons, etc.) are present in the corpus; our assumption is that these tokens carry distinctive information for personality traits recognition. Thus, there was no attempt to remove or normalize them. Using weka's "string to word vector", the text was converted into feature vector using TF-IDF [282] as a feature value. The training set's dictionary obtained using this scheme contains $15,268$ features; the same dictionary was used for the test set. TF-IDF feature valued representation was selected for the fact that it outperformed Boolean feature valued representation on exploratory experiments.

12.3 Experimental Methodology

For the experiments, we used SMO with linear kernel, BLR, and MNB sparse model. The choice of algorithms is driven by their different properties for classification. SMO is chosen due to its fast optimization during the training of SVM and it has a better generalization capability. Another reason for SMO is the high classification accuracy on different tasks reported in the literature [38, 237, 283] on personality traits recognition. BLR uses different priors (e.g., Laplace and Gaussian) in order to avoid overfitting and it produces sparse predictive models for text data. Moreover, it is also widely applied in text categorization. The key idea of BLR is to use prior probability distribution that favors sparseness in the fitted model. Whereas, MNB sparse model is an extension of Multinomial Naïve Bayes generative model where a sparse representation is used to reduce space and time complexity. For the feature extraction and the classification, we used weka [164].

The performance of the system had been evaluated using myPersonality test set. In the shared task guidelines it is suggested to use precision, recall, F1 as evaluation metrics. Additionally, we computed weighted average (WA) and un-weighted average (UA), which are used in recent paralinguistic classification tasks [10]. UA is the average of true positive rate and true negative rate where the average of both poles is considered, whereas WA is the accuracy (Acc).

Even though the suggestion is to use precision, recall, and F1, we have computed macro-averaged precision, recall, and F1 to consider both poles.

Another motivation is that macro-averaged precision, recall, and F1 are inline with UA and WA metrics. Hence, we use the terms Pre-Avg, Re-Avg, F1-Avg, Acc (WA) in this study. Since UA is the same as the average of recall, it is not reported. Pre-Avg, Re-Avg and F1-Avg are computed using the equations 12.1,12.2, and 12.3.

$$Pre(Avg) = \frac{1}{2}\left(\frac{tp}{tp+fp} + \frac{tn}{tn+fn}\right) \tag{12.1}$$

$$Re(Avg) = \frac{1}{2}\left(\frac{tp}{tp+fp} + \frac{tn}{tn+fn}\right) \tag{12.2}$$

$$F1(Avg) = 2 \times \left(\frac{Pre(Avg) \times Re(Avg)}{Pre(Avg) + Re(Avg)}\right) \tag{12.3}$$

where tp, tn, fp, fn are the number of true positives, true negatives, false positives and false negatives, respectively.

12.4 Results and Discussion

In this section, we report and discuss the performances of the classification algorithms on personality traits recognition task. Table 12.2 reports results for SMO, where chance (%) is the accuracy computed by randomly drawing class labels using prior distribution. It is computed 100 times with seed (1-100) and the mean is measured. The results of BLR and MNB sparse classifiers are reported in Tables 12.3 and 12.4, respectively. All classification algorithms perform above chance level baseline across all categories.

Additional to training and test set evaluation, we run 10-folds cross-validation on the training set to predict variability. Table 12.5 reports mean±standard deviation values on the cross-validation run of MNB sparse model.

From the study of the personality traits recognition on the social network data (Facebook status messages), it is observed that MNB sparse generative

Table 12.2: Results on the test set using the SMO (linear kernel) classifier. Chance (%) is the mean accuracy obtained by randomly drawing labels 100 times using the prior distribution.

Class	**Pre-Avg**	**Re-Avg**	**F1-Avg**	**Acc**	**Chance (%)**
O	57.46	58.28	57.68	65.84	61.78
C	58.02	58.09	57.99	58.16	50.36
E	57.47	57.57	57.49	58.21	51.05
A	58.40	58.41	58.40	58.45	50.10
N	56.89	56.99	56.92	59.25	52.94
Mean	57.65	57.87	57.70	59.98	53.25

Table 12.3: Results on the test set using the Bayesian Logistic Regression (BLR).

Class	**Pre-Avg**	**Re-Avg**	**F1-Avg**	**Acc**
O	55.03	55.86	55.02	62.57
C	56.99	57.06	56.90	57.00
E	56.06	56.17	56.02	56.58
A	57.79	57.71	57.68	57.95
N	55.38	55.52	55.41	57.59
Mean	56.25	56.46	56.21	58.34

Table 12.4: Results on the test set using Multinomial Naïve Bayes (MNB) sparse model.

Class	**Pre-Avg**	**Re-Avg**	**F1-Avg**	**Acc**
O	59.83	59.71	59.77	69.48
C	59.06	59.11	59.07	59.34
E	57.99	58.13	57.98	58.57
A	59.09	58.71	58.49	59.16
N	58.84	57.90	57.95	62.40
Mean	58.96	58.71	58.65	61.79

Table 12.5: Results (mean ± standard deviation) on 10-folds cross validation run of the train set using MNB sparse model. Last row represents the overall mean ± standard deviation.

Class	Pre-Avg	Re-Avg	F1-Avg	Acc
O	58.6±1.6	58.4±1.4	58.4±1.5	68.5±1.7
C	59.2±1.4	59.2±1.3	59.2±1.3	59.4±1.4
E	58.2±1.6	58.3±1.6	58.1±1.6	58.6±1.5
A	57.2±1.6	56.9±1.5	56.7±1.5	57.6±1.5
N	59.6±2.1	58.5±1.7	58.6±1.7	63.0±1.9
Over all	58.5±0.9	58.3±0.8	58.2±0.9	61.4±4.5

model performs better than discriminative models, SMO and BLR. Comparing the cross-validation results on the training set (Table 12.4) and the test set results (Table 12.5) using MNB sparse model, the conclusion is that the test set results are within the statistical variation.

Since there are no published results on this particular data set, we report results on other corpora used in the personality traits recognition literature. First, [237] report classification accuracy ranging from 52.75 to 62.52 and the overall of 57.10 with SMO classifier on the essay corpus. Second, [283] reports precision-recall results for both poles and the average recall ranges from 49.00 to 64.50 with the overall of 58.30 on the modern Greek spontaneous text corpus with SMO. Thus, the performance of classifiers on myPersonality data reported in this paper is within the expected range.

We obtained overall macro-averaged precision - 58.96, recall - 58.71, F1 - 58.65 and accuracy 61.79 with our best model. The results of MNB are statistically significant with $p < 2.20E - 16$ when compared to SMO and BLR using Pearson's Chi-squared test. In all of the experiments we used classifiers' default parameters; additional parameter tuning might increase the performance. Additionally, we have conducted an experiment by leave one user group out (LOUGO) cross validation method using all the data set and the obtained results are reported in Table 12.6. The data was randomly

Table 12.6: Results using LOUGO cross validation method using all data with MNB sparse model.

Class	Pre-Avg	Re-Avg	F1-Avg	Acc-Avg
O	59.75	59.88	59.80	69.12
C	60.44	60.42	60.40	60.71
E	59.16	59.29	59.14	59.74
A	59.10	58.73	58.57	59.33
N	59.86	58.94	59.06	63.23
Mean	59.66	59.45	59.39	62.43

split into 10 user groups.

An extension of this study would be combining different classifiers' results where an upper bound of the overall accuracy would be 76.45 ± 2.63, which was obtained using an oracle experiment.

12.5 Summary

In this chapter, we present our baseline study to automatically recognize BIG-5 personality traits on the social network data (Facebook status messages), which was self-reported. We explored different classification methods. We observed that MNB sparse model performs better than SMO and BLR. We report system performances using macro-averaged precision, recall, F1, and accuracy (WA). Future directions of this study include integrating syntactic, semantic and statistical features; studying feature selection and classifier combination methods, which may lead to provide more information to recognize personality traits.

Chapter 13

Personality in Broadcast News and Spoken Conversation

In spoken language communication, speech signal provides important information for analyzing and modeling human behavior. This speech signal carries rich information about a variety of linguistic and paralinguistic phenomena, which is encoded with different behavioral cues [6], including emotion, intent, traits.

This chapter presents the study of Big-5 personality traits classification by exploring different domain and modalities with focusing on personality trait recognition and perception. The domain includes broadcast news and human-human call center spoken conversation.

13.1 Corpora

13.1.1 Speaker Personality Corpus

Speaker Personality Corpus (SPC) was obtained from the organizers of the Interspeech 2012 Speaker Trait Challenge [10]. The data set consists of training, development and test set. Each instance is labeled with Big-5 traits and each trait is mapped into two classes, positive and negative. This corpus consists of 640 audio files, that were randomly collected from the French news bulletins, broadcasted in February 2005, with the quality of 16 bit, 8kHz sample rate. Out of those clips, professional speakers were produced 307 clips and 333 clips were from 210 non-professional speakers. Only one speaker was used for each audio clip and there were altogether 322 individual speakers. By utilizing the *observer rating* instruments, the corpus was assessed by 11 judges by listening to all the clips and individually evaluated the clips using BFI-10 [256]. The judges did not understand French, so the personality assessment could only be motivated by the nonverbal behavior. The dataset

Table 13.1: Train, development and test splits of SPC corpus. A number of instances on each set and for each class and their distribution. O-openness, C-conscientiousness, E-extraversion, A-agreeableness, N-neuroticism. Y and N represent positive and negative classes for each trait.

Class	**Train**		**Dev**		**Test**	
	Y (%)	**N(%)**	**Y (%)**	**N(%)**	**Y (%)**	**N(%)**
O	97 (39.27)	159 (40.46)	70 (28.34)	113 (28.75)	80 (32.39)	121 (30.79)
C	110 (37.93)	146 (41.71)	81 (27.93)	102 (29.14)	99 (34.14)	102 (29.14)
E	121 (37.81)	135 (42.19)	92 (28.75)	91 (28.44)	107 (33.44)	94 (29.38)
A	139 (43.03)	117 (36.91)	79 (24.46)	104 (32.81)	105 (32.51)	96 (30.28)
N	140 (44.03)	116 (36.02)	88 (27.67)	95 (29.50)	90 (28.30)	111 (34.47)

also consists of extracted acoustic features from those speech files and then later speech has been manually transcribed for lexical based experiments [163]. The SPC train, development and test set consists of 256, 183 and 201 instances respectively. The distribution of the corpus is quite balanced as presented in Table 13.1. We used this corpus for the *personality perception* experiment as presented in Chapter 13.

13.1.2 Persia Corpus

Personable and Intelligent virtual Agents (PerSIA) [11] corpus is an Italian human-human spoken dialog corpus, recorded in a simulated tourist call center. Speakers played randomly the “customer” and the “agent’s” role over a telephone conversation. Each customer was given a tourism task to perform and the agent provided relevant answers. The task scenarios’ difficulty ranged from easy to no-solution [11]. Out of the 24 speakers, 12 were users and 12 were agents. Personality label was assigned based on the *self-assessment* questionnaire during the data collection. At the end, out of 144 (each user and agent) calls, 119 calls of Agent sub-corpus were used in the experiment. A distribution of the corpus is presented in Table 13.2.

Table 13.2: A number of instances and class distribution of the Persia corpus. O-openness, C-conscientiousness, E-extraversion, A-agreeableness, N-neuroticism. Y and N represent positive and negative classes for each trait.

Class	**Total**	**Y**	**N**	**Y %**	**N %**
O	119	74	45	0.62	0.38
C	119	100	19	0.84	0.16
E	119	59	60	0.50	0.50
A	119	78	41	0.66	0.34
N	119	59	60	0.50	0.50

13.2 Experiments: Acoustic Features

We present a comparative study of automatic speaker personality trait recognition from speech corpora that differ in the source speaking style (broadcast news vs. conversational) and experimental context. We evaluated different feature selection algorithms such as information gain, relief and ensemble classification methods to address the high dimensionality issues. We trained and evaluated ensemble methods to leverage base learners, using three different algorithms such as SMO (Sequential Minimal Optimization for Support Vector Machine), RF (Random Forest) and Adaboost. After that, we combined them using majority voting and stacking methods. Our study shows that performance of the system greatly benefits from feature selection and ensemble methods across corpora.

This study follows previous research [11,237,260] on designing algorithms to extract features from the speech that best predict SPTs as well as machine learning algorithms that tackle the high-dimensionality and variability of the classification problem. In particular, this study comparatively evaluates Speaker Personality Traits (SPTs) automatic recognition algorithms on two speech corpora drawn from different speaking styles and data collection conditions. We evaluated the SPTs specific feature selection algorithms as well as their impact on the base and the ensemble classification systems.

We conducted several experiments for this comparative study and to examine the performance of different feature selection and classification algorithms. For the experiment, we first extracted acoustic features and then used feature selection algorithms to select a subset of features. After that, we applied ensemble methods as opposed to say ‘classifier combination methods’ for the final classification, which is explained in Section 13.2.3. It is evident that ensemble methods have also been studied for emotion and personality traits recognition from speech [260, 284]. A conceptual design of the system is given in Figure 13.1.

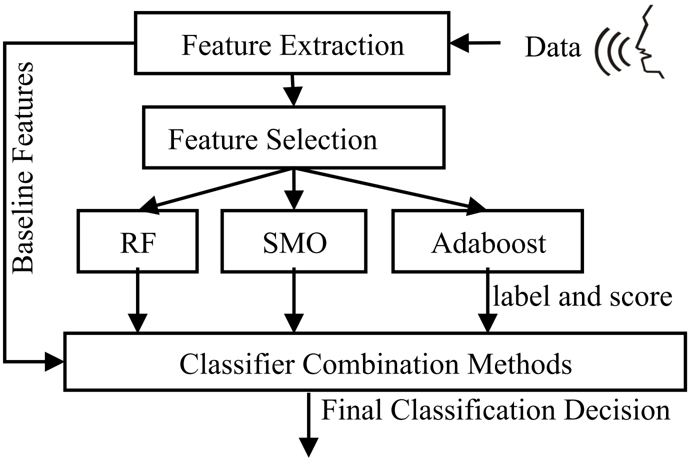


Figure 13.1: System architecture for the classification of personality traits. Exploited different classification and feature algorithms.

13.2.1 Acoustic Features

We extracted acoustic features using openSMILE [285] with the predefined configuration file provided in the Interspeech-2012 Speaker trait evaluation campaign. The low-level acoustic features extracted with approximately 100 frames per second with 10-30ms per frame. These low-level descriptors (LLDs) were then projected on single scalar values by descriptive statistical functionals [38]. More details of the acoustic features can be found in [10]. For this study, we denote these acoustic features as our baseline features.

13.2.2 Feature Selection

We have high-dimensional problems p>>N, the number of features p, (6125) is much larger than the number of instances N. Therefore, to avoid high variance and overfitting we worked on two different feature selection techniques such as Information Gain (IG) [286] and Relief [158] along with equal frequency discretization method. Feature values were discretized into 10 equal frequency bins before applying feature selection algorithms. All acoustic features were continuous valued and converted into discrete value. This is because some feature selection algorithms like IG is not able to handle continuous value. Additionally, we applied discretization for relief feature selection as we were getting better results after applying discretization.

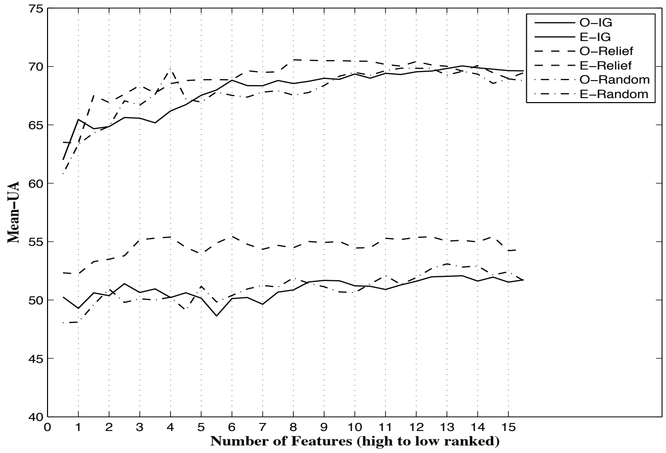


Figure 13.2: IG, relief and random feature selection learning curves with SMO classifier for the O and C categories, which shows different patterns. In x-axis, each point represents multiple of 400 top ranked features from left to right, whereas y-axis represents mean-UA of the LSGO cross validation.

