



Modeling of Emotional Subjectivity in Affective-Based Predictive Systems

by

Hassan Hayat

ORCID: [0000-0003-0979-5623](https://orcid.org/0000-0003-0979-5623)

Supervisor(s): Prof. Agata Lapedriza Garcia
Prof. Carles Ventura Royo

A thesis submitted in total fulfillment for the
degree of Doctor of Philosophy in
Network and Information Technologies (NIT)
Department of IT, Multimedia and Telecommunications (IMT)
Doctoral School
Universitat Oberta de Catalunya

February 15, 2023

Abstract

One of the goals of affective computing is to develop affective technologies that can understand humans emotionally and make their life better. Human emotions are highly subjective in nature. This is why systems that consider affective along with subjective information play a significant role not only in mimicking an individual's cognitive process but also in an individual's interaction with others. This thesis targets emotional subjectivity in affect-related tasks. In particular, this thesis studies subjectivity from two different perspectives: **(I)** subjectivity in the annotations, and **(II)** subjectivity according to personality traits. Regarding annotations, in supervised machine learning, affective systems are trained and tested on annotated datasets. Usually, these annotations are the aggregation of multiple subjective annotations which basically represent each annotator's subjective emotional perception. The common practice to get aggregated annotations is by computing the average score and majority voting of multiple subjective annotations. These aggregated labels lose subjective information. Systems that are trained and tested based on these aggregated annotations have poor generalization capabilities for predicting subjective emotional perception. To tackle this problem, we proposed a Multi-Task (MT) learning approach that has the capability to learn each subjective emotional perception available in the annotations separately. The results show that our MT approach (that considers all subjective annotations separately) has more generalization capabilities as compared to approaches that are trained only on aggregated annotations. The second part of the thesis presents the study in the context of dialogues. Concretely, we studied the problem of predicting subjective emotional responses for the upcoming utterance with respect to each speaker in the conversation. We developed a Multi-Task (MT) learning approach that has the capability to predict multiple subjective emotional responses in the conversation using the personality information of each speaker. The results show that separate modeling of each speaker's emotional responses using joint modeling (i.e. Multi-Task learning) is better than combined modeling of all speaker's emotional responses.

The last part of this thesis focuses on the automatic recognition of personality traits in human speech signals. We proposed an interpretable model that takes audio speech as input to predict the personality traits of the speaker based on Big five representation. With our interpretable model, we found distinct frequency patterns for each Big five personality trait in human speech. The developed model performed better, is more lightweight, and has interpretable properties.

Resum

Un dels objectius de la computació afectiva és desenvolupar tecnologies afectives que puguin entendre els humans emocionalment i millorar la seva vida. Les emocions humanes són de naturalesa altament subjectiva. És per això que els sistemes que consideren la informació afectiva juntament amb la subjectiva juguen un paper important no només a l'hora d'imitar el procés cognitiu d'un individu, sinó també en la interacció d'un individu amb els altres. Aquesta tesi té com a objectiu estudiar la subjectivitat emocional en tasques relacionades amb l'afecte o les emocions. En particular, aquesta tesi estudia la subjectivitat des de dues perspectives diferents: **(I)** subjectivitat en les anotacions, i **(II)** subjectivitat segons trets de personalitat. Pel que fa a les anotacions, en l'aprenentatge automàtic supervisat, els sistemes afectius s'entrenen i es testegen en conjunts de dades anotats. Normalment, aquestes anotacions són l'agregació de múltiples anotacions subjectives que representen bàsicament la percepció emocional subjectiva de cada anotador. La pràctica habitual per obtenir anotacions agregades és calcular la puntuació mitjana o la votació majoritària de múltiples anotacions subjectives. Aquestes etiquetes agregades perden informació subjectiva. Els sistemes entrenats i testejats a partir d'aquestes anotacions agregades tenen poques capacitats de generalització per predir la percepció emocional subjectiva. Per fer front a aquest problema, vam proposar un enfocament d'aprenentatge multitasca (MT) que té la capacitat d'aprendre cada percepció emocional subjectiva disponible a les anotacions per separat. Els resultats mostren que el nostre enfocament MT (que considera totes les anotacions subjectives per separat) té més capacitats de generalització en comparació amb els enfocaments que només s'entrenen en anotacions agregades. La segona part de la tesi presenta l'estudi en el context dels diàlegs. Concretament, hem estudiat el problema de predir respostes emocionals subjectives respecte cada participant de la conversa. Hem desenvolupat un enfocament d'aprenentatge multitasca (MT) que té la capacitat de predir múltiples respostes emocionals subjectives a la conversa utilitzant la informació de la personalitat de cada parlant. Els resultats mostren que el modelatge separat de les respostes emocionals de cada parlant mitjançant el modelatge conjunt (és a dir, l'aprenentatge multitasca) és millor que el modelatge combinat de les respostes emocionals de tots els parlants. L'última part d'aquesta tesi se centra en el reconeixement automàtic de trets de personalitat en els senyals de la parla humana.