To identify the top ranked most informative features using these feature selection algorithms we generated feature learning curves by incrementally adding top ranked features. These learning curves were generated using our chosen classification algorithms – RF, SMO, and Adaboost. From the feature learning curves we were able to figure out what range of feature we should select for different categories of personality traits. Figure 13.2 shows an example of feature learning curve for SPC using IG, relief and random feature selection with SMO classifier, where random feature selection was considered as a baseline study. In each learning point we also computed standard deviation from the cross validation results to see the statistical variation. Each of the feature selection algorithms behaves differently for each personality trait with different classification algorithms. Therefore, for different personality traits and for different feature selection algorithms we selected different number of features.

13.2.3 Ensemble of Classifiers

For the classification of personality traits we conducted experiments with ensemble (classifier combination) methods where to design base learners we used RF, SMO and Adaboost (Ada). Ensemble methods were chosen due to their higher generalization ability [287] than just a single base learner. We choose three different classification algorithms in ensemble methods because of their different characteristics in classification. SMO [136] is chosen for its fast optimization during the training of SVM and it has a better generalization capability. RF [137] is a combination of tree predictors and it builds a series of classification trees and each tree on its own makes a prediction. These predictions vote to make the RF prediction. RF reduces variances in classification by randomizing features and training instances. Adaptive Boosting (Adaboost) [138] is a meta-learner that uses greedy search for a linear combination of classifiers by overweighting the examples that are misclassified by each classifier. Similar to RF, Adaboost also reduces variances by randomizing the training instances. We used weka [288] for feature selection and classification. As combiners in the ensemble methods, we conducted an experiment using majority voting and stacking. Voting is the most popular and fundamental combination method for nominal outputs and the majority

Table 13.3: Baseline results on the SPC dev set using baseline features with RF, SMO and Ada.

Class	UA-RF	UA-SMO	UA-Ada
O	58.5	60.4	60.5
C	71.6	71.6	72.2
E	81.9	82.0	78.7
A	65.8	66.3	59.0
N	68.7	68.7	62.7
Mean	69.3	69.8	66.6

vote [164] is computed using the equation 4.3.

Stacking [287] is a general procedure where a learner is trained to combine the base learners and the combiner is called second level learner or meta-learner. To train the meta-learner we used LSGO (leave speaker group out) cross validation. In LSGO, speakers were drawn randomly to make groups and the instances of the speaker groups were selected for the train and test set by leaving speaker-group-out approach. Base level classifier's decision and class probability were used as features in the meta-learner and we designed meta-learner using multi-response linear regression (MLR) [164].

13.2.4 Results and Discussion

We evaluated BIG-5 personality traits binary classification models on both the SPC and the Persia corpora.

Baseline Results

Baseline results were measured using all the acoustic features (baseline features) for both the SPC and the Persia corpora. The SPC corpus was evaluated using the SPC dev set and we obtained baseline results using baseline features with RF, SMO and Ada as shown in Table 13.3. We estimated the performance of the SPC dev set by using LSGO cross validation on the SPC train set. For the evaluation of the Persia corpus, we used micro-averaged LOSO cross validation. Table 13.4 shows the results using baseline features

Table 13.4: Micro-averaged baseline results on the LOSO cross validation using baseline features of the Persia corpus. Chance (%) is the performance of the randomly drawing labels.

Class	UA-RF	UA-SMO	UA-Ada	Chance %
O	44.5	45.5	26.6	53.0
C	54.5	52.1	73.2	73.2
E	56.4	58.9	58.9	50.0
A	53.7	63.3	56.2	54.8
N	48.1	45.4	44.6	50.0
Mean	51.4	53.0	51.9	56.2

with RF, SMO, Adaboost (Ada) and Chance [11]. Chance (%) is the performance computed by randomly drawing labels using the prior distribution, more details can be found in [11].

Feature Selection Results

After applying feature selection methods IG and relief on the SPC corpus we obtained improved results using relief feature selection with SMO. Table 13.5 shows the results on the SPC dev set using relief feature selection where we obtained better results with SMO. However, performance had been dropped in the agreeableness category. Similarly, for the Persia corpus, we obtained improved results using relief feature selection with SMO as shown in Table 13.6. Though, after feature selection, performance had been reduced in conscientiousness category using RF and Adaboost, and in neuroticism category using Adaboost.

Ensemble Methods

Table 13.7 shows the results of the SPC corpus with the ensemble of majority vote where classifier ensemble is formed by the best models of three classification algorithms: baseline features for RF, relief feature selection for SMO and baseline features for Adaboost. We used same models for stacking and obtained mean-UA: 69.0 for Big-5 traits.

With the Persia corpus, the ensembles (majority vote and stacking) of the

Table 13.5: Results on the SPC dev set with relief feature selection. Feat-* represents number of features selected for RF, SMO and Adaboost.

Class	Feat-RF	UA-RF	Feat-SMO	UA-SMO	Feat-Ada	UA-Ada
O	1200	61.2	1200	63.4	600	56.3
C	2200	73.2	2200	75.5	1000	66.0
E	1000	79.2	3200	84.2	1200	74.9
A	400	63.4	3800	65.4	800	56.4
N	400	65.6	1800	69.8	400	61.5
Mean		68.5		71.6		63.0

Table 13.6: Micro-averaged results on the LOSO cross validation using relief feature selection on the Persia corpus

Class	Feat-RF	UA-RF	Feat-SMO	UA-SMO	Feat-Ada	UA-Ada
O	2200	47.2	2400	47.0	600	46.7
C	4800	50.6	800	74.6	1200	47.6
E	3600	64.8	200	64.7	200	58.8
A	1400	56.8	2400	71.8	200	69.7
N	3200	51.3	3000	54.6	1600	42.1
Mean		54.1		62.6		53.0

Table 13.7: Results on the SPC dev set using the ensemble of the majority vote, which is comparable with [1]. With UA: SPC train is the mean results across the LSGO cross validation runs and for all traits we obtained 63.2±3.7 (mean±std).

	Our results		**Results [22]**
Class	**UA: SPC train**	**UA: SPC-dev**	**UA: SPC-dev**
O	52.5	65.2	67.0
C	67.2	75.3	73.2
E	70.8	83.0	80.9
A	57.8	66.0	69.0
N	67.7	69.2	71.0
Mean	63.2	71.7	72.2

three best models (relief feature selection with three different classifiers, Table 13.6) we obtained mean-UA: 56.4 with a majority vote, and mean-UA: 49.1 with stacking. However, we obtained improved results with ensemble methods on extraversion (majority voting: UA-67.3) and agreeableness (stacking: UA-80.2) categories. The reason of poor performance is the higher correlation between lower performing classifiers (e.g., RF and Ada). Applying weighted majority voting could probably alleviate this problem, where proper weight needs to assign to the individual classifier.

The results of SPC on dev set are comparable with the results in [1], where our system performs better in conscientiousness and extraversion categories. However, overall, in five categories of OCEAN traits, our results are close to their results in [1]. From the cross validation on the SPC-train set it is observed that our results are within statistical variation 63.2 ± 3.7 (mean±std) in all traits. Another difference is that, in [1], they obtained their best results by considering the best models and using the majority voting of all of their models they did not obtain better results compared to this study. For the Persia corpus, the results in [11] showed the performance in terms of WA_{micro} where they obtained overall 57.5 and we obtained 64.4 with our best system (SMO with relief feature selection).

13.2.5 Conclusion

We investigated automatic recognition of SPTs from speech using two different corpora – conversation and broadcast news. We studied different feature selection techniques such as IG and relief with different classification algorithms. It is observed that relief with SMO performs better than other models on both corpora and also relief feature selection performs well than IG. We obtained better results using majority voting ensemble method on the SPC corpus. Moreover, the stacking ensemble method did not perform well in any corpus with all personality traits categories. Future directions of this study include integrating linguistic information, understanding feature overlap in different feature selection algorithms and studying the contextual information.

13.3 Experiments: Multiple Feature sets

In the previous section, we presented different feature selection and ensemble methods for the personality trait perception experiment. In this section, we present our study of different types of speech features to the automatic recognition of Speaker Personality Trait (SPT) using the broadcast news and spoken conversation speech corpora. We have extracted acoustic, linguistic, and psycholinguistic features and modeled their combination as input to the classification task. For the classification, we used Sequential Minimal Optimization for Support Vector Machine (SMO) together with Relief feature selection.

Following our previous study, our goal here is (a) to understand the prediction capability of linguistic and psycholinguistic features in addition to acoustic features, (b) analyze the feature fusion technique to get the best prediction and c) evaluate our algorithms across different speech corpora. There are several studies that show how personality manifests in word usage [237, 289, 290]. This has indeed motivated us to use linguistic and psycholinguistic features in this context.

For the study of broadcast news and spoken conversation speaker personality traits (SPC) and Persia corpus has been used. Different feature sets such as acoustic, linguistic and psycholinguistic (LIWC) has been extracted, and then generated and evaluated models for each feature set. Then experiments have been conducted with the different feature fusion techniques and eventually selected a fusion technique where all feature vectors were combined to form a single vector and apply feature selection followed by classification. A conceptual design of the system is given in Figure 13.3.

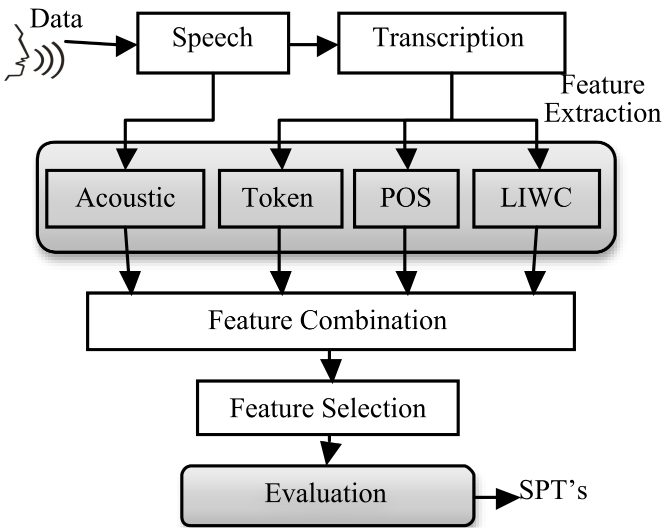


Figure 13.3: System architecture for the classification of personality traits using different feature sets.

13.3.1 Features Type

Acoustic Features

These features were extracted using openSMILE [285] with the predefined configuration file (IS2012.conf), which was provided in the Interspeech-2012 speaker trait evaluation campaign. The low-level acoustic features were extracted with approximately 100 frames per second, with 10-30 milliseconds per frame. These low-level descriptors (LLDs) were then projected onto single scalar values by descriptive statistical functionals [38]. More detail on the acoustic features can be found in [10].

Linguistic Features

Bag-of-words is the most widely used approach in document categorization. It is also commonly used in behavioral signal processing [284]. Bag-of-words and bag-of parts-of-speech (POS) associated words has been extracted separately, then each set was transformed with term-frequency and inverse-document-frequency (TF-IDF).

Psycholinguistic (LIWC) Features

As mentioned in Section 4.1.3, Linguistic Inquiry Word Count (LIWC) [159] system has been developed Pennebaker & King in order to study gender, age, personality, and health and the correlation between these attributes and word uses. There is a total of 81 word categories, a few of which are family, cognitive mechanism, affect, occupation, body, article, and function words. It analyzes language (in our case utterance) on a word-by-word basis. The system has master dictionaries for different languages. LIWC counts the words in the utterance sample that match each of the categories in the dictionary. Scores for each category are expressed as percentages or a proportion of words that match the total number of words used. For example, if an utterance used 10 words that fall into the word category "anger" and the utterance contained 100 words, then the word category's score for "anger" would be 0.10. We used the dictionaries that are available with LIWC for Italian and French with Persia and SPC, respectively.

13.3.2 Feature Selection

To understand the contribution of each feature set, before feature combination phase as shown in Figure 13.3, we tried to reduce dimension for acoustic and token feature vectors. Dimensionality reduction has been used because of the assumption that higher dimension may overfit and may reduce the performance of unseen examples. For the dimensionality reduction, we applied the relief feature selection technique [158], which is the approach that was used in a previous study [291]. No feature selection has been applied to the POS and psycholinguistic feature sets. For the study of POS, only tags have been used that were extracted from tokens. For this study tree-tagger [292] has been used for the Persia corpus and Stanford POS tagger [293] for the SPC corpus.

After evaluating each feature set, we combined the different baseline feature vectors into a single vector, which, as a result, introduced high-dimensional problems. We thus applied the same relief feature selection approach as that we used in [291], in order to avoid high variance and over-fitting. Before the feature selection phase, feature values were discretized into equal frequency bins. All continuous-valued features were transformed into discrete valued features.

13.3.3 Classification and Evaluation

We generated our classification models using SMO [136] with its linear kernel. The main reasons for choosing SMO were (a) its higher generalization capability and (b) the fact that we obtained better results using it, compared to Adaboost and Random Forest algorithms discussed in Section 13.2. The linear kernel was chosen in order to alleviate the problem of higher dimensions. In all of the classification settings, we used SMO's defaults parameters, whereas in a previous study [163] (discussed in section 13.2) we tuned those parameters.

The performance of the system was measured in terms of Weighted Average (WA) and Un-weighted Average (UA), which have recently been used in the paralinguistic tasks [10]. However, for the sake of simplicity, we present only UA.

To evaluate the performance of the SPC development set (dev), we used the SPC training set (train) to generate the model. To evaluate the performance of the SPC test set, we generated a model by combining the SPC training and development sets (training set: train + dev). In each case, performance was estimated using Leave Speaker Group Out (LSGO) cross-validation method on the training set, with macro-averaging. In macro-averaging, UA and WA were calculated for each cross validation fold and their average was computed.

For the Persia corpus, we used Leave One Speaker Out (LOSO) cross-validation with micro-averaging to measure the performance of the system.

Micro-averaged values were calculated by first constructing a global confusion matrix from each cross-validation fold, and then by computing UA_{micro} and WA_{micro}, as shown in equations 13.1 and 13.2. Imbalance class distribution of the Persia corpus was the main reason for choosing micro-average.

$$UA_{micro} = \frac{1}{2}\left\{\frac{\sum_{i=1}^{F} TP_i}{\sum_{i=1}^{F} TP_i + FN_i} + \frac{\sum_{i=1}^{F} TN_i}{\sum_{i=1}^{F} TN_i + FP_i}\right\} \tag{13.1}$$

$$WA_{micro} = \frac{1}{2}\left\{\frac{\sum_{i=1}^{F} TP_i + TN_i}{\sum_{i=1}^{F} TP_i + FP_i + TN_i + FN_i}\right\} \tag{13.2}$$

13.3.4 Results and Discussion

BIG-5 personality traits binary classification models have been evaluated on both corpora. We present the performance of the system for each feature set, their combination, and oracle. The results presented on the acoustic and bag-of-words (token) features are obtained after applying feature selection. We obtained the results on the combined feature set by combining the baseline feature vectors into one vector and then applying the feature selection technique. Oracle performance gives an upper bound on our model performance based on current single feature type models.

Classification results on the SPC dev and test sets are given in Tables 13.8 and 13.9, respectively. The feature combination provides comparable results with the state-of-the-art, even when using SMO's default parameters. The mean UA results overall Big-5 categories show that the acoustic feature set contributes most to the classification decision, whereas the psycholinguistic feature set appears to contribute the second most.

The annotation of the SPC corpus was based on paralinguistic cues (i.e., annotators did not understand the language). However, it seems that lexical-prosodic information coexists here. This means that words, perhaps salient, representing the prosodic information convey some information. Therefore,

Table 13.8: UA results on SPC dev set using different feature sets. Tok.: token, POS: parts-of-speech, Psyc: Psycholinguistic, AC: acoustic, Comb: Feature combination with feature selection, Ora: oracle performance.

Class	Tok	POS	Psyc	AC	Comb	Ora
O	51.6	50.0	56.1	59.1	67.7	92.1
C	55.5	54.7	65.2	74.1	73.7	96.1
E	52.0	53.7	63.4	83.6	84.1	98.4
A	52.3	46.8	54.0	64.0	64.9	97.1
N	51.3	50.0	49.7	63.5	66.3	97.9
Mean	52.5	51.1	57.7	68.8	71.4	96.3

Table 13.9: UA results on SPC test set of different feature sets.