Vam proposar un model interpretable que pren la parla d'àudio com a entrada per predir els trets de personalitat del parlant basant-se en la representació dels Big Five. Amb el nostre model interpretable, hem trobat diferents patrons de freqüència per a cada tret de personalitat dels Big Five en la parla humana. El model desenvolupat funciona millor, és més lleuger i té propietats interpretables.

Resumen

Uno de los objetivos de la computación afectiva es desarrollar tecnologías afectivas que puedan comprender emocionalmente a los humanos y mejorar su vida. Las emociones humanas son de naturaleza altamente subjetiva. Esta es la razón por la que los sistemas que consideran la información afectiva junto con la subjetiva juegan un papel importante no solo en la imitación del proceso cognitivo de un individuo, sino también en la interacción de un individuo con los demás. Esta tesis se enfoca en el estudio de la subjetividad emocional en tareas relacionadas con el afecto. En particular, esta tesis estudia la subjetividad desde dos perspectivas diferentes: *(I)* subjetividad en las anotaciones, y *(II)* subjetividad según los rasgos de personalidad. Con respecto a las anotaciones, en el aprendizaje automático supervisado, los sistemas afectivos se entrenan y testean en conjuntos de datos anotados. Por lo general, estas anotaciones son la agregación de múltiples anotaciones subjetivas que básicamente representan la percepción emocional subjetiva de cada anotador. La práctica común para obtener anotaciones agregadas es calcular la anotación promedio o la votación mayoritaria de múltiples anotaciones subjetivas. Estas etiquetas agregadas pierden información subjetiva. Los sistemas que se entrenan y testean en función de estas anotaciones agregadas tienen capacidades de generalización deficientes para predecir la percepción emocional subjetiva. Para abordar este problema, propusimos un enfoque de aprendizaje multitarea (MT) que tiene la capacidad de aprender cada percepción emocional subjetiva disponible en las anotaciones por separado. Los resultados muestran que nuestro enfoque MT (que considera todas las anotaciones subjetivas por separado) tiene más capacidades de generalización en comparación con los enfoques que se entrenan solo en anotaciones agregadas. La segunda parte de la tesis presenta un estudio en el contexto de los diálogos. Concretamente, estudiamos el problema de predecir respuestas emocionales subjetivas para el próximo enunciado con respecto a cada hablante en la conversación. Desarrollamos un enfoque de aprendizaje multitarea (MT) que tiene la capacidad de predecir múltiples respuestas emocionales subjetivas en la conversación utilizando la información de personalidad de cada hablante. Los resultados muestran que el modelado separado de las respuestas emocionales de cada hablante mediante el modelado conjunto (es decir, el aprendizaje de tareas múltiples) es mejor que el modelado combinado de las respuestas emocionales de todos los hablantes. La última parte de esta tesis se centra en

el reconocimiento automático de rasgos de personalidad en las señales del habla humana. Propusimos un modelo interpretable que toma el habla de audio como entrada para predecir los rasgos de personalidad del hablante en función de la representación de los Big Five. Con nuestro modelo interpretable, encontramos patrones de frecuencia distintos para cada rasgo de personalidad de los Big Five en el habla humana. El modelo desarrollado funcionó mejor, es más ligero y tiene propiedades interpretables.

Declaration of Authorship

I, HASSAN HAYAT, declare that this thesis titled, 'MODELING OF EMOTIONAL SUBJECTIVITY IN AFFECTIVE-BASED PREDICTIVE SYSTEMS' and the work presented in it are my own. I confirm that:

- The thesis comprises only my original work toward the Doctor of Philosophy in Network and Information Technologies at Universitat Oberta de Catalunya.
- Any part of this thesis has not been previously submitted for a degree and other qualification at Universitat Oberta de Catalunya or any other institute.
- The parts of the thesis have been published before submission. References are clearly mentioned in the published parts of the thesis.

Author's Signed: **Hassan Hayat**

Certified By: **Prof. Agata Lapedriza Garcia**

Certified By: **Prof. Carles Ventura Royo**

Date: **February 15, 2023**

Preface

This thesis presents my work on the topic of Emotional Subjectivity in Affective Computing during the period of 2018 to 2023. The journey has been very educational in both personal and professional ways. I have had the privilege to work under very talented and supportive people. I want to thank my both supervisors who were with me on my whole journey.

Affective computing is all about human emotion, sentiment, and feelings. Emotions are very much a part of our experience also known as an intrinsic element of human experience. Because of this subjective nature of emotions, they can easily blind affective systems in predicting subjective emotions. This study introduced how we incorporate subjectivity in affective-based systems. The thesis has two parts:

The Part I presents a technique for modeling subjective emotional perceptions in affect-related tasks. The proposed technique was first published in *Affective Computing and Intelligent Interaction (ACII-2021)* and later a more detailed study was published in the *Sensors* journal in 2022.