Class	Tok	POS	Psyc	AC	Comb	Ora
O	49.4	49.6	52.8	63.1	62.5	93.6
C	64.6	48.8	69.8	78.6	79.6	94.0
E	56.5	56.0	61.6	77.2	78.2	97.3
A	48.8	51.4	56.2	62.5	65.1	94.2
N	50.4	49.4	50.1	65.6	66.9	91.0
Mean	53.9	51.0	58.1	69.4	70.4	94.0

the feature sets extracted from transcription show quite improved results when combined with acoustic features.

A closer investigation was done after applying feature selection to understand which types of features are important among feature sets in different Big-5 categories. For Big-5 categories, feature selection method selects different ranges of features. However, overall reduction appears to be from 35% to 62% on the SPC train + dev sets, out of $\sim 9.5K$ features. Study of SPC feature sets reveals that for different Big-5 categories, the feature selection method selects and rank different types of features. For example, in the openness category, MFCC-based features appear to have a higher ranking within acoustic features. Within the psycholinguistic feature set, personal pronoun, articles, social and affective categories appear in ranked order. In the POS feature set, it appears that pronouns, verbs, and adverbs have greater significance, and in that order.

The results of our previous study on the SPC dev set are comparable with the feature combination results of the same data set. The performance of the present system was improved by 2.5% and 1.1% in openness and extraversion categories, respectively.

The results of the SPC test are comparable with the baseline results presented in [10], noting only one difference – the baseline results were obtained using tuned parameters whereas our results are obtained using SMO's default parameters. However, our results on the SPC test set outperform the baseline results in all categories except the conscientiousness category. We performed cross-validation on the training set (train + dev) and obtained 68.1 ± 2.7 (mean ± standard deviation) in all traits and it is evident that our results are within statistical variation.

In the study of Persia corpus, we obtained a similar improvement using our feature combination method. Compared to the study [163] discussed in section 13.2, we obtained an improvement of 3.8% on the extraversion and 5.9% on the agreeableness categories. However, performance drops on the conscientiousness and agreeableness categories. An interesting finding here is that in the openness category, using majority voting ensemble method, we obtained 50.2, which is 2.7% better than the feature combination method.

13.3.5 Conclusion

For this comparative study of different types of feature sets, we obtained comparable results when we combined these feature sets into a single vector. Psycholinguistic features, extracted using LIWC, give better results when compared with the token and POS feature sets, whereas acoustic features outperform the other feature sets. However, oracle performance suggests that there is room for improvement in the feature or decision combination

Table 13.10: UA$_{micro}$ results on Persia of different feature sets using LOSO cross validation.

Class	Tok	POS	Psyc	AC	Comb	Ora
O	57.1	44.3	41.5	45.9	47.5	81.6
C	37.5	46.6	37.6	72.6	68.5	81.1
E	37.9	28.6	43.8	52.1	67.3	90.0
A	32.1	48.0	75.1	71.8	74.9	96.9
N	53.7	69.8	64.7	52.1	60.5	96.7
Mean	43.7	47.5	52.6	58.9	63.7	89.2

approach.

13.4 Summary

We presented our contribution to the automatic recognition of speaker personality traits from speech using two different corpora – conversation and broadcast news. We investigated acoustic features using different classification algorithms such as SMO, Random Forest and Adaboost in combination with different feature selection methods such as Information Gain and Relief. We also investigated different decision combination methods such as stacking and majority voting. We found that Relief feature selection with SMO performs better than other feature selection and classification approaches. In regards to the decision combination, majority voting performs better than stacking.

Following our findings on acoustic only feature sets and different feature selection and classification algorithms, later, we investigated other feature sets such as lexical parts-of-speech and psycholinguistic. We obtained comparable results using feature fusion i.e., combining different feature vectors into a single vector. In this comparative study, acoustic features outperform the other feature sets, psycholinguistic features provided better results compared to token and POS feature sets. For an in-depth understanding, we computed oracle results among the feature sets results, which shows new avenue for the research for feature combinations.

Chapter 14

Personality: Multi-Model Corpus - Youtube-Blog

Recently, research in behavioral signal processing has focused on automatically measuring personality traits using different behavioral cues that appear in our daily communication. The computing research aimed at building classifiers generated using a supervised machine learning approach that learns the patterns from social interactions appearing in different forms such as speech, visual-expressions or textual content. A typical approach is to use linguistic or acoustic features, or a combination of both. Linguistic features include lexical features using the bag-of-ngram approach and in some cases using Parts-Of-Speech (POS) or psycholinguistic features [237], whereas acoustic features include statistical functionals applied to low-level descriptors [163]. In most cases, the goal is to find the most relevant features, learning algorithms [163] or the correlation between the lexical features and traits [237].

In this study, we present an approach to automatically recognize personality traits using a video-blog (vlog) corpus, consisting of transcription and extracted audio-visual features. We analyzed linguistic, psycholinguistic and emotional features in addition to the audio-visual features provided with the dataset. We also studied whether we can better predict a trait by identifying other traits. Using our best models we obtained very promising results compared to the official baseline.

14.1 Background

Researchers in psychology have been trying to understand human personality in the last century and it has become one of the sub-fields of psychology. Later, the study of personality became one of the central concerns in different fields such as business, social science, health-care and education [53].

Since the late 19th century, psychologists have been trying to define theories, rating scales, and questionnaires by analyzing lexical terms or biological phenomena [53, 209]. In personality psychology, personality trait is defined as the coherent pattern of affect, behavior, cognition and desire over time and space, which are used to characterize a unique individual.

The advancement of personality traits theories and simplified inventories opened the window for its automatic processing. Hence, in the last few years, automatic personality traits recognition has become one of the mainstream topics in the field of speech and natural language processing to ease the process of interaction between human and virtual agents. This is because it adds value in different areas, such as virtual assistants, healthcare such as mood detection, detection of personality disorder, recommender systems such as customer profiling.

Automatic processing of personality traits from different modalities is a challenging problem and there are many open research issues to solve, such as the types of features, long or short term history of a user, small datasets with imbalanced class labels, combination methods for multimodal information. In this study, we investigate the usefulness of different feature sets using a Youtube dataset released in the Workshop on Computational Personality Recognition (Shared Task) 2014 (WCPR14). The main contributions of this study are the following:

- Studying audio-visual, lexical, POS, psycholinguistic and emotional features and their combinations
- Using predicted traits as features

We used predicted traits as features to predict a trait in a cascaded classification system in order to show that traits can be used as predictors in automatic classification task.

14.2 Corpus: Youtube-Blog

Youtube-Blog corpus has been released in the Workshop on Computational Personality Recognition (Shared Task) 2014 – WCPR14 [13,215]. The shared task consists of two tracks: 1) close task with two competitions - participants are allowed to use multimodal information using one of the datasets and transcriptions from the Youtube dataset and 2) open task - participants can use any external resources. Tasks also include solving both classification and regression problems. Our contributions were comprised of both tracks, however, we focused on only solving the classification problem using the Youtube dataset. The corpus consists of vlogs collected from Youtube, where a single person talks by looking at the camera with their face and shoulders showing and the vloggers talk about a product or an event. Annotation of the vloggers' personality traits has been obtained using Amazon Mechanical Turk. For the shared task, the dataset has been released in the form of extracted audio-visual features, along with the automatic transcription. It contains 348 training, 56 test instances, consisting of 404 vlogs in total, where 194 (48%) are male and 210 (52%) are female vloggers. Train and test splits of the dataset and their distribution are presented in Table 14.1.

Table 14.1: Train and test splits of Youtube-Blog corpus. A number of instances of each and their distribution. O-openness, C-conscientiousness, E-extraversion, A-agreeableness, N-neuroticism. Y and N represent positive and negative classes for each trait with percentage within parenthesis.

Class	Train-set		Test-set	
	Y (%)	N (%)	Y %	N %
O	123 (35.34)	225 (64.66)	18 (32.14)	38 (67.86)
C	155 (44.54)	193 (55.46)	23 (41.07)	33 (58.93)
E	146 (41.95)	202 (58.05)	17 (30.36)	39 (69.64)
A	274 (78.74)	74 (21.26)	42 (75.00)	14 (25.00)
N	157 (45.11)	191 (54.89)	28 (50.00)	28 (50.00)

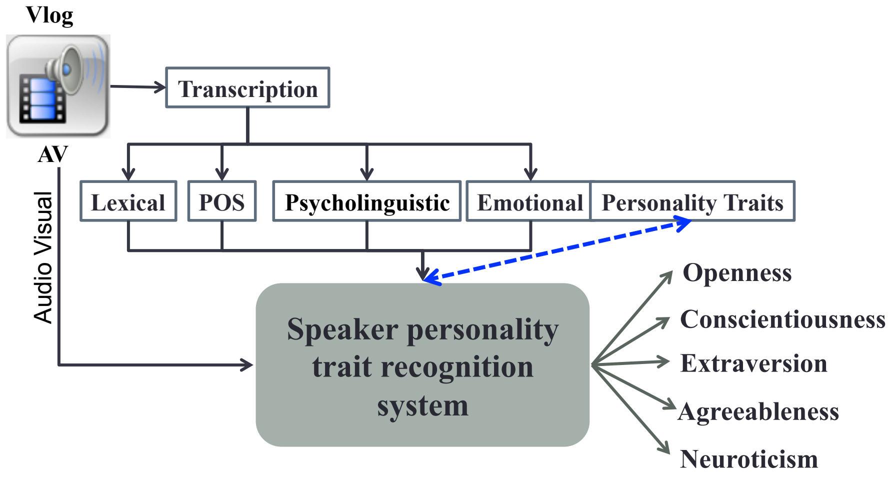


Figure 14.1: The architecture of personality traits classification system using youtube-blog.

14.3 Experimental Design

For the study, we experimented with audio-visual features that had been released with the dataset and also extracted lexical, POS, psycholinguistic and emotional features from the transcription. We also investigated the use of trait labels as features. A conceptual design of the system is given in Figure 14.2 and a details design of the classification system is shown in Figure 14.2. In the following sub-sections, we describe the details of each feature set, feature selection, and classification method.

14.4 Features

Audio-visual features (AV): Different groups of audio-visual features are acoustic, visual and multimodal [215]. The acoustic features include *speech activity* - speaking time, an average length of the speaking segments and a number of speaking turns and *prosodic cues* - voice rate, a number of autocorrelation peaks, spectral entropy, energy, Δ-energy and different variation of pitch. The visual features include *looking activity and pose* - looking time, an average length of the looking segments, a number of looking turns, proximity to the camera and vertical framing and *visual activity* -

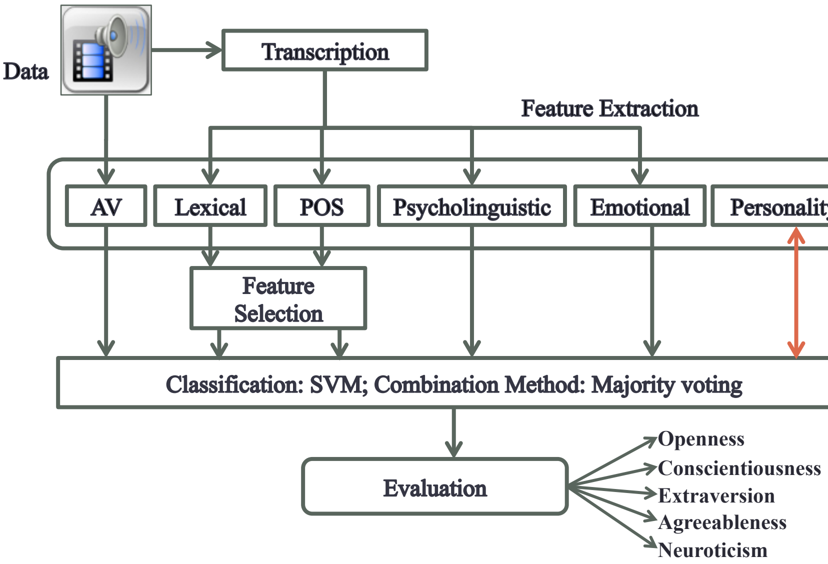


Figure 14.2: Design of the classification system.

statistical descriptors of the body activity. The multimodal features are the combination of speaking and looking ratio.

14.4.1 Lexical features (Lex)

From the transcription, we extracted lexical features (tokens) and then transformed them into a bag-of-words, vector space model. This is a numeric representation of text that has been introduced in text categorization [130] and is widely used in behavioral signal processing [163]. We computed frequencies and then transformed them into logarithmic term frequency (TF) multiplied with inverse document frequency (IDF). To use the contextual benefit of n-grams, we extracted token trigram features, which eventually results in a very large dictionary, however, we reduced them by selecting the top 10K frequent features and filtering out lower frequent features.

14.4.2 POS features (POS)

To extract POS features we used Stanford POS Tagger [294] and used the same approach of lexical features for the transformation and reduction of the POS feature set.

14.4.3 Psycholinguistic features (LIWC)

Pennebaker et al. designed psycholinguistic word categories using most frequent words and developed the Linguistic Inquiry Word Count (LIWC) [159]. It has been used to study gender, age, personality, and health in order to understand the correlation between these attributes and word uses. The word categories include family, cognitive mechanism, affect, occupation, body, article, and function words. We extract 81 features using LIWC and also include gender information with this feature set, is available with the dataset.

14.4.4 Emotional features (Emo)

We considered emotional categories and sentiment predictions as emotional features extracted from different resources. These resources include NRC lexicon [257], WordNet-Affect [295], SentiWordNet [296] and Stanford-sentiment tool [297]. To extract information for emotional categories, we used NRC lexicon and WordNet-Affect where the list of words are annotated with emotional categories. We calculated the frequency of an emotional category by matching the words belonging to this category with the words in the instance of the transcription. The NRC categories include *anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise* and *trust* whereas the WordNet-Affect categories include *anger, disgust, fear, joy, sadness* and *surprise.* There are overlaps between categories of these two lexicons. However, we have not combined them as the designing processes of these two lexicons are different. We computed sentiment scores using the SentiWordNet, which computes scores based on the positive and negative sentiment scores defined in the lexicon and sentiment decision using the Stanford-sentiment tool. Apart from that, we also use two additional *neutral* categories. One *neutral* category is composed of the list of words from NRC that do not belong to any of the NRC emotional categories and the other *neutral* category includes the words of an instance that do not belong to any emotional category. Therefore, we have 20 features - 10 NRC, 6 WordNet-Affect, 1 SentiWordNet, 1 Stanford-sentiment and 2 neutral.

14.4.5 Traits as features (Traits)

To design a model for a trait we used other four traits' labels as features and to obtain the traits labels for the test set we designed a two-level cascaded classification system. In the cascaded system, the first level model is selected from the models we generated using different feature sets and by using that we generated the features (traits labels) for the test set. Then, we designed the second level model.

14.5 Feature selection

We extracted high-dimensional features for lexical and POS sets, which is one of the reasons of overfitting. Therefore, to avoid high variance and overfitting and to improve the performance, we performed feature selection using Relief (see [163] and the reference therein) algorithm with 10-fold cross-validation on the training set, following the same approach used in [163]. Before the feature selection, feature values were discretized into 10 equal frequency bins.

14.6 Classification and Evaluation

We generated our classification models using Sequential Minimal Optimization (SMO) for Support Vector Machine (SVM) [136] for each feature set as described above. SMO is a variant of SVM, which solves the Quadratic Optimization (QP) problems analytically and avoids time-consuming numerical QP optimizations. We used different kernels for different feature sets, such as linear kernel for lexical (Lex) and POS features and polynomial kernel for audio-visual (AV), psycholinguistic (LIWC), emotional (Emo) and traits (Traits) features. The Linear kernel was chosen in order to alleviate the problem of higher dimensions for lexical and POS feature sets. Sometimes, however, it also gives optimal results for a small set of features. We have tuned the parameters to obtain a better performance on each feature set using 10-folds cross-validation on the training set. Feature selection has been applied for lexical and POS feature set as mentioned earlier (see Section 14.5). The performance of each classification model has been measured in terms of average precision 12.1, recall 12.2 and F1 12.3, which are the evaluation metrics specified for the shared task. However, for the reasons of brevity, we only present F1 scores.

For the combination of different models of the feature sets we used decision fusion as shown in equation 4.3 and combined the decisions from the models of five feature sets. As a combiner, we applied majority voting. We first designed a model by combining the decisions from the models generated using five feature sets, named it as **Maj-5** model - majority voting of the five models of five feature sets. After that, we designed another model discarding the model of emotional features from the combination, named this model as **Maj-4** - majority voting of the four best models out of the models of five feature sets.

To understand the usefulness of the traits as features, we designed a two-level cascaded classification system. In the cascaded system, we generated the traits labels for the test set using the best combined model (**Maj-5**) as the average performance of this model is best among the models. We designed the second level model using the predicted traits as features (see Section 14.4.5) and used SMO with its default parameters, named this model as **Maj-5-Traits**. To obtain the baseline of this feature set, we trained models using the traits labels of the training set and then evaluated them using the traits labels on the test set as shown the results in Table 14.3, named it as **Ref**.

Table 14.2: Results on test set using different feature sets. Baseline: Official baseline, AV: Audio-Visual, Lex: Lexical, POS: Parts-Of-Speech, LIWC: psycholinguistic, Emo: Emotion, Maj-5: Majority voting of the five models, Maj-4: Majority voting of the four best models, Maj-5-Traits: Generated traits labels using Maj-5 model

Model	**O**	**C**	**E**	**A**	**N**	**Avg**
Baseline	40.4	42.9	41.1	33.3	37.1	39.0
AV	63.4	42.9	70.4	67.7	55.7	60.0
Lex	59.9	49.4	60.4	65.8	56.7	58.4
POS	57.3	54.3	57.8	69.6	**61.9**	60.2
LIWC	55.0	56.0	66.2	71.4	46.8	59.1
Emo	49.3	52.5	53.5	59.4	40.1	51.0
Maj-5	**65.0**	57.4	69.4	**76.7**	59.4	65.6
Maj-4	61.5	**61.9**	68.8	74.7	57.1	64.8
Maj-5-Traits	59.2	41.7	**71.0**	62.2	52.6	57.3
Best model	**65.0**	**61.9**	**71.0**	**76.7**	**61.9**	**67.3**

14.7 Results and Discussion

We present the performance of the classification models designed using different feature sets, decision combination, traits features and their best F1 on each trait in Table 14.2, in addition to the official baseline. In the close shared task, using audio-visual features, we obtained an average of $F1$: 1.6% better than using lexical features and an average of F1: 21% better than official baseline. We obtained comparative results among the AV, Lex, POS, and LIWC feature sets. The emotional feature set (Emo) does not perform well individually. An extension of this research work could be examining the representation of these features in the vector form, either as frequency or relative frequency or any other transformation.

The decision combination provides better results compared to the results of any single feature set. We obtained an average of F1: 65.6% using the model Maj-5 and F1: 64.8% using the model Maj-4, which implies that emotional features also contribute to improve the performance in combination. The performance of traits features is lower compared to the Maj-5 model, however, we obtain better results on *extraversion* category.

Table 14.3: Results on test set using traits as features. Ref: Reference labels of the test set. Maj-5-Traits: Generated traits labels using Maj-5 model

Model	**O**	**C**	**E**	**A**	**N**	**Avg**
Ref	77.1	41.7	77.1	58.6	58.6	62.6
Maj-5-Traits	59.2	41.7	71.0	**62.2**	52.6	57.3

In Table 14.2, we show the performance of the traits feature set using the reference labels and Maj-5-Traits. The results of the Maj-5-Traits model are better in *agreeableness* category compared to the model using reference labels. We will investigate the traits features further on different datasets to understand their significance. Our observation is that performance of each trait varies for different feature sets, which implies that the same feature set or architecture might not work for all traits. We might have to use the model, which performs best for a particular trait. The best models are marked in bold-form in Table 14.2 for the traits and the last row of Table 14.2 shows the best results where we obtained an average of F1: 67.3%.

Significance test: We conducted statistical significance test of our best models with the second best models using the binomial test. The test revealed that the results of the best models are statistically significant with $p < 0.05$ for extraversion and with $p < 0.01$ for other categories.

14.8 Summary

In this chapter, we presented our contribution to the automatic recognition of personality traits from a video-blog corpus by studying different types of feature sets. The feature sets include audio-visual, lexical, POS, LIWC, emotional features and their combinations using majority voting. In addition, we also used predicted traits as features and designed a cascaded classification system. We obtained very promising results compared to the official baseline. The performance of the model using emotional feature set is very low compared to the other feature sets, however, it helps in combination.

Chapter 15

Personality, Mood and Communication Style

Participating in social media has become a mainstream part of our daily lives – we read articles, comments, other people's statuses and provide feedback in terms of emotions, likes, dislikes, and other social signals through written content. Since currently our social participation is mostly done through social media platforms, the online content, including social media and newspapers' content, is growing very rapidly. It creates unprecedented opportunities for businesses and individuals, as well as it poses new challenges to process and generate concrete summaries out of it. The challenges include automatic processing of semi-structured or unstructured data in different dimensions such as linguistic style, interaction, sentiment, mood, personality and other social signals. Finding the collective information of such signals requires automatic processing, which will be useful for various professionals, specifically psychologists and social and behavioral scientists. Among other affective dimensions, mood, personality and communication style has also been studied for the analysis of the consumer behavior towards brands and products [298–300].

In this chapter, we address the question of how personality types and communication styles of Twitter users are related to the selection of contents they share in Twitter, affecting the diffusion of a positive or negative mood. Our goal was to use publicly available tools and self-reported annotations to design and evaluate the computation models in order to find their association.

15.1 Background

In the online news and social media, people read and share links to news articles or other multimedia contents, that are related to their emotions, tastes and identity [301]. The exposure to contents generated by others can give rise to different emotions like indignation, joy, anger or sadness [302]. Sometimes these contents may be shared or retweeted, indicating the users' will to participate in a diffuse conversation [303] and share their emotions with others. Researchers [304], [305] have discovered that such media consumption and sharing is affected by the personality type of the user. Different personality types are associated to different psychological dimensions [306], such as linguistic functions, attentional focus, emotionality and social relationships.

In this study, we aim at finding the relationships between personality, communicative style and mood sharing; the best predictors of mood and the performance in the classification of positive and negative mood sharers among Twitter users. We formalize the problem into three tasks:

1. correlation analysis
2. feature selection
3. classification

We identify the data sources in Corriere[1], an Italian news platform that provides mood metadata annotated by the readers on a voluntary basis, and Twitter[2], that is widely used as an information diffusion platform. We annotate the data with personality and communication style labels, then we predict the average mood of the articles shared on Twitter by the users. The main contributions of this work to the research community are: 1) the development of an aligned corpus of Tweets and news articles, automatically annotated with personality types, communication styles and gold standard mood labels; 2) the analysis of the influence of Twitter users' metadata, personality and communication style in the diffusion of mood; 3) the prediction

[1] http://corriere.it

[2] http://twitter.com

of mood of a news article from personal data.

We exploit *mood* metadata annotated directly by news readers in *Corriere.it* on a voluntary basis, to analyze the role of the users in spreading moods in a social network like Twitter. In corriere, there are 5 context-independent mood states: **amused, satisfied, disappointed, worried and indignated**. Each one of them can have a strength value between 0 and 100.

To define *personality types*, we adopt the most popular personality model in psychology: the Big Five [307], that defines 5 bipolar traits: **extraversion** (sociable *vs* shy); **emotional stability/neuroticism** (secure *vs* neurotic); **agreeableness** (friendly *vs* ugly); **conscientiousness** (organized *vs* careless) and **openness to experience** (insightful *vs* unimaginative).

To define *communication styles* we adopt the classes provided by Analyzewords, a tool for tweet analysis based on Linguistic Inquiry and Word Count (LIWC) [308]. Analyzewords defines 11 communicative dimensions, namely: **upbeat** (positive words and large use of "we"), **worried** (use of anxious language and short questions), **angry** (large use of captions and hostile words), **depressed** (use of self-reference and negative words), **plugged-in** (use energy words and include many mentions in tweets), **personable** (use positive words and often refers to others), **distant** (use action words and do not refer to self much), **spacy** (use excited words and a lot of exclamation marks), **analytic** (use long words and complex conjunctions) **sensory** (use many feeling words and reference to self), **in the moment** (use mainly verbs at present and hashtags).

15.2 Related Work

It is well known that mood has an impact on social media and spreads through social networks. Bollen *et al.* [309] predicted mood states (tension, depression, anger, vigor, fatigue, and confusion) from tweets and compared the results to a record of popular events gathered from media, finding a significant correlation between them. Other works focussed on information spread, virality and retweeting of messages. This kind of research reached contradictory conclusions: while some researchers concluded that the most important features to predict retweeting is the level of influence of the source of the tweet and the retweeter [310], others discovered that message virality is connected to the content of the message being shared, rather than to the influencers who share it [311] [312].

Recent works that put together emotions and information spread, found that emotionally charged tweets tend to be retweeted more often and more quickly compared to neutral ones [313]. Viral messages containing the six primary emotions (surprise, joy, sadness, anger, fear, and disgust) are very effective on recipients' emotional responses to viral marketing campaigns. However, emotional content can evoke different reactions based also on the gender of the audience. Dobele *et al.* [314] discovered that male recipients were more likely to forward disgust-based and fear-based campaigns that their female counterparts. The effectiveness of mood as a feature has been proven for tasks like author profiling [315] and cyberpedophilia [316]. Hill *et al.* provided formal evidence that positive and negative emotional states behave like infectious diseases spreading across social networks over long periods of time [317]. As for the relationship between sentiment and personality, previous literature [318] reports a little improvement in the classification of sentiment exploiting personality types. Other related works include sentiment analysis [319], mood annotation [320], or mood assessment [321].

15.3 Methodology

In Figure 15.1, we present the functional architecture of our workflow. We collected tweets who shared corriere articles and user's metadata from

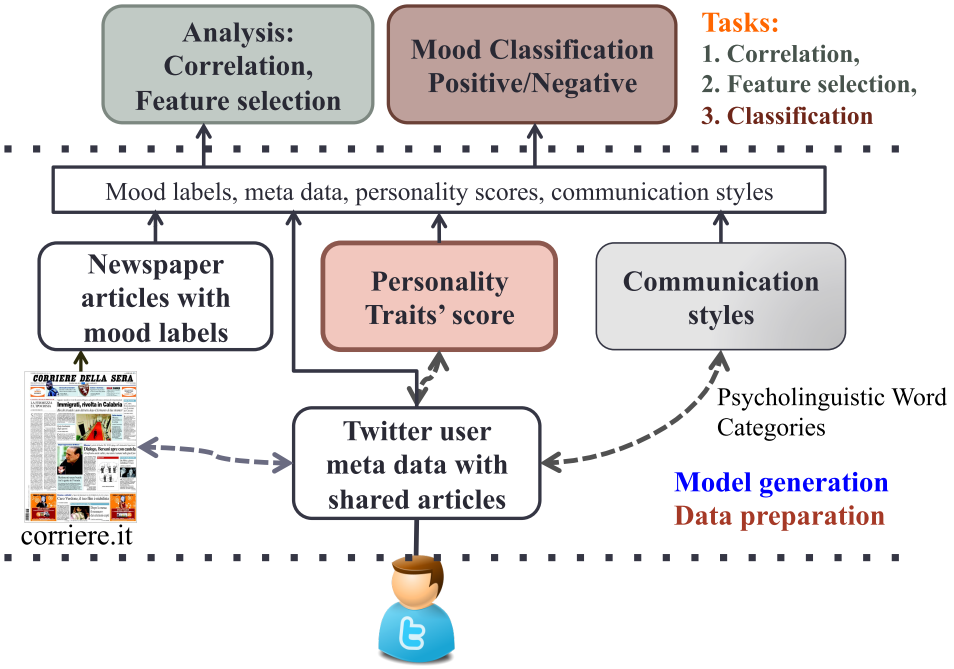


Figure 15.1: Conceptual design of the workflow.

the twitter. For the same user who shared corriere articles, we also collected metadata containing mood information from corriere. We automatically labeled the tweets with personality traits and communication styles, and used mood information for the analysis (i.e., correlation and feature selection) and classification tasks. In the following subsections, we discuss the details of each component of the system.

15.3.1 Data Collection and Annotation

Twitter is a very popular micro-blogging web service, which allows users to post short text messages, called "tweets", up to 140 characters. Common practices in Twitter are the "mentions", to converse with other users, "retweets" - to share information [322], and "hashtags" - to aggregate messages by topic. In recent years a lot of works have focussed on data mining from Twitter. For example, for sentiment analysis from emoticons [323], irony detection [324], ranking algorithm for extracting topic keyphrases from tweets [325] and of course personality recognition [326] [327], [328]. Corriere

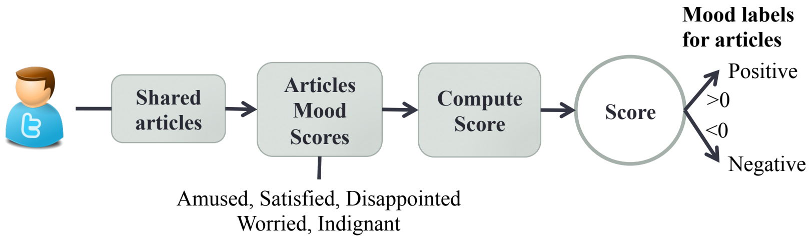


Figure 15.2: Reference labels generation for the mood classification task.

is one of the most popular Italian daily newspapers, and its online platform is structured as a social network, according to the definition in Boyd & Ellison [329]. In particular, the website of corriere provides 1) a semi-public profile for each registered user, 2) articulates a list of users connected by a relationship of interest and 3) allows to view their list of connections to other registered users.

We sampled about 2500 users from Twitter who shared at least two articles from *corriere.it*. We limited the number of tweets sampled from the APIs to 3000 per user. We computed the ratio between the number of articles shared and the number of tweets posted, cutting the tail in the fourth quartile (tweet-shared articles ratio above 0.32), in order to remove the accounts of Corriere.it, journalists of Corriere and bots that retweet corriere articles.

To compute average mood class, first, we subtracted the sum of "disappointed", "worried" and "indignated" scores from the sum of "amused" and "satisfied", obtaining a unique polarity score, as presented in Figure 15.2. We turned this polarity score into two classes: above and below zero, removing 21 instances with a score equal to 0. After this process, we have 2042 unique users. A summary of the distribution of all features is reported in Figure 15.3. Hashtag score, mention score and articles shared score are computed as the ratios of hashtags ($\frac{hashtags}{tweets}$), mentions ($\frac{all@-self@}{tweets}$) and Corriere articles ($\frac{articles}{tweets}$) over the number of Tweets sampled. All the other features are real values: count or scores from personality traits and communication style prediction.

15.3.2 Dataset for the Evaluation of Personality

In order to evaluate the annotation of personality types, we recruited 210 Twitter users with an advertising campaign targeted at the followers of Corriere in Twitter, we assessed their personality types by means of the short BFI-10 personality test [330] online[3]. In this way we obtained gold standard personality labels for the training and evaluation. We used the short test (it takes less than 5 minutes to be completed) and we recruited only volunteers in order to have the full attention of the users [331]. In the sample we have 118 males and 92 females aged between 14 and 65 years. A summary of the distribution of gold standard personality types is reported in Table 15.1.

Trait	**Min**	**Mean**	**Max**
Open	-0.2	0.21	0.5
Cons	-0.2	0.18	0.5
Extr	-0.3	0.18	0.5
Agre	-0.3	0.14	0.5
Neuro	-0.3	0.12	0.5

Table 15.1: Summary of gold standard personality types distribution. Open: Openness, Cons: Conscientiousness, Extr: Extraversion, Agre: Agreeableness, Neuro: Neuroticism.

15.3.3 Automatic Label for Personality Traits

In order to perform the automatic annotation of personality types, we trained a supervised model on the gold standard labeled dataset we collected from Twitter. We split the data into training (180 Twitter users) and test set (30 users). Then, trained the model using bag-of-n-grams as features and Random Forest as la earning algorithm. We obtained an average Root mean Squared Error of 0.18, as reported in detail in Table 15.2.

This result is comparable to the study of Golbeck [328], who obtained an average Mean Absolute Error of 0.15.