The second part (Part II) focuses on the role of personality traits in predicting subjective emotions. The considered problem for this is to predict the subjective emotional responses to the next utterance during the conversation. The research findings support that personality traits play a significant role in predicting subjective emotional responses in conversation. The study is submitted to the 11th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2023), and it is currently under review.

To understand the causation of distinct emotions due to personality traits, a visual representation of personality traits in human speech is presented in this thesis. For this, an interpretable machine-learning model is designed for human speech. This model discovers the distinct frequency patterns that are associated with each big five personality trait (extraversion, agreeableness, openness, conscientiousness, and neuroticism) in human speech signals. The research findings were published at the conference (Conference of the Catalan Association for Artificial Intelligence-2019) proceedings book, titled “Frontiers in Artificial Intelligence and Applications”

Acknowledgements

I would like to thank my supervisors Prof. Agata Lapedriza and Prof. Carles Ventura for their guidance, patience, and support throughout my Ph.D. journey. I have benefited greatly from your wealth of knowledge and meticulous editing. I am extremely grateful that you took me on as a Ph.D. student and continued to have faith in me over the years.

Lastly, my family deserves endless gratitude; my mother has continuously given me moral support throughout the years, my sisters boosted me to fulfill my dreams, and my brother always kept in touch with me. Thank you for all of your love and for always reminding me of the end goal.

Contents

Abstract	i
Declaration of Authorship	vii
Preface	viii
Acknowledgements	ix
List of Figures	xiii
List of Tables	xv
Introduction	1
I Incorporate Subjectivity Using Soft Labels	5
1 Emotional Subjectivity in Affect Labeling	6
1.1 Annotator Subjectivity	9
2 Literature Review	13
2.1 Emotion Representation Models	14
2.1.1 Categorical Models	14
2.1.2 Dimensional Models	16
2.2 Affect-related Datasets	17
2.2.1 Subjective Annotations in Affect-related Datasets	26
2.3 Measuring Annotator Agreement	26
2.3.1 Pearson’s Correlation Coefficient	27
2.3.2 Spearman’s Correlation Coefficient	28
2.3.3 Cohen’s Kappa	29
2.3.4 Krippendorff’s Alpha	30
2.3.5 Scott’s Pi	31
2.3.6 Gwet AC1	31
2.3.7 Fleiss Kappa	32
2.3.8 Rosenberg and Binkowski Kappa	33

2.4	Approaching Agreement/Disagreement from a Machine Learning Perspective	34
2.4.1	Subjectivity as Noise	35
2.4.2	Subjectivity as Information	37
3	A Multi-Task (MT) Learning Approach to Model Subjectivity in Affect-related Tasks	39
3.1	An Introduction to Multi-Task (MT) Learning	40
3.2	Modeling Subjective Annotations with Multi-Task (MT) Learning	43
3.2.1	Single-Task (ST) Architecture	43
3.2.2	Multi-Task (MT) Architecture	45
3.3	The MT Approach in the Context of Related Work	46
4	Experiments and Results	49
4.1	Implementation Details	50
4.1.1	ST Architecture	51
4.1.2	MT Architecture	52
4.2	Evaluation Metrics	53
4.2.1	Classification Accuracy	53
4.2.2	Unweighted Average Recall (UAR)	53
4.3	Experiments with Synthetic Dataset	54
4.3.1	Data generation	55
4.3.2	Results	55
4.4	Experiments with Human-Annotated Datasets	57
4.4.1	COGNIMUSE Dataset	59
4.4.1.1	Data Distribution and Annotator Agreement Analysis	60
4.4.1.2	Backbone Architecture	60
4.4.1.3	Results	63
4.4.1.4	Qualitative Analysis	64
4.4.2	IEMOCAP Dataset	66
4.4.2.1	Data Distribution and Annotator Agreement Analysis	66
4.4.2.2	Backbone Architecture	67
4.4.2.3	Results	68
4.4.2.4	Qualitative Analysis	70
4.4.3	SemEval_2007 Dataset	71
4.4.3.1	Data Distribution and Annotator Agreement Analysis	72
4.4.3.2	Backbone Architecture	73
4.4.3.3	Results	73
4.4.3.4	Qualitative Analysis	75
5	Discussion	76