[3] http://personality.altervista.org/personalitwit.php

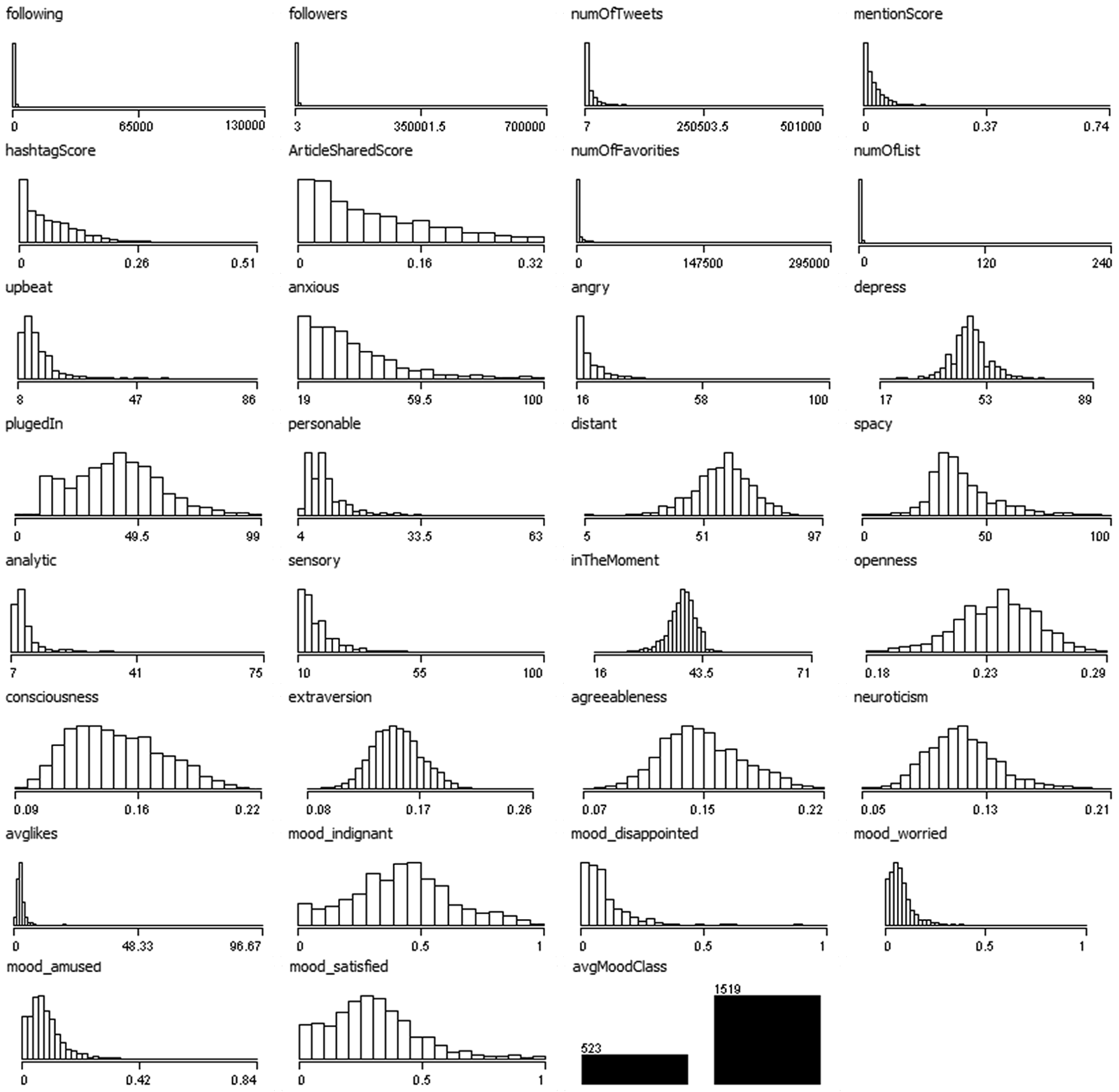


Figure 15.3: Distribution of features in the dataset for experiments.

Class	Baseline	RMSE
Open	0.19	0.18
Cons	0.16	0.15
Extr	0.22	0.17
Agre	0.17	0.17
Neuro	0.24	0.24
Avg	**0.19**	**0.18**

Table 15.2: Results of personality score evaluation. Open: Openness, Cons: Conscientiousness, Extr: Extraversion, Agre: Agreeableness, Neuro: Neuroticism.

15.3.4 Automatic Label for Communication Styles

We also labeled the dataset with communication styles, defined in section 15.2, by exploiting a freely available tool – Analyzewords[4]. This tool provides a representation of Tweets based on the psycholinguistic dimensions in LIWC, which is based on expert knowledge.

15.4 Results

15.4.1 Correlation Analysis

First of all we computed correlations between all the dimensions we retrieved, and we report the heatmap in Figure 15.4.

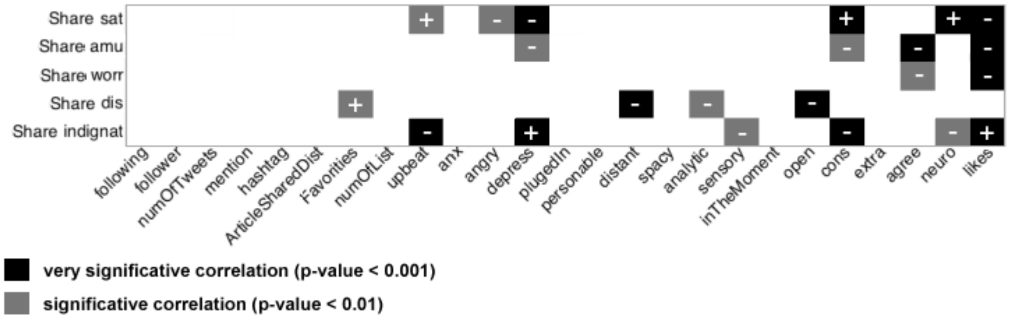


Figure 15.4: Heatmap of the correlations between all the dimensions we retrieved (Twitter metadata, corriere metadata) and generated (personality types, communication styles).

Many interesting relationships emerge from this experiment: first of all,

[4] http://www.analyzewords.com/

the correlations between Twitter metadata and the action of sharing a specific mood are very few and weak. The only significative correlation is between the number of favorite Tweets and the tendency to share articles that arouse disappointment. This can be explained that users tend to read and collect news and tweets that attract their attention arousing disappointment.

Among communication styles, it is very interesting to note that the upbeat style is in a strong negative correlation to sharing articles that arouse indignation, and in a positive correlation with the action of sharing satisfaction. On the contrary, a depressed communicative style is strongly correlated to sharing indignation and negatively correlated to sharing satisfaction. Surprisingly, a distant communicative style is negatively correlated to sharing disappointing articles. We find the same negative correlation, although weaker, also for the users with an analytic communication style. Moreover, an angry communicative style is not correlated to sharing indignation, but it is just negatively correlated to sharing satisfaction.

Among personality types, openness to experience is negatively correlated to sharing disappointment, just like the distant communicative style. An explanation for this is that open-minded users like to understand things and do not like to share articles arousing disappointment. Conscientiousness is positively correlated to sharing satisfaction and negatively correlated to sharing indignation, and also negatively correlated to sharing amusement, although with less strength. A surprise is that also agreeableness is negatively correlated to sharing articles arousing amusement, but it is also negatively correlated to sharing articles that arouse worry or concern. Unsurprisingly, emotional stability/neuroticism is strongly correlated to sharing satifaction and negatively correlated to sharing indignation. Surprisingly, extraversion is not correlated to any mood sharing action, although strongly correlated to an upbeat communication style.

Crucially, the number of likes on the articles is strongly correlated to arti-

Feature	Infomation Gain
Avglikes	0.0886
NumOfTweets	0.0706
ArticleSharedScore	0.0689
Depress	0.0681
Consciousness	0.0653
Angry	0.0604
PlugedIn	0.0544
Upbeat	0.0528
NumOfFavorities	0.0499
HashtagScore	0.0477

Table 15.3: Results of feature selection.

cles that arouse indignation, while it is negatively correlated to articles arousing worry, amusement and satisfaction. It is not easy to explain why the "like" action is strongly associated with a negative emotion. We suggest this may be connected to the fact that indignation is a social emotion [332] triggered by people's tendency to view others' behavior in relation to self-behaviour. Under this perspective, the "like" action is an expression of support [333] to indignant people.

15.4.2 Informative Features

In the feature selection experiment, we want to find the best predictors of the average mood shared on Twitter. We ran feature selection with information gain ranking as algorithm and 10-fold cross validation as the evaluation method. This algorithm evaluates the worth of the features by measuring the information gain of each attribute with respect to the class:
$InfoGain(Class, Attribute) = H(Class) - H(Class|Attribute)$
where H is the entropy. The results reported in Table 15.3, show that the best features are the average article like score, which is not really surprising because it depends directly on the article content. Crucially, the best communication style predictor is depression and the best personality predictor is conscientiousness, in line with the findings in previous work [318].

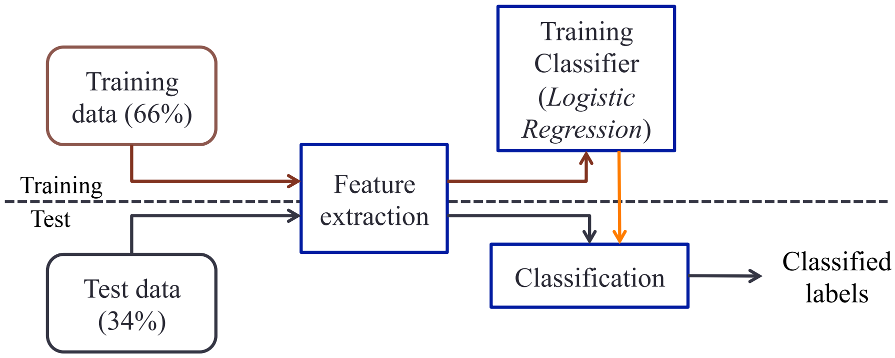


Figure 15.5: Mood classification system.

Class	P	R	F1
Baseline	0.50	0.50	0.50
Positive	0.62	0.66	0.64
Negative	0.63	0.57	0.69
Avg.	0.62	0.62	0.62

Table 15.4: Results of classification of positive and negative mood sharers in Twitter.

15.4.3 Mood Classification

In Figure 15.5, we present the mood classification system. We performed the classification task to predict the class labels for mood and recognize automatically the positive and negative mood sharers on Twitter. As a classification algorithm, we used a Logistic Regression, with 66% training and 33% test split. We balanced the two classes with a weighting scheme, in order to preserve the number of instances, and used all the features. The results, reported in Table 15.4, show that it is possible to predict correctly about 60% of positive and negative mood sharers in Twitter using personality types and communication styles. In particular, positive mood sharers can be detected with more recall and negative mood sharers with more precision.

15.5 Summary

We explored the correlations between personality, communication style and Twitter metadata and we successfully predicted the users who shared articles arousing positive and negative moods. We found some correlations, such as the one between sharing satisfaction and an upbeat communicative style. We also found surprisingly significant correlations, like the fact that open minded people tend not to share disappointment. We conclude that these findings can be very interesting for the works about virality and Social Network Analysis: some personality types and some communicative styles correlate with what is being shared, and this is something to keep into account when modeling the diffusion of news or emotions trough social networks.

Chapter 16

Summary: Personality

In the second part of the work, we discussed our contributions on designing computational models for personality traits. We presented the review of current state-of-art, which follows our contributions to the personality traits computing research. We presented how social media conversations, such as Facebook statuses, can be employed to automatically predict user's personality traits. At the same time, we discussed the challenges to deal with social media conversations. Following that we explored human-human dyadic spoken conversations and broadcast news where we presented how vocal non-verbal and verbal cues can be exploited for the automatic prediction task. Also, we presented how both verbal i.e., in terms of lexical and psycholinguistic features, and vocal non-verbal, i.e., in terms of acoustic features can be combined to make a final prediction. While doing so we also explored different machine learning algorithms to understand the algorithm's prediction power for a certain task. We then studied users' audio-visual expressions where users' discuss a product or present an experience in youtube blog. We explored how audio-visual information can be exploited in order to automatically predict users' personality traits. After that, we investigated the correlation between users' personality traits, mood and communication style in which we exploited different sources of information such as tweets, and newspaper articles with users' self-reported labels. We believe our findings will broaden the scope of the current state-of-the-art in personality traits computing research.

Chapter 17

Conclusion

17.1 Contributions

The motivation of this thesis was to design computational models of affective behavior and personality traits, which can facilitate domain experts, help in designing intelligent interface for human-machine communication. Our contributions towards such a goal are divided into two parts.

In the *first part of the thesis*, we focused on speech as the only input modality for designing automatic models of affective behavior. We exploited real call center conversations for designing models for detecting agent's *empathy* and customer's *agner, frustration, satisfaction* and *dissatisfaction.* For annotating the manifestations of emotional states, we developed annotation guidelines for the annotators, who are expert psycholinguistics. Our annotation guidelines are based on the modal model of emotion by Gross [4], which is based on the appraisal theory. By following the annotation guidelines the annotators annotated 1894 conversations.

For our study of designing computational models, we have conducted both conversation and segment level classification experiments. For the conversation level, we designed model for detecting the presence of absence of an emotional state in a conversation with a binary classification setting. The task was detecting empathy on the agent channel and other four emotional states on the customer channel. For a more in-depth understanding, we then focused on the segment level classification task. The goal was to automatically segment the conversation into speech and non-speech segments then classify the speech segment with emotional states. Our goal was to design the computational models, which does not require any human intervention.

For designing such systems, we investigated speaker's verbal and vocal non-verbal cues in term of lexical, psycholinguistic and acoustic features. The input to our system is an audio signal, however, in order to obtain the verbal content i.e., transcriptions from audio, we employed ASR system.

Our experimental findings suggest that lexical and acoustic features provide comparative results. Using the acoustic features the segment level empathy detection model provide an average recall of 68.1% and the model for customer's emotion provides 47.4%. The scenario is different with lexical features, we obtained 65.6% with empathy detection model and 56.5% with the customer's emotion model. The results of empathy detection model, i.e., agent's emotion and customer's emotion model are not exactly comparable due to the different classification settings in two systems. For the agent's model, the setting is binary classification whereas the for the customer's emotion model the setting is multiclass classification. Due to the complexity of the multiclass classification setting, the performance is lower compared to the binary classification. For both systems, we obtained better results with decision combination using majority voting i.e, 70.1% for agent's model and 56.9% for the customer's model. It is needed to mention that the use of acoustic features is an ideal scenario when no transcription available.

One of the important problems that we needed to deal with is the skewed class distribution. We proposed a two steps sampling approach to reduce the class-imbalance problem. The *first step* is to downsample the instances of majority class e.g, neutral. While downsampling the instances we used different bins by considering the segment length and randomly selected equal number (e.g., $N = 100$) of segments from each bin. The number of the bin was predefined, for example, bin_1 contains segments with length ≥ 0 and ≤ 3 seconds. The predefined number of the bin has been empirically found optimal on the development set. The purpose was to have an equal number of variable length segment in the training set, which can capture the patterns on the test set and at the same time reduce the class-imbalance problem. In the *second step*, we upsampled the instances of the minority class(es), e.g., empathy, using SMOTE algorithm. This algorithm generates synthetic instances based on its nearest neighbors. It takes the difference between the feature vector under consideration (i.e., instance using which we will generate a synthetic instance) and its nearest neighbor instance. Then, it multiplies this difference by a random number between 0 and 1, and add it to the feature vector under consideration.

For designing the segment level classification model, we also explored the HMM based generative model for segmenting and classifying segments. We investigated two corpora such as SISL behavioral corpus and FAU-Aibo robot corpus. This study is in a very early stage, however, it will open a new avenue for future research.

Following our work on segment level classification systems, we focused on finding the emotional sequence for the whole dyadic spoken conversation. We defined the term *affective scene* for representing the emotional sequence. From the emotional sequence, one can analyze different patterns to get insights of the agent and customers' affective behavior. Such analysis can be useful to the domain experts such as call center managers. Even though we investigated call center conversations, however, the presented systems can also be useful in other area of research such as health-care, and teacher-student tutoring systems.

In any machine classification task, generalizability is one of the important problems. We wanted to understand how our designed models can perform in other domain or corpora. For doing this study, we utilized SISL behavioral corpus and FAU-Aibo robot corpus and conducted binary and multi-class classification experiments. Our findings show that there is a drop in performance while evaluating system across corpora. This study also opens a new avenue of research in future by employing domain adaptation or transfer learning strategy.

In the *second part of the thesis* we discussed our contributions of the personality traits. Our motivation was to advance the current state-of-the-art of personality computing research. We investigated social media conversations such as Facebook statuses, human-human spoken conversations, broadcast news, and video-blog. While studying social media conversations we explored different feature representation strategies using bag-of-ngram approach. It includes boolean representation, word-frequency, and TF-IDF representation. At the same time, we explored different machine learning algorithms. We obtained better performance, an average across recall of 58.71%, with tf-idf representation with Multinomial Naïve Bayes model.

In studying broadcast news and human-human spoken conversations, we explored the verbal and vocal non-verbal content in term of lexical and acoustic features. While doing so we also studied different feature selection and classification algorithms. The feature selection algorithm includes Information gain and Relief, and machine learning algorithms include BLR, MNB, SVM, Random Forest, and Adaboost. We obtained higher performance with a combination of Relief and SVM. In addition, we also explored parts-of-speech and psycholinguistic features extracted from the transcriptions. We obtained the best performance when we combined the results with decision level combination using majority voting.

Our study on personality traits using youtube video shows that video blogger expresses their personality with the visual expression too. We independently evaluated lexical, syntactic (part-of-speech), psycholinguistic, audio-visual features and trait labels as features. Again, with decision level combination we obtained better results.

Finally, we explored how personality traits of the twitter user are associated with user's mood and communication style. We studied it in terms of correlation analysis, informative features, and mood classification using personality traits and communication style features.

Our contributions of personality traits will facilitate to design the user's persona and finding similar minded people, which might also have many business potentials.

17.2 Future Directions

In regards to the study of affective behavior, one important study can be explored in future is to investigate an adaptation or transfer learning approach to exploit unlabeled data. It is also necessary to understand how the computational model works across corpora in order to understand the generability of the model.

Currently, the segment level system is comprised of speech *vs* non-speech segmenter followed by the segment classification models. To design a single model to deal with both segmentation and labeling tasks, a significant amount of research needed to be done in future.

The study of the affective scene in terms of the emotional sequence is a new avenue of research. It can be explored not only in dyadic conversations but also multi-party conversations. This thesis only utilized speech modality, and other modalities can also be combined.

For the study of personality traits, more studies need to be done in order to use them in real applications in terms of summarizing users' long terms traits. How these long-term traits can reflect the short term states. This thesis only explored such associations with social media conversations. Speech combined with other modalities can also be explored.

Bibliography

[1] C. Chastagnol and L. Devillers, “Personality traits detection using a parallelized modified sffs algorithm,” *computing*, vol. 15, p. 16, 2012.

[2] K. R. Scherer, “Psychological models of emotion,” *The neuropsychology of emotion*, vol. 137, no. 3, pp. 137–162, 2000.

[3] K. R. Scherer, T. Johnstone, and G. Klasmeyer, “Vocal expression of emotion,” *Handbook of affective sciences*, pp. 433–456, 2003.

[4] J. J. Gross, “The emerging field of emotion regulation: An integrative review,” *Review of General Psychology*, vol. 2, no. 3, p. 271, 1998.

[5] R. W. Picard, *Affective computing.* MIT press, 2000.

[6] S. Narayanan and P. G. Georgiou, “Behavioral signal processing: Deriving human behavioral informatics from speech and language,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[7] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing.* John Wiley & Sons, 2013.

[8] F. Weninger, M. Wöllmer, and B. Schuller, “Emotion recognition in naturalistic speech and language—a survey,” *Emotion Recognition: A Pattern Analysis Approach*, pp. 237–267, 2015.

[9] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan, “Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions.” in *Proc. of Interspeech*, 2009, pp. 1983–1986.

[10] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge." in *Proc. of INTERSPEECH*, 2012.

[11] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, "Recognition of personality traits from human spoken conversations." in *INTERSPEECH*, 2011, pp. 1549–1552.

[12] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition: Shared task," in *AAAI*, 2013.

[13] F. Celli, B. Lepri, J.-I. Biel, D. Gatica-Perez, G. Riccardi, and F. Pianesi, "The workshop on computational personality recognition 2014," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 1245–1246.

[14] M. A. Schmuckler, "What is ecological validity? a dimensional analysis," *Infancy*, vol. 2, no. 4, pp. 419–436, 2001.

[15] K. R. Scherer, "Appraisal considered as a process of multilevel sequential checking," *Appraisal processes in emotion: Theory, methods, research*, pp. 92–120, 2001.

[16] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation," in *Affect and emotion in human-computer interaction*. Springer, 2008, pp. 75–91.

[17] A. Batliner, F. Burkhardt, M. Van Ballegooy, and E. Nöth, "A taxonomy of applications that utilize emotional awareness," *Proc. of IS-LTC*, pp. 246–250, 2006.

[18] G. Riccardi and D. Hakkani-Tür, "Grounding emotions in human-machine conversational systems," *Lecture Notes in Computer Science, Springer-Verlag*, pp. 144–154, 2005.

[19] K. Forbes-Riley and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," *Speech Communication*, vol. 53, no. 9, pp. 1115–1136, 2011.

[20] S. Yildirim, C. M. Lee, S. Lee, A. Potamianos, and S. Narayanan, "Detecting politeness and frustration state of a child in a conversational computer game." in *INTERSPEECH*, 2005, pp. 2209–2212.

[21] T. Zhang, M. Hasegawa-Johnson, and S. E. Levinson, "Cognitive state classification in a spoken tutorial dialogue system," *Speech communication*, vol. 48, no. 6, pp. 616–632, 2006.

[22] S. Yildirim, S. Narayanan, and A. Potamianos, "Detecting emotional state of a child in a conversational computer game," *Computer Speech &amp; Language*, vol. 25, no. 1, pp. 29–44, 2011.

[23] H. Ai, D. J. Litman, K. Forbes-Riley, M. Rotaru, J. R. Tetreault, and A. Pur, "Using system and user performance features to improve emotion detection in spoken tutoring dialogs," 2006.

[24] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[25] G. Orrù, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli, "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review," *Neuroscience &amp; Biobehavioral Reviews*, vol. 36, no. 4, pp. 1140–1152, 2012.

[26] C. Jones and M. Jonsson, "Using paralinguistic cues in speech to recognise emotions in older car drivers," in *Affect and Emotion in Human-Computer Interaction.* Springer, 2008, pp. 229–240.

[27] C. Jones and A. Deeming, "Affective human-robotic interaction," in *Affect and emotion in human-computer interaction.* Springer, 2008, pp. 175–185.

[28] S. Steidl, *Automatic classification of emotion-related user states in spontaneous children's speech.* University of Erlangen-Nuremberg, 2009.

[29] L.-Y. Lin, "The relationship of consumer personality trait, brand personality and brand loyalty: an empirical study of toys and video games buyers," *Journal of Product & Brand Management*, vol. 19, no. 1, pp. 4–17, 2010.

[30] M. Bosnjak, M. Galesic, and T. Tuten, "Personality determinants of online shopping: Explaining online purchase intentions using a hierarchical approach," *Journal of Business Research*, vol. 60, no. 6, pp. 597–605, 2007.

[31] N. A. Anaza, "Personality antecedents of customer citizenship behaviors in online shopping situations," *Psychology & Marketing*, vol. 31, no. 4, pp. 251–263, 2014.

[32] R. Giorgio and V. Alessandro, "Personality in computational advertising: A benchmark," in *ACM RecSys - Emotions and Personality in Personalized Systems, (EMPIRE 2016)*, 2016.

[33] P. J. Rentfrow and S. D. Gosling, "The do re mi's of everyday life: the structure and personality correlates of music preferences." *Journal of personality and social psychology*, vol. 84, no. 6, p. 1236, 2003.

[34] B. Ferwerda, E. Yang, M. Schedl, and M. Tkalcic, "Personality traits predict music taxonomy preferences," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems.* ACM, 2015, pp. 2241–2246.

[35] G. Matthews and I. J. Deary, *Personality traits.* Cambridge University Press, 1998.

[36] L. D. Goodstein and R. I. Lanyon, "Applications of personality assessment to the workplace: A review," *Journal of Business and Psychology*, vol. 13, no. 3, pp. 291–322, 1999.

[37] D. Frauendorfer, M. S. Mast, L. S. Nguyen, and D. Gatica-Perez, "The role of the applicant nonverbal behavior in the job interview and job performance," in *16th Congress of the European Association of Work and Organizational Psychology (EAWOP)*, Münster, Germany, May 2013.

[38] B. Schuller, "Voice and speech analysis in search of states and traits," in *Computer Analysis of Human Behavior.* Springer, 2011, pp. 227–253.

[39] F. Celli, E. Stepanov, and G. Riccardi, "Tell me who you are, i'll tell whether you agree or disagree: Prediction of agreement/disagreement in news blog," in *In Proceedings. of NLPMJ, in conjunction to IJCAI 2016.*, 2016.

[40] E. Stepanov, B. Favre, F. Alam, S. Chowdhury, K. Singla, J. Trione, F. Béchet, and G. Riccardi, "Automatic summarization of call-center conversations," in *In Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015.

[41] R. R. McCrae and P. T. Costa Jr, “A five-factor theory of personality,” *Handbook of personality: Theory and research*, vol. 2, pp. 139–153, 1999.

[42] W. Fisher, R. Groff, and H. Roane, “Applied behavior analysis: History, philosophy, principles, and basic methods,” *Handbook of applied behavior analysis*, pp. 3–13, 2011.

[43] S. Robbins, T. A. Judge, B. Millett, and M. Boyle, *Organisational behaviour.* Pearson Higher Education AU, 2013.

[44] P. N. Juslin and K. R. Scherer, “Vocal expression of affect,” *The new handbook of methods in nonverbal behavior research*, pp. 65–135, 2005.

[45] J. J. Gross and R. A. Thompson, “Emotion regulation: Conceptual foundations,” *Handbook of Emotion Regulation*, vol. 3, p. 24, 2007.

[46] N. H. Frijda, “Moods, emotion episodes, and emotions.” 1993.

[47] L. F. Barrett, M. Lewis, and J. M. Haviland-Jones, *Handbook of emotions.* Guilford Publications, 2016.

[48] H. M. Weiss and R. Cropanzano, “Affective events theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work.” 1996.

[49] M. L. Hoffman, “Empathy and prosocial behavior,” *Handbook of Emotions*, vol. 3, pp. 440–455, 2008.

[50] C. McCall and T. Singer, “Empathy and the brain,” *Understanding Other Minds: Perspectives from developmental social neuroscience*, pp. 195–214, 2013.

[51] A. Perry and S. Shamay-Tsoory, “Understanding emotional and cognitive empathy: A neuropsychological,” in *Understanding Other Minds:*

Perspectives from developmental social neuroscience. Oup Oxford, 2013, p. 178.

[52] M. Danieli, G. Riccardi, and F. Alam, "Emotion unfolding and affective scenes: A case study in spoken conversations," in *Proc. of Emotion Representations and Modelling for Companion Systems (ERM4CT) 2015,*. ICMI, 2015.

[53] O. P. John, R. W. Robins, and L. A. Pervin, *Handbook of personality: Theory and research.* G. Press, 2010.

[54] M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, *Handbook of emotions.* Guilford Press, 2010.

[55] P. Ekman, W. V. Friesen, M. O'Sullivan, and K. Scherer, "Relative importance of face, body, and speech in judgments of personality and affect." *Journal of Personality and Social Psychology*, vol. 38, no. 2, p. 270, 1980.

[56] C. E. Izard, "Basic emotions, relations among emotions, and emotion-cognition relations." 1992.

[57] C. E. Izard, *Human emotions.* Springer Science & Business Media, 2013.

[58] R. Plutchik, "The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[59] S. S. Tomkins, "Affect theory," *Approaches to emotion*, vol. 163, p. 195, 1984.

[60] W. Wundt, "Fundamentals of psychology," *Liepzig. Engelman*, 1905.

[61] R. Plutchik, *Emotion: A psychoevolutionary synthesis.* Harpercollins College Division, 1980.

[62] J. A. Russell, "Pancultural aspects of the human conceptual organization of emotions." *Journal of personality and social psychology*, vol. 45, no. 6, p. 1281, 1983.

[63] J. Panksepp, "The neurobiology of emotions: Of animal brains and human feelings." 1989.

[64] A. Landa, Z. Wang, J. A. Russell, J. Posner, Y. Duan, A. Kangarlu, Y. Huo, B. A. Fallon, and B. S. Peterson, "Distinct neural circuits subserve interpersonal and non-interpersonal emotions," *Social neuroscience*, vol. 8, no. 5, pp. 474–488, 2013.

[65] C. Darwin, *The expression of the emotions in man and animals.* University of Chicago press, 1965, vol. 526.

[66] S. S. Tomkins, "Affect, imagery, consciousness: Vol. i. the positive affects." 1962.

[67] P. Eckman, "Universal and cultural differences in facial expression of emotion," in *Nebraska symposium on motivation*, vol. 19. University of Nebraska Press Lincoln, 1972, pp. 207–284.

[68] C. Izard, "Human emotions,," 1977.

[69] K. Oatley and P. N. Johnson-Laird, "Towards a cognitive theory of emotions," *Cognition and emotion*, vol. 1, no. 1, pp. 29–50, 1987.

[70] K. Oatley and P. N. Johnson-Laird, "The communicative theory of emotions: Empirical tests, mental models, and implications for social interaction." 1996.

[71] J. Panksepp, *Affective neuroscience: The foundations of human and animal emotions.* Oxford university press, 1998.

[72] A. Ortony, G. Clore, and A. Collins, "The cognitive structure of emotions. 10.1017," *CBO9780511571299*, 1988.

[73] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowledge: further exploration of a prototype approach." *Journal of personality and social psychology*, vol. 52, no. 6, p. 1061, 1987.

[74] A. Moors and K. R. Scherer, "The role of appraisal in emotion," *Handbook of cognition and emotion*, pp. 135–155, 2013.

[75] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research.* Oxford University Press, 2001.

[76] J. J. Gross, *Handbook of emotion regulation.* Guilford Press, 2011.

[77] M. Shaikh, H. Prendinger, and M. Ishizuka, "A linguistic interpretation of the occ emotion model for affect sensing from text," *Affective Information Processing*, pp. 45–73.

[78] K. Stueber, "Empathy," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., 2014.

[79] E. B. Titchener, *Lectures on the Experimental Psychology of the Thought-processes.* Macmillan, 1909.

[80] C. Rogers, *A way of being.* Houghton Mifflin Harcourt, 1995.

[81] V. Gallese and A. Goldman, "Mirror neurons and the simulation theory of mind-reading," *Trends in Cognitive Sciences*, vol. 2, no. 12, pp. 493–501, 1998.

[82] S. Baron-Cohen, M. Lombardo, H. Tager-Flusberg, and D. Cohen, Eds., *Understanding Other Minds: Perspectives from Developmental Social Neuroscience*, 3rd ed. Oxford University Press, 2013.

[83] E. Fehr and U. Fischbacher, "The nature of human altruism," *Nature*, vol. 425, no. 6960, pp. 785–791, 2003.

[84] M. L. Hoffman, *Empathy and moral development: Implications for caring and justice.* Cambridge University Press, 2001.

[85] J. J. Gross, "Emotion and emotion regulation: Personality processes and individual differences." 2008.

[86] K. McLaren, *The Art of Empathy: A Complete Guide to Life's Most Essential Skill.* Sounds True, 2013.

[87] N. Eisenberg and R. A. Fabes, "Emotion, regulation, and the development of social competence." 1992.

[88] P. L. Lockwood, A. Seara-Cardoso, and E. Viding, "Emotion regulation moderates the association between empathy and prosocial behavior," *PloS one*, vol. 9, no. 5, p. e96555, 2014.

[89] D. Matsumoto, D. Keltner, M. N. Shiota, M. O'Sullivan, and M. Frank, "Facial expressions of emotion," *Handbook of emotions*, vol. 3, pp. 211–234, 2008.

[90] K. R. Scherer, "Vocal affect expression: a review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[91] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[92] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1, pp. 227–256, 2003.

[93] L. Leinonen, T. Hiltunen, I. Linnankoski, and M.-L. Laakso, "Expression of emotional–motivational connotations with a one-word utterance," *The Journal of the Acoustical society of America*, vol. 102, no. 3, pp. 1853–1863, 1997.

[94] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[95] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological bulletin*, vol. 129, no. 5, p. 770, 2003.