II	Emotion Subjectivity and Personality Traits	80
6	Introduction	81
7	A Study on Modeling Subjective Emotion Expression using Personality Traits in the Context of Dialogue Systems	85
7.1	Related work	89
7.2	Dialogue Datasets with Emotion Labels	91
7.3	Proposed Approach	95
7.3.1	Single-Task (ST) Architecture	96
7.3.2	Multi-Task (MT) Architecture	97
7.3.3	Implementation Details	98
7.4	Experiments and Results	102
7.4.1	PELD Dataset	102
7.4.2	Results	103
7.4.2.1	Predicting Response Emotions with 3 Categories	104
7.4.2.2	Predicting Response Emotions with 7 Categories	106
7.4.2.3	Predicting Response Emotions without Personality	110
7.5	Qualitative Analysis	112
7.6	Discussion	122
8	Automatic Recognition of Personality Traits	125
8.1	Related Work	128
8.1.1	Automatic Personality Recognition	128
8.1.2	Audio Classification	129
8.2	End-to-End Approach for Personality Trait Recognition from Audio	130
8.2.1	Audio Pre-Processing	130
8.2.2	Architectures and Implementation Details	131
8.3	Experiments	132
8.3.1	First-Impression Dataset	132
8.3.2	Evaluation metric	133
8.3.3	Results	134
8.4	Interpretability for Personality Trait Recognition from Audio	135
8.4.1	Interpretable CNN for Clip-Spectrogram (Our-ICS) and Summary-Spectrogram (Our-ISS)	137
8.4.1.1	Results	137
8.4.2	Interpretable CNN for Raw Audio Data (Our-IRA)	140
8.4.2.1	Results	142
8.5	Discussion	142
	Conclusion	146

List of Figures

1.1	An example of labeling step: multiple annotators annotate each movie clip into Negative (0) and Positive (1) emotional categories. . .	10
2.1	Ekman’s model of six basic emotion	15
2.2	Plutchik’s wheel of emotions	15
2.3	PAD (Pleasure, Arousal, Dominance) emotion representation model	16
2.4	2D (pleasure and arousal) emotion representation model	17
2.5	An interactive clip between human and SAL agent in SEMAINE Dataset	18
2.6	A sample video clip of the RECOLA Dataset	18
2.7	A video clip annotated in multiple affect classes in HUMAINE Dataset	19
2.8	A video sample available in FlimStim Dataset	20
2.9	A sample video clip from LIRIS-ACCEDE Dataset	21
2.10	A sample video clip from MediaEval 2015 Dataset	21
2.11	A sample utterance was annotated into different affect dimensions in the IEMOCAP Dataset	22
2.12	A movie clip annotated by multiple annotators in COGNIMUSE Dataset	23
2.13	A Newline annotated into two different dimensions in <i>SemEval_2007</i>	24
2.14	A tweet sample from GoEmotion dataset	25
2.15	A sample of MovieGraph dataset	25
3.1	Soft parameter sharing multi-task learning	41
3.2	Hard parameter sharing multi-task learning	42
3.3	Architecture of the Single-Task (ST) model	43
3.4	Architecture of the Multi-Task (MT) model	46
4.1	Comparison of ST and MT for aggregated annotator, including all 12 configurations	59
4.2	COGNIMUSE: Distributions of negative and positive examples with respect to each individual and aggregated annotations.	61
4.3	Distribution of positive and negative examples for each movie w.r.t each individual annotator	61
4.4	Pair-wise inter-annotator Cohen’s kappa of COGNIMUSE dataset. . .	62
4.5	Qualitative analysis of Multi-Task (MT) learning for COGNIMUSE dataset	65

4.6	IEMOCAP: Represents the number of samples in each category annotated by each annotator also with aggregated annotations. . . .	67
4.7	Pair-wise inter-annotator Cohen’s kappa of IEMOCAP dataset. . . .	68
4.8	SemEval_2007: It represents the number of samples in each category annotated by each annotator also with aggregated annotations. . . .	72
4.9	Pair-wise inter-annotator Cohen’s kappa of SemEval_2007 dataset. . .	73
7.1	Demonstrating the problem of predicting the emotion category of an upcoming utterance	88
7.2	An example in DailyDialog dataset	92
7.3	Twitter conversation in MOJITALK dataset	92
7.4	Tweets with emotion hashtags in CBET dataset	93
7.5	Examples from EMPATHETICDIALOGUES dataset	93
7.6	Profile information of a speaker in PERSONA-CHAT	94
7.7	Tweet annotated with an emotional category in SemEval_2018 dataset	94
7.8	Triple example in PELD	95
7.9	Personality-based Single-Task (ST) Learning	96
7.10	Personality-based Multi-Task (MT) Learning	98
7.11	Data distribution w.r.t each speaker in PELD dataset	103
8.1	Audio-based CNN for big five personality prediction using CS and SS	132
8.2	Samples of First Impressions dataset	133
8.3	Snapshots of the interface for labeling videos	133
8.4	Discriminative frequency patterns of a personality trait	136
8.5	Interpretable CNN	137
8.6	CAM of extraversion trait in 2D for a video	139
8.7	CAM of extraversion trait in the frequency domain for a video . . .	139
8.8	CAM of big five personality traits in 2D	140
8.9	CAM of big five personality traits in frequency domain	141
8.10	Interpretable CNN for raw audio as input	141
8.11	Frequency-based class activation maps of each personality traits . .	143

List of Tables

2.1	Detail description of some frequently used affective datasets	26
4.1	Synthetic data configuration	56
4.2	Cross-validation folds synthetic data	57
4.3	ST and MT results synthetic data Config_1	57
4.4	ST and MT results synthetic data Config_2	58
4.5	ST and MT results synthetic data Config_11	58
4.6	ST and MT results synthetic data Config_12	58
4.7	ST vs. MT comparison on the COGNIMUSE dataset. Results of ST and MT when modeling single and aggregated annotators with cross-validation evaluation.	64
4.8	MT comparison with SOTA on the COGNIMUSE dataset	64
4.9	Result comparison of ST, MT with SOTA for IEMOCAP dataset	69
4.10	The number of training samples per emotion category with respect to each individual and aggregated annotator of IEMOCAP dataset.	70
4.11	ST and MT Results per each emotion category using IEMOCAP.	70
4.12	Qualitative analysis of MT for IEMOCAP dataset	71
4.13	ST vs. MT comparison on the SemEval_2007 dataset. Mean accuracies obtained with cross-validation.	74
4.14	Comparison of ST and MT with SOTA	74
4.15	Qualitative analysis of MT learning for SemEval_2007 dataset	75
7.1	Detail distribution of datasets used for affective modeling in dialogues	91
7.3	Personalities traits of speakers in PELD dataset	100
7.4	Number of dialogues per each speaker in PELD dataset	102
7.5	7 emotions to 3 emotions categories conversion	105
7.6	Number of dialogues for the Train, Val, and Test set for 3 emotions categories with respect to each speaker	105
7.7	Emotional responses for the next utterance in 3 categories using Single-Task (ST) learning. F1-score is measured to predict negative, neutral, and positive emotion categories with respect to each speaker. The macro ($m - avg$) and weighted average ($w - avg$) are also measured.	106

7.8	Emotional responses for the next utterance in 3 categories using Multi-Task (MT) learning. F1-score is measured to predict negative, neutral, and positive emotion categories with respect to each speaker. The macro (<i>m - avg</i>) and weighted average (<i>w - avg</i>) are also measured.	106
7.9	Comparison of MT and ST approaches with SOTA for predicting emotional responses in 3 categories.	106
7.10	Predicting 7 categories emotional responses for the next utterance in a dialogue using Single-Task (ST) learning. F1-score is measured to predict each emotion category with respect to each speaker. . .	108
7.11	Predicting 7 categories emotional responses for the next utterance in a dialogue using Multi-Task (MT) learning. F1-score is measured to predict each emotion category with respect to each speaker. . .	108
7.12	Number of dialogues for the Train, Val, and Test for emotions with respect to each speaker	109
7.13	Comparison of MT and ST approaches with SOTA	110
7.14	Predicting 3 categories emotional responses without personality vector for the next utterance in a dialogue using Multi-Task (MT) learning. F1-score is measured to predict each emotion category with respect to each speaker.	111
7.15	Predicting 7 categories emotional responses without personality for the next utterance in a dialogue using Multi-Task (MT) learning. F1-score is measured to predict each emotion category with respect to each speaker.	111
7.16	Comparison of MT with and without considering personality information for predicting 3 categories emotional responses. F1-score is measured to predict each emotion category with respect to each speaker.	112
7.17	Comparison of MT with and without considering personality information for predicting 7 categories emotional responses. F1-score is measured to predict each emotion category with respect to each speaker.	112
7.18	Qualitative analysis of Personality based MT for 3-categories emotion prediction	113
7.19	Qualitative analysis of Personality based MT for 7-categories emotion prediction	114
7.20	Qualitative analysis of MT without personality for 3-categories emotion prediction	115
7.21	Qualitative analysis of MT without personality for 7-categories emotion prediction	116
7.22	Qualitative analysis of MT considering all outputs for 3-categories emotion prediction	118
7.23	Qualitative analysis of MT considering all outputs for 7-categories emotion prediction	119
7.24	Qualitative analysis of MT when replacing personality of speaker for 3-categories emotion prediction	120

7.25	Qualitative analysis of MT when replacing personality of speaker for 7-categories emotion prediction	121
8.1	Accuracies Evaluation with SOTA	134
8.2	Accuracies Evaluation with SOTA	138

Introduction

“Emotions change how we see the world and how we interpret the actions of others.”

Paul Ekman

Introduction

As human lives become increasingly assisted by technologies, there is a need to better understand the role of human emotions in Human-Computer Interactions (HCI). The advancement of artificial intelligence technologies (AI) evolved affective computing also known as emotional artificial intelligence. The concept of Affective Computing was introduced by Prof. Rosalind Picard of the Massachusetts Institute of Technology (MIT) in 1997 [1]. Affective Computing studies how to design machines with human-like capabilities of observation, interpretation, and generation of affective features. It is an important topic for harmonious Human-Computer Interaction; by increasing the quality of human-computer communication and improving the intelligence of the computer. The research on affect or emotion can be traced from nowadays to 19th century [2]. Traditionally, “affect” was seldom linked to lifeless machines, and was normally studied by psychologists. Researchers in the last decade have obtained dozens of scientific findings [3–8] illuminating the important roles of emotion in intelligent human functioning, even when it looks like a person is showing no emotion. These findings have reshaped scientific understanding of emotion and have inspired a number of researchers to consider that emotional mechanisms might be more valuable than previously believed. Consequently, a number of researchers have charged ahead with building machines that have several affective abilities, especially: recognizing, expressing, modeling, communicating, and responding to emotion [1, 9, 10]. And, within these areas, a number of new criticisms and challenges have arisen.

One of the main criticism faced by affective computing systems is *“People’s expression of emotions are variable and there is little hope of accurately recognizing an individual’s emotional state from the available data [11].”* Since emotions are highly subjective in nature and every person experiences emotions differently, which means the same instance can provoke different emotions in different individuals. To accurately recognize any individual’s emotions the affective systems should be aware of the emotional subjectivity. Psychologists already considered emotions as subjective responses/feelings [12]. In contrast, in affective computing, emotional subjectivity is rarely considered by researchers when designing and developing emotional recognition systems [13, 14].

Introduction

It is one of the core challenging problems faced by affective computing systems: how to model the subjective nature of experienced human emotions in the data.

A proper emotion model is the fundamental theoretical challenge for all affective computing studies. The emotion model is used to label the emotion experienced by a person. A common practice is to arbitrarily choose one from the mainstream emotion models proposed by psychologists. On the one hand, the categorical emotion model divides emotion into six basic emotions (anger, disgust, fear, joy, sadness, and surprise) [15, 16]. On the other hand, the dimensional emotion model, in which emotions are presented as a point or a region within a two-dimensional space (Arousal and Valence) [17] and the final classification is a vector, where each item corresponds to a specific emotion and each value corresponds to emotional intensity.

Researchers in affective computing use different modalities such as facial expression, body posture, voice, text, etc. to infer emotions. For example, Jung et al. [18] and Hasani et al. [19] classified face expressions into multi-class emotional categories. Xie et al. [20] and Wei et al. [21] classified speech signals into multi-class emotional categories. Noroozi et al. [22] and Sapiński et al. [23] recognized emotions using body movements. Zhong et al. [24] and Seyeditabari et al. [25] classified text data into multi-class emotion categories.

This thesis presents the research work conducted to develop affect machine learning systems that are aware of emotional subjectivity and recognize the affect perception of each individual in the available data. The study consist of the investigation of emotional subjectivity from two different perspectives:

(A) Emotion recognition systems are often trained on labeled datasets, which come from multiple human labelers. Since every single human has his own emotional perception which, as a result, produces emotional subjectivity in labels. The generalization capabilities of trained systems are highly dependent on the labeled data. This is why emotional subjectivity plays a critical role in designing emotion recognition systems. This thesis presents the previous studies on emotional subjectivity in annotations, their weaknesses, and the proposed solution

Introduction

which overcomes the existing shortcoming in the technology. These aspects are addressed in Part I of the thesis.

(B) Part II studies the relationship between emotional subjectivity and personality traits. The objective is to model affect expressions using a parametrization of personality traits as extra information. Concretely, in our experiments, we model emotional subjectivity in predicting automatic emotional responses to the next utterance in dialogues. For modeling subjective emotional perception, personality information along with the dialogue context is used and the results show that the addition of personality information boosts the generalization capabilities of each individual speaker as compared to the state-of-the-art.

Further in this Part, we presented research work on how to automatically recognize personality traits using human speech. The study developed an interpretable machine learning model that can visualize the discriminative frequency patterns in human speech based on the big five personality traits. The reason to design an interpretable model is that we wanted to understand the causation of personality traits with the model prediction. Furthermore, such types of visualizations give a deep understanding of the model's behavior and build trust in predictions with respect to the input data [26–32].

Part I

Incorporate Subjectivity Using Soft Labels

Chapter 1

Emotional Subjectivity in Affect Labeling

“All experience is subjective”

Gregory Beteson

Chapter 1: Emotional Subjectivity in Affect Labeling

This chapter introduces the role of emotional subjectivity in annotating affect labels, procedures for getting single labels from multiple affect labels, and the potential weakness of the affect recognition systems that are trained on single hard labels.

In supervised machine learning, the algorithm improves the generalization capabilities using the label and reiterating until the algorithm reaches a desired level of accuracy. In almost all machine learning algorithms, there is a cost function or objective function. The cost function is typically a measure of the error between the label and the algorithm prediction. By minimizing the cost function, we train our model to produce estimates that are close to the labels. This procedure is known as optimization. Minimization of the cost function is usually achieved using the gradient descent technique [33–36]. We have M training samples, and each one of them is labeled. The representation of the data is in pairs (x, y) , where (x) represents the input data and (y) represents the label. The input data (x) can be an N – *dimensional*, whereas each dimension corresponds to a feature or a variable. In a nutshell, supervised machine learning is basically learning a function that maps inputs to an output based on the labeled data. In fact, most real-world applications of machine learning are based on supervised machine learning [37]. In this work, we only targeted supervised machine learning. Therefore, the introduction of other machine-learning techniques is out of the scope of our study.

A large portion of the supervised machine learning research focuses on what is done once the ground truth/gold standard of the data is available, i.e. after the data labeling or annotation. Data annotation is the process of providing labels to raw data such as images, audio, video, text, etc. After data annotation, a machine learning model can learn from the annotated data, which is the collection of samples and their associated labels [38]. We particularly focus on human-labeled or human-annotated data, in which one or more individuals decide which category (Y) is associated with each item (X). Human annotation raised many issues, one important issue is called annotator bias. Geiger et al. [39] provide a review of the existing work that addresses the human-annotation process of the

data. Furthermore, the authors also discussed “*What is the best practice in a human annotation?*” They [39] summarize:

A “coding scheme” is defined as a set of labels, annotations, or codes that items in the corpus may have. Schemes include formal definitions or procedures, and often include examples, particularly for borderline cases. Next, coders are trained with the coding scheme, which typically involves interactive feedback. Training sometimes results in changes to the coding scheme, in which the first round becomes a pilot test. Then labelers independently review at least a portion of the same items throughout the entire process, with a calculation of “inter-rater reliability” (IRR) or “inter-annotator agreement” (IAA). Finally, there is a process of “reconciliation” for disagreements, which is sometimes by majority vote without discussion and other times discussion-based.

Most of the existing research work around human annotation is related to the reliability of the annotations [39–41]. The approaches that deal with annotation reliability basically assume that disagreement is a source of the noise. As mentioned in [39], the best practice to deal with disagreement is to get aggregated annotations. This is why very few approaches consider disagreement as a source of information (see Chapter 2). When we consider disagreement as a piece of information it means that we are considering each annotator’s subjective perception, which is also called annotator subjectivity. The reason behind considering each annotator’s subjective perception is that supervised machine learning handles multiple tasks. Some tasks are less subjective in nature but others are highly subjective. For example, giving an image and labeling it whether this image contains a cat or a dog is less subjective compared to giving an image and identifying whether an utterance contains positive or negative emotions.

1.1 Annotator Subjectivity

Figure 1.1 shows an example of labeling the emotion associated with a video clip where not all annotators agree on assigning the same emotion category. This disagreement is natural because each human has a subjective perception. This means, there is a possibility that significant variation exists in the worldview of annotators regarding any task.

Here, a question arises, i.e. how does the annotator’s disagreeing behavior affects the performance of machine learning models? Since in supervised machine learning the loss is computed between the predicted outcomes and the true labels, these true labels play a critical part in learning the mapping between inputs and outputs. Later, the models are used to anticipate unseen or future observations. This is why the generalization performance of the trained models is highly dependent upon the true labels. The common practice to get true labels is to use aggregated labels. These aggregated labels are always biased towards the majority group of annotators and hence lose the subjective information of each annotator. As a result, the trained models are also biased. Similarly, the generalization capabilities of the trained models are different for different groups, i.e. good performance for the majority group and bad for the subjective ones. Approximately all the tasks contain some level of subjectivity. Preserving the diversity of multiple interpretations (considering them as a source of information instead of a source of error or noise) could be an advantage in developing algorithms for subjective tasks.

Research Question

How could we incorporate multiple subjective perceptions in designing affect recognition systems?

In recent years, subjectivity has gotten significant attention, and researchers proposed different approaches to tackle subjectivity [40, 41]. Researchers assume subjectivity broadly in two categories: (i) Subjectivity as Noise and (ii) Subjectivity as Information. We discussed approaches that lie under both assumptions in

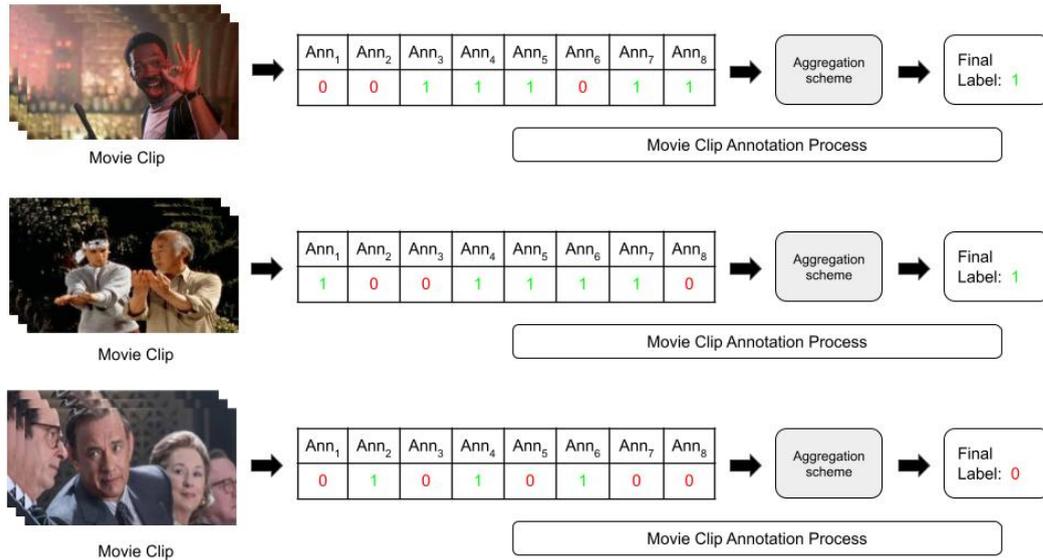


FIGURE 1.1: An example of labeling step: multiple annotators annotate each movie clip into Negative (0) and Positive (1) emotional categories.

detail (see Chapter 2). Subjectivity is highly task-dependent. For example, classifying an image into cat or dog is less subjective as compared to classifying an image into different emotional classes. Most of the research work was carried out by considering the subjectivity as noise and only targeting the tasks that are less subjective such as Breast malignancy [42], Quality Estimated of translated sentences [43], Image classification (cats and dogs [44] /multi-category like highways, streets, forests, etc. [45]), and human activity recognition (like walking, standing, etc.) [46]. In our research work, we considered subjectivity as information and targeted emotional subjective perception.

Research Objective

The main research objective is to develop a deep neural architecture that can consider subjective perception and has the following capabilities:

- The network can handle subjective as well as aggregated perceptions simultaneously.
- The network can generate subjective and aggregated predictions separately.

- The network should be able of handling multiple modalities such as Image, Text, Audio, etc.

Application Areas

There exist multiple application areas which could leverage the information from subjective annotations such as health (AI-based psychotherapy systems), education (affective-based e-learning systems), monitoring (affective analysis systems), marketing (affective-based recommender systems), and entertainment (affective media retrieval). In this section, we will review some of these areas where our proposed network could be useful to handle the subjectivity problem.

AI-based Psychotherapy Systems: These systems [47–49] provide psychological recommendations as psychological therapy to assist people whenever they face an immediate emotional conflict in their surrounding environment with the use of artificial intelligence. The recommendations provided by these systems are subjective, which means every single patient receives recommendations based on his/her cognitive and affective understanding.

Affective-based E-Learning Systems: The objectives of these systems [50–52] are to understand the emotional states (such as boredom, anger, anxiety, enjoyment, surprise, sadness, frustration, pride, hopefulness, hopelessness, shame, confusion, happiness, natural emotion, fear, joy, disgust, interest, relief, and excitement) of students during learning and provide the subjective responses to each individual student for more effective learning. For example, the frustrated emotional state of a student can either influence the learning rate or the dropout rate. In this scenario, the system will understand the emotional state and recommend subjective motivation strategies that are based on the student’s learning styles and cognition.

Affective Analysis Systems: With the fast expansion of social media applications, users have the privilege to access and broadcast their perception about any incident that happens anywhere around the world. The subjective perception may contain high to low emotional content that may be good for one group but

bad for another group of people [53, 54]. To understand the emotional perception of different groups of people we need a system that is capable to handle subjective information.

Affective-based Recommender Systems: As the advancement in machine learning is going, affective-based recommender systems were introduced in [55–57] that can use the affective state of an individual and recommend the content including news/posts/movies. For example, rather than recommend a movie based on previously watched movies, an affective recommender system will recommend a movie based on the emotional state of the viewer.

Affective Media Retrieval: On the internet, every day thousands of multimedia content has been uploaded. In order to properly utilize these data, retrieval systems are also constantly evolving. Currently, most of the retrieval systems use different modalities including text or image [58, 59]. However, the multimedia content also represents the affective intensity from high to low and may affect the emotional state of the user. Retrieving the best suitable content that matches the emotional state of the user could be a promising application. The application can adapt to the emotional state of the user while retrieving multimedia content.

Chapter 2

Literature Review

*“Your branches can only reach
high if your roots go deep”*

Brain