[96] R. Van Bezooijen, *Characteristics and recognizability of vocal expressions of emotion.* Walter de Gruyter, 1984, vol. 5.

[97] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog." in *INTERSPEECH.* Citeseer, 2002.

[98] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of Artificial Neural Networks in Engineering*, vol. 710, 1999.

[99] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, 2005.

[100] S. J. Mozziconacci and D. J. Hermes, “Expression of emotion and attitude through temporal speech variations.” in *INTERSPEECH*, 2000, pp. 373–378.

[101] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[102] B. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge.” in *Proc. of Interspeech*, 2009, pp. 312–315.

[103] E. Douglas-Cowie, “Wp5 members,“preliminary plans for exemplars: Databases,”,” technical report, The HUMAINE Assoc, Tech. Rep., 2004.

[104] R. Cowie, E. Douglas-Cowie, and C. Cox, “Beyond emotion archetypes: Databases for emotion modelling using neural networks,” *Neural networks*, vol. 18, no. 4, pp. 371–388, 2005.

[105] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, “Design, recording and verification of a danish emotional speech database.” in *Proc. of Eurospeech*, 1997.

[106] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech.” in *Proc. of INTERSPEECH*, vol. 5, 2005, pp. 1517–1520.

[107] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, “Getting started with susas: a speech under simulated and actual stress database.” in *Eurospeech*, vol. 97, no. 4, 1997, pp. 1743–46.

[108] O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The enterface’05 audio-visual emotion database,” in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on.* IEEE, 2006, pp. 8–8.

[109] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, “Audiovisual behavior modeling by combined feature spaces,” in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 2. IEEE, 2007, pp. II–733.

[110] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D’Arcy, M. J. Russell, and M. Wong, “" you stupid tin box"-children interacting with the aibo robot: A cross-linguistic emotional speech corpus.” in *LREC*, 2004.

[111] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, “Emotional speech: Towards a new generation of databases,” *Speech communication*, vol. 40, no. 1, pp. 33–60, 2003.

[112] K. R. Scherer and G. Ceschi, “Lost luggage: a field study of emotion–antecedent appraisal,” *Motivation and emotion*, vol. 21, no. 3, pp. 211–235, 1997.

[113] L. Devillers, L. Vidrascu, and L. Lamel, “Challenges in real-life emotion annotation and machine learning based detection,” *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[114] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, “Being bored? recognising natural interest by extensive audiovisual integration for real-life application,” *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[115] M. Grimm, K. Kroschel, and S. Narayanan, “The vera am mittag german audio-visual emotional speech database,” in *Multimedia and Expo, 2008 IEEE International Conference on.* IEEE, 2008, pp. 865–868.

[116] K. Laskowski, “Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings,” in *Acoustics, Speech*

and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009, pp. 4765–4768.

[117] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold, "Development of user-state conventions for the multimodal corpus in smartkom," in *LREC Workshop on" Multimodal Resources", Las Palmas, Spain*, 2002.

[118] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[119] T. Bänziger, V. Tran, and K. R. Scherer, "The geneva emotion wheel: A tool for the verbal report of emotional reactions," *Poster presented at ISRE*, vol. 149, pp. 271–294, 2005.

[120] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1, pp. 5–32, 2003.

[121] J. Cohen, "A coefficient of agreement for nominal scales. educational and psychosocial measurement, 20, 37-46," 1960.

[122] A. Batliner, S. Steidl, D. Seppi, and B. Schuller, "Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach," *Advances in Human-Computer Interaction*, vol. 2010, p. 3, 2010.

[123] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on.* IEEE, 2005, pp. 474–477.

[124] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes." in *INTERSPEECH*, 2004, pp. 205–211.

[125] M. T. Shami and M. S. Kamel, "Segment-based approach to the recognition of emotions in speech," in *ICME 2005.* IEEE, 2005, pp. 4–pp.

[126] J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen, "Segment-based emotion recognition from continuous mandarin chinese speech," *Computers in Human Behavior*, vol. 27, no. 5, pp. 1545–1552, 2011.

[127] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[128] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the 21st ACM international conference on Multimedia (ACMM).* ACM, 2013, pp. 835–838.

[129] S. J. Young and S. Young, *The HTK hidden Markov model toolkit: Design and philosophy.* University of Cambridge, Department of Engineering, 1993.

[130] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features.* Springer, 1998.

[131] J. Pohjalainen, S. Kadioglu, and O. Räsänen, "Feature selection for speaker traits." in *INTERSPEECH*, 2012.

[132] T. Polzehl, K. Schoenenberg, S. Moller, F. Metze, G. Mohammadi, and A. Vinciarelli, "On speaker-independent personality perception and prediction from speech," 2012.

[133] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[134] D. Wu, "Genetic algorithm based feature selection for speaker trait classification." in *INTERSPEECH*, 2012.

[135] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "Combining efforts for improving automatic classification of emotional user states," *Proc. IS-LTC*, pp. 240–245, 2006.

[136] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft Research, Tech. Rep., 1998.

[137] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[138] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational learning theory.* Springer, 1995, pp. 23–37.

[139] M. H. Sanchez, A. Lawson, D. Vergyri, and H. Bratt, "Multi-system fusion of extended context prosodic and cepstral features for paralinguistic speaker trait classification." in *INTERSPEECH*, 2012.

[140] M. Karahan, D. Hakkani-Tur, G. Riccardi, and G. Tur, "Combining classifiers for spoken language understanding," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on.* IEEE, 2003, pp. 589–594.

[141] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of perfor-

mances," in *Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2009, pp. 552–557.

[142] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features," *Speech Communication*, vol. 55, no. 1, pp. 1–21, 2013.

[143] M. H. Goodwin and C. Goodwin, "Emotion within situated activity," *Communication: An arena of development*, pp. 33–53, 2000.

[144] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Networks*, vol. 18, no. 4, pp. 317–352, 2005.

[145] M. Danieli, G. Riccardi, and F. Alam, "Emotion unfolding and affective scenes: A case study in spoken conversations," in *Proc. of Emotion Representations and Modelling for Companion Systems (ERM4CT) 2015,*. ICMI, 2015.

[146] J. Liscombe, G. Riccardi, and D. Hakkani-Tur, "Using context to improve emotion detection in spoken dialog systems," in *Proc. of Interspeech*, 2005, pp. 1845–1848.

[147] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[148] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[149] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. of 10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[150] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[151] T. Schmidt and K. Wörner, "Extensible markup language for discourse annotation (exmar-alda)," 2004.

[152] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[153] M. Davies and J. L. Fleiss, "Measuring agreement for multinomial data," *Biometrics*, pp. 1047–1051, 1982.

[154] S. Abrilian, L. Devillers, S. Buisine, and J.-C. Martin, "Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces," in *HCI International*, 2005.

[155] S. A. Chowdhury, G. Riccardi, and F. Alam, "Unsupervised recognition and clustering of speech overlaps in spoken conversations," in *Proc. of Workshop on Speech, Language and Audio in Multimedia - SLAM2014*, 2014, pp. 62–66.

[156] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011, pp. 1–4.

[157] E. Pianta, C. Girardi, and R. Zanoli, “The textpro tool suite.” in *LREC*. Citeseer, 2008.

[158] I. Kononenko, “Estimating attributes: analysis and extensions of relief,” in *Proc. of Machine Learning: European Conference on Machine Learning (ECML)*. Springer, 1994, pp. 171–182.

[159] J. W. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count: Liwc 2001,” *Mahway: Lawrence Erlbaum Associates*, vol. 71, 2001.

[160] F. Alam and G. Riccardi, “Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition,” in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 955–959.

[161] F. Alparone, S. Caso, A. Agosti, and A. Rellini, “The italian liwc2001 dictionary.” LIWC.net, Austin, TX, Tech. Rep., 2004.

[162] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: Liwc and computerized text analysis methods,” *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[163] F. Alam and G. Riccardi, “Comparative study of speaker personality traits recognition in conversational and broadcast news speech,” in *Proc. of Interspeech*. ISCA, 2013, pp. 2851–2855.

[164] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[165] J. Platt, *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. MIT Press, 1998. [Online]. Available: http://research.microsoft.com/~jplatt/smo.html

[166] NIST, *The 2009 RT-09 RIch transcription meeting recognition evaluation plan*, NIST, 2009.

[167] D. Castán, A. Ortega, A. Miguel, and E. Lleida, "Audio segmentation-by-classification approach based on factor analysis in broadcast news domain," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–13, 2014.

[168] D. Morena, R. Giuseppe, E. Barker, J. Foster, A. Funk, R. Gaizauskas, M. Hepple, E. Kurtic, M. Poesio, L. Molinari, and V. Giliberti, *Preliminary Version of Use Case Design. SENSEI project deliverable D1.1*, D. Morena and R. Giuseppe, Eds. University of Trento, 2014. [Online]. Available: http://www.sensei-conversation.eu/

[169] S. D. Preston and A. J. Hofelich, "The many faces of empathy: Parsing empathic phenomena through a proximate, dynamic-systems view of representing the other in the self," *Emotion Review*, vol. 4, no. 1, pp. 24–33, 2012.

[170] C. D. Batson, *The Social Neuroscience of Empathy.* MIT press, 2009, ch. These things called empathy: eight related but distinct phenomena.

[171] J. Decety and C. Lamm, "Human empathy through the lens of social neuroscience," *The Scientific World Journal*, vol. 6, pp. 1146–1163, 2006.

[172] P. R. Gesn and W. Ickes, "The development of meaning contexts for empathic accuracy: Channel and sequence effects." *Journal of Personality and Social Psychology*, vol. 77, no. 4, p. 746, 1999.

[173] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from

the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[174] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. of 10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[175] S. Kumano, K. Otsuka, D. Mikami, and J. Yamato, "Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings," in *Proc. of IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 2011, pp. 43–50.

[176] B. Xiao, P. Georgiou, and S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *Proc. of Asia Pacific Signal & Inf. Process. Assoc.*, 2012, pp. 1–4.

[177] B. Xiao, Bo, S. Daniel, I. Maarten Van, A. Zac E., G. David C., N. Panayiotis G., and S. S., "Modeling therapist empathy through prosody in drug addiction counseling," in *Proc. of Interspeech*, 2014, pp. 213–217.

[178] S. A. Chowdhury, M. Danieli, and G. Riccardi, "Annotating and categorizing competition in overlap speech," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[179] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proc. of Interspeech*, 2013.

[180] M. D. Pell and S. A. Kotz, "On the time course of vocal emotion recognition," *PLoS One*, vol. 6, no. 11, p. e27256, 2011.

[181] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition." in *INTERSPEECH*, 2006.

[182] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing," in *Affective Computing and Intelligent Interaction.* Springer, 2007, pp. 139–147.

[183] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[184] J. H. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011.* IEEE, 2011, pp. 4940–4943.

[185] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan, "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features." in *Proc. of Multimedia Signal Processing, 2007 (MMSP 2007)*, 2007, pp. 48–51.

[186] M. Mansoorizadeh and N. M. Charkari, "Speech emotion recognition: Comparison of speech segmentation approaches," *Proc of IKT, Mashad, Iran*, 2007.

[187] E. Weiste and A. Peräkylä, "Prosody and empathic communication in psychotherapy interaction," *Psychotherapy Research*, vol. 24, no. 6, pp. 687–701, 2014.

[188] J. L. Coulehan, F. W. Platt, B. Egener, R. Frankel, C.-T. Lin, B. Lown, and W. H. Salazar, "Let me see if i have this right…: words that help build empathy," *Annals of Internal Medicine*, vol. 135, no. 3, pp. 221–227, 2001.

[189] S. Barrett, *Keep Them Calling: Superior Service on the Telephone.* American Media Inc, 1996.

[190] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, pp. 321–357, 2002.

[191] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[192] P. Ladefoged and K. Johnson, *A course in phonetics.* Cengage learning, 2014.

[193] L. R. Rabiner and R. W. Schafer, *Theory and application of digital speech processing.* Pearson, 2009.

[194] T. Butko and C. Nadeu, "Audio segmentation of broadcast news in the albayzin-2010 evaluation: overview, results, and discussion," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, no. 1, pp. 1–10, 2011.

[195] A. Ortega, D. Castan, A. Miguel, and E. Lleida, *The Albayzin 2014 Audio Segmentation Evaluation*, 2014.

[196] M. K. Sönmez, M. Plauché, E. Shriberg, and H. Franco, "Consonant discrimination in elicited and spontaneous speech: a case for signal-adaptive front ends in asr." in *INTERSPEECH*, 2000, pp. 325–328.

[197] S. Tsakalidis, "Linear transforms in automatic speech recognition: Estimation procedures & integration of diverse acoustic data," Ph.D. dissertation, The Johns Hopkins University, 2005.

[198] D. Gildea, "Corpus variation and parser performance," in *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 2001, pp. 167–202.

[199] N. Webb and M. Ferguson, "Automatic extraction of cue phrases for cross-corpus dialogue act classification," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, ser. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1310–1317. [Online]. Available: http://dl.acm.org/citation.cfm?id=1944566.1944716

[200] D. Wang and Y. Liu, "A cross-corpus study of unsupervised subjectivity identification based on calibrated em," in *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, ser. WASSA '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 161–167. [Online]. Available: http://dl.acm.org/citation.cfm?id=2107653.2107674

[201] M. Shami and W. Verhelst, "Automatic classification of emotions in speech using multi-corpora approaches," in *Proc. of the second annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS 2006), Antwerp, Belgium*, 2006, pp. 3–6.

[202] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: a multi-corpus study," in *Speaker classification II*. Springer, 2007, pp. 43–56.

[203] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-corpus classification of realistic emotions–some pilot experiments," in

Proc. 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Valetta, 2010, pp. 77–82.

[204] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, “Cross-corpus acoustic emotion recognition: variances and strategies,” *Affective Computing, IEEE Transactions on*, vol. 1, no. 2, pp. 119–131, 2010.

[205] I. Lefter, L. J. Rothkrantz, P. Wiggers, and D. A. Van Leeuwen, “Emotion recognition from speech by combining databases and fusion of classifiers,” in *Text, Speech and Dialogue.* Springer, 2010, pp. 353–360.

[206] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, “Unsupervised learning in cross-corpus acoustic emotion recognition,” in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on.* IEEE, 2011, pp. 523–528.

[207] U. Gut, *Introduction to English phonetics and phonology.* Peter Lang, 2009, vol. 1.

[208] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan *et al.*, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” 2015.

[209] C. E. Izard, D. Z. Libero, P. Putnam, and O. M. Haynes, “Stability of emotion experiences and their relations to traits of personality.” *Journal of personality and social psychology*, vol. 64, no. 5, p. 847, 1993.

[210] W. Revelle and K. R. Scherer, “Personality and emotion,” *Oxford companion to emotion and the affective sciences*, pp. 304–306, 2009.

[211] D. J. Ozer and V. Benet-Martinez, “Personality and the prediction of consequential outcomes,” *Annu. Rev. Psychol.*, vol. 57, pp. 401–421, 2006.

[212] B. W. Roberts, N. R. Kuncel, R. Shiner, A. Caspi, and L. R. Goldberg, “The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes,” *Perspectives on Psychological Science*, vol. 2, no. 4, pp. 313–345, 2007.

[213] T. W. Bickmore and R. W. Picard, “Establishing and maintaining long-term human-computer relationships,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no. 2, pp. 293–327, 2005.

[214] M. Mount and M. Barrick, “Five reasons why the ‘big five’ article has been frequently cited,” *Personnel Psychology*, vol. 51, no. 4, pp. 849–857, 1998.

[215] J. Biel and D. Gatica-Perez, “The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs,” *IEEE Trans. on Multimedia*, vol. 15, no. 1, pp. 41–55, Jan 2013.

[216] T. Chamorro-Premuzic and A. Furnham, *Personality and intellectual competence.* Psychology Press, 2014.

[217] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr, “A very brief measure of the big-five personality domains,” *Journal of Research in personality*, vol. 37, no. 6, pp. 504–528, 2003.

[218] S. Whelan and G. Davies, “Profiling consumers of own brands and national brands using human personality,” *Journal of Retailing and Consumer Services*, vol. 13, no. 6, pp. 393–402, 2006.

[219] D. Schultz and S. Schultz, *Theories of personality.* Cengage Learning, 2016.

[220] R. Ewen, *An introduction to theories of personality.* Psychology Press, 2014.

[221] W. A. GORDON, "Personality: A psychological interpretation." 1937.

[222] R. B. Cattell, *Personality and mood by questionnaire.* Jossey-Bass, 1973.

[223] H. J. Eysenck, "Personality and experimental psychology: The unification of psychology and the possibility of a paradigm." *Journal of Personality and social Psychology*, vol. 73, no. 6, p. 1224, 1997.

[224] L. R. Goldberg, "The development of markers for the big-five factor structure." *Psychological assessment*, vol. 4, no. 1, p. 26, 1992.

[225] B. De Raad, *The Big Five Personality Factors: The psycholexical approach to personality.* Hogrefe &amp; Huber Publishers, 2000.

[226] B. E. de Raad and M. E. Perugini, *Big five assessment.* Hogrefe & Huber Publishers, 2002.

[227] G. Saucier and L. R. Goldberg, "Lexical studies of indigenous personality factors: Premises, products, and prospects," *Journal of personality*, vol. 69, no. 6, pp. 847–879, 2001.

[228] P. T. Costa and R. R. MacCrae, *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual.* Psychological Assessment Resources, 1992.

[229] C. Chiorri, F. Bracco, T. Piccinno, C. Modafferi, and V. Battini, "Psychometric properties of a revised version of the ten item personality inventory," *European Journal of Psychological Assessment*, 2015.

[230] G. J. Boyle and E. Helmes, *The Cambridge Handbook of Personality Psychology.* Cambridge University Press Cambridge, 2009, ch. Methods of personality assessment.

[231] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction." *Journal of Experimental Psychology: Applied*, vol. 7, no. 3, p. 171, 2001.

[232] C. I. Nass and S. Brave, *Wired for speech: How voice activates and advances the human-computer relationship.* MIT press Cambridge, 2005.

[233] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one*, vol. 8, no. 9, p. e73791, 2013.

[234] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander, "Large scale personality classification of bloggers," in *ACII.* Springer, 2011, pp. 568–577.

[235] S. Love and J. Kewley, "Does personality affect peoples' attitude towards mobile phone use in public places?" in *Mobile Communications.* Springer, 2005, pp. 273–284.

[236] J. M. Balmaceda, S. Schiaffino, and D. Godoy, "How do personality traits affect communication among users in online social networks?" *Online Information Review*, vol. 38, no. 1, pp. 136–153, 2014.

[237] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text." *J. Artif. Intell. Res.(JAIR)*, vol. 30, pp. 457–500, 2007.

[238] R. G. Ivanov A. V., "Automatic turn segmentation in spoken conversations," in *Proc. of Interspeech*, 2010.

[239] S. Butt and J. G. Phillips, "Personality and self reported mobile phone use," *Computers in Human Behavior*, vol. 24, no. 2, pp. 346–360, 2008.

[240] L. M. Batrinca, "Multimodal personality recognition from audiovisual data," Ph.D. dissertation, University of Trento, 2013.

[241] F. Celli, "Adaptive personality recogntion from text," Ph.D. dissertation, University of Trento, 2012.

[242] K. Kalimeri, "Traits, states and situations: Automatic prediction of personality and situations from actual behavior," Ph.D. dissertation, University of Trento, 2013.

[243] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems.* ACM, 2011, pp. 253–262.

[244] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Egloff, and S. D. Gosling, "Facebook profiles reflect actual personality, not self-idealization," *Psychological Science*, vol. 21, no. 3, pp. 372–374, 2010.

[245] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell, "Personality and patterns of facebook usage," in *Proceedings of the 3rd Annual ACM Web Science Conference.* ACM, 2012, pp. 24–32.

[246] G. Farnadi, S. Zoghbi, M.-F. Moens, and M. De Cock, “How well do your facebook status updates express your personality?” in *Proceedings of the 22nd edition of the annual Belgian-Dutch conference on machine learning (BENELEARN)*, 2013.

[247] G. Farnadi, S. Zoghbi, M.-F. Moens, and M. De Cock, “Recognising personality traits using facebook status updates,” in *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13)*, 2013.

[248] G. Farnadi, G. Sitaraman, M. Rohani, M. Kosinski, D. Stillwell, M.-F. Moens, S. Davalos, and M. De Cock, “How are you doing? study of emotion expression from facebook status updates with users’ age, gender, personality and time,” in *In Proc. of the EMPIRE Workshop on UMAP*, 2014.

[249] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, “Predicting personality from twitter,” in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on.* IEEE, 2011, pp. 149–156.

[250] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, “Our twitter profiles, our selves: Predicting personality with twitter,” in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on.* IEEE, 2011, pp. 180–185.

[251] E. Kafeza, A. Kanavos, C. Makris, and P. Vikatos, “T-pice: Twitter personality based influential communities extraction system,” in *Big Data*

(BigData Congress), 2014 IEEE International Congress on. IEEE, 2014, pp. 212–219.

[252] D. O. Olguın, P. A. Gloor, and A. S. Pentland, "Capturing individual and group behavior with wearable sensors," in *Proceedings of the 2009 aaai spring symposium on human behavior modeling, SSS*, vol. 9, 2009.

[253] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Who's who with big-five: Analyzing and classifying personality traits with smartphones," in *Wearable Computers (ISWC), 2011 15th Annual International Symposium on.* IEEE, 2011, pp. 29–36.

[254] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Mining large-scale smartphone data for personality studies," *Personal and Ubiquitous Computing*, vol. 17, no. 3, pp. 433–450, 2013.

[255] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *Affective Computing, IEEE Transactions on*, vol. 5, no. 3, pp. 273–291, 2014.

[256] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[257] M. Saif. and K. Svetlana, "Using nuances of emotion to identify personality," in *ICWSM-WCPR*, 2013.

[258] F. Iacobelli and A. Culotta, "Too neurotic, not too friendly: Structured personality classification on textual data," in *AAAI*, 2013.

[259] E. Sapir, "Speech as a personality trait," *American Journal of Sociology*, pp. 892–905, 1927.

[260] K. Audhkhasi, A. Metallinou, M. Li, and S. Narayanan, "Speaker personality classification using systems based on acoustic-lexical cues and an optimal tree-structured bayesian network." in *INTERSPEECH*, 2012.

[261] J. Wagner, F. Lingenfelser, and E. André, "A frame pruning approach for paralinguistic recognition tasks." in *Proc. of Interspeech*, 2012.

[262] M. D. Robinson, E. R. Watkins, and E. Harmon-Jones, *Handbook of cognition and emotion.* Guilford Press, 2013.

[263] Y. Attabi and P. Dumouchel, "Anchor models and wccn normalization for speaker trait classification," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[264] C. Montacié and M.-J. Caraty, "Pitch and intonation contribution to speakers' traits classification." in *INTERSPEECH*, 2012.

[265] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker, "Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life." *Journal of personality and social psychology*, vol. 90, no. 5, p. 862, 2006.

[266] A. J. Gill and R. M. French, "Level of representation and semantic distance: Rating author personality from texts," *Proc. Euro Cogsci, Delphi, Greece*, 2007.

[267] J. Oberlander and A. Gill, "Individual differences and implicit language: personality, parts-of-speech and pervasiveness," *Proceedings of the 26th AnnualConference of the Cognitive Science Society*, pp. 1035–1040, 2004.

[268] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, "Lexical predictors of personality type," in *In Proceedings of the Joint Annual*

Meeting of the Interface and the Classification Society of North America. Citeseer, 2005.

[269] J. Oberlander and S. Nowson, "Whose thumb is it anyway?: classifying author personality from weblog text," in *Proceedings of the COLING/ACL on Main conference poster sessions.* Association for Computational Linguistics, 2006, pp. 627–634.

[270] A. Minamikawa and H. Yokoyama, "Blog tells what kind of personality you have: egogram estimation from japanese weblog," in *Proceedings of the ACM 2011 conference on Computer supported cooperative work.* ACM, 2011, pp. 217–220.

[271] A. J. Gill, S. Nowson, and J. Oberlander, "What are they blogging about? personality, topic and motivation in blogs." in *ICWSM*, 2009.

[272] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks: towards socially and personality aware visual surveillance," in *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis.* ACM, 2010, pp. 37–42.

[273] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proceedings of the 10th international conference on Multimodal interfaces.* ACM, 2008, pp. 53–60.

[274] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, "A multimodal annotated corpus of consensus decision making meetings," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 409–429, 2007.

[275] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, "Please, tell me about yourself: automatic personality assessment using short

self-presentations," in *Proceedings of the 13th international conference on multimodal interfaces.* ACM, 2011, pp. 255–262.

[276] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion&amp;# x2014; a systematic study," *Affective Computing, IEEE Transactions on*, vol. 3, no. 4, pp. 443–455, 2012.

[277] J. Fitzgerald Steele, D. C. Evans, and R. K. Green, "Is your profile picture worth 1000 words? photo characteristics associated with personality impression agreement." *landscape*, vol. 2, p. 139, 2009.

[278] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina, "Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis," in *Proceedings of the 21st ACM international conference on Multimedia.* ACM, 2013, pp. 213–222.

[279] F. Celli, E. Bruni, and B. Lepri, "Automatic personality and interaction style recognition from facebook profile pictures," in *Proceedings of the ACM International Conference on Multimedia.* ACM, 2014, pp. 1101–1104.

[280] R. Srivastava, J. Feng, S. Roy, S. Yan, and T. Sim, "Don't ask me what i'm like, just watch and listen," in *Proceedings of the 20th ACM international conference on Multimedia.* ACM, 2012, pp. 329–338.

[281] M. Rojas, D. Masip, A. Todorov, and J. Vitria, "Automatic prediction of facial trait judgments: Appearance vs. structural models," *PloS one*, vol. 6, no. 8, p. e23323, 2011.

[282] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval.* Cambridge University Press Cambridge, 2008, vol. 1.

[283] K. L. Kermanidis, "Mining authors' personality traits from modern greek spontaneous text," in *4th International Workshop on Corpora for Research on Emotion Sentiment &amp; Social Signals, in conjunction with LREC12.* Citeseer, 2012, pp. 90–94.

[284] B. Schuller, R. Müller, M. K. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles." in *Proc. of INTERSPEECH*, 2005, pp. 805–808.

[285] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia.* ACM, 2010, pp. 1459–1462.

[286] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, vol. 97, 1997, pp. 412–420.

[287] Z.-H. Zhou, *Ensemble methods: foundations and algorithms.* CRC Press, 2012.

[288] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[289] J. W. Pennebaker, "The secret life of pronouns," *New Scientist*, vol. 211, no. 2828, pp. 42–45, 2011.

[290] L. A. Fast and D. C. Funder, "Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior." *Journal of personality and social psychology*, vol. 94, no. 2, p. 334, 2008.

[291] F. Alam and R. Zanoli, "A combination of classifiers for named entity recognition on transcription," *Evaluation of Natural Language and Speech Tools for Italian*, pp. 107–115, 2013.

[292] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of the international conference on new methods in language processing*, vol. 12. Citeseer, 1994, pp. 44–49.

[293] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[294] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of HLT-NAACL 2003*. ACL, 2003, pp. 173–180.

[295] C. Strapparava and A. Valitutti, "Wordnet affect: an affective extension of wordnet." in *LREC*, vol. 4, 2004, pp. 1083–1086.

[296] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 2200–2204.

[297] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of EMNLP*. Citeseer, 2013, pp. 1631–1642.

[298] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[299] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior," *Journal of Management Information Systems*, vol. 29, no. 4, pp. 217–248, 2013.

[300] G. Roffo and A. Vinciarelli, "Personality in computational advertising: A benchmark," in *4 th Workshop on Emotions and Personality in Personalized Systems (EMPIRE) 2016*, 2016, p. 18.

[301] H. Liu, "Social network profiles as taste performances," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 252–275, 2007.

[302] E. Cambria, C. Havasi, and A. Hussain, "Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis." in *FLAIRS conference*, 2012, pp. 202–207.

[303] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on.* IEEE, 2010, pp. 1–10.

[304] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell, "Personality and patterns of facebook usage," in *Proceedings of the ACM Web Science Conference.* ACM New York, NY, USA, 2012, pp. 36–44.

[305] M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, and T. Graepel, "Manifestations of user personality in website choice and behaviour on online social networks," *Machine Learning, Volume 95, Issue 3*, pp. 357–380., 2013.

[306] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems.* ACM, 2011, pp. 253–262.

[307] P. T. Costa and R. R. McCrae, "The revised neo personality inventory (neo-pi-r)," *In G.J. Boyle, G Matthews and D. Saklofske (Eds.). The SAGE handbook of personality theory and assessment*, vol. 2, pp. 179–198, 2008.

[308] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: Liwc and computerized text analysis methods,” *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[309] J. Bollen, H. Mao, and A. Pepe, “Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena,” in *ICWSM*, 2011, pp. 1–10.

[310] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern, “Predicting information spreading in twitter,” in *Workshop on computational social science and the wisdom of crowds, nips*, 2010, pp. 599–601.

[311] M. Guerini, C. Strapparava, and G. Özbal, “Exploring text virality in social networks,” in *Proceedings of ICWSM*, 2011, pp. 1–5.

[312] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, “Want to be retweeted? large scale analytics on factors impacting retweet in twitter network,” in *Social computing (socialcom), 2010 ieee second international conference on*. IEEE, 2010, pp. 177–184.

[313] S. Stieglitz and L. Dang-Xuan, “Emotions and information diffusion in social media?sentiment of microblogs and sharing behavior,” *Journal of Management Information Systems*, vol. 29, no. 4, pp. 217–248, 2013.

[314] A. Dobele, A. Lindgreen, M. Beverland, J. Vanhamme, and R. Van Wijk, “Why pass on viral messages? because they connect emotionally,” *Business Horizons*, vol. 50, no. 4, pp. 291–304, 2007.

[315] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, “Automatically profiling the author of an anonymous text,” *Communications of the ACM*, vol. 52, no. 2, pp. 119–123, 2009.

[316] D. Bogdanova, P. Rosso, and T. Solorio, “Exploring high-level features for detecting cyberpedophilia,” *Computer Speech & Language*, vol. 28, no. 1, pp. 108–120, 2014.

[317] A. L. Hill, D. G. Rand, M. A. Nowak, and N. A. Christakis, “Emotions as infectious diseases in a large social network: the sisa model,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 277, no. 1701, pp. 3827–3835, 2010.

[318] F. Celli and C. Zaga, “Be conscientious, express your sentiment!” in *proceedings of ESSEM, in conjunction with AIXIA2013, Turin*, 2013, pp. 52–56.

[319] E. Cambria, C. Havasi, and A. Hussain, “Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis,” in *FLAIRS Conference*, 2012, pp. 202–207.

[320] J. Staiano and M. Guerini, “Depeche mood: a lexicon for emotion analysis from crowd annotated news,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, vol. 2. The Association for Computer Linguistics, 2014, pp. 427–433.

[321] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, “Profile of mood states (poms),” in *STOP, THAT and One Hundred Other Sleep Scales.* Springer, 2012, pp. 285–286.

[322] D. Boyd, S. Golder, and G. Lotan, “Tweet, tweet, retweet: Conversational aspects of retweeting on twitter,” in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on.* IEEE, 2010, pp. 1–10.

[323] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” in *LREC*, 2010, pp. 1320–1326.

[324] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in twitter," *Language resources and evaluation*, vol. 47, no. 1, pp. 239–268, 2013.

[325] W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li, "Topical keyphrase extraction from twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* Association for Computational Linguistics, 2011, pp. 379–388.

[326] F. Celli and L. Rossi, "The role of emotional stability in twitter conversations," in *Proceedings of the Workshop on Semantic Analysis in Social Media.* Association for Computational Linguistics, 2012, pp. 10–17.

[327] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom).* IEEE, 2011, pp. 180–185.

[328] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom).* IEEE, 2011, pp. 149–156.

[329] D. Boyd and N. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[330] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english

and german," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[331] T. Buchanan, J. A. Johnson, and L. R. Goldberg, "Implementing a five-factor personality inventory for use on the internet," *European Journal of Psychological Assessment*, vol. 21, no. 2, pp. 115–127, 2005.

[332] C. H. Miller, "Indignation, defensive attribution, and implicit theories of moral character," 2000.

[333] C. Gerlitz and A. Helmond, "Hit, link, like and share. organising the social and the fabric of the web." *proceedings of Digital Methods Winter Conference*, pp. 1–29, January 2011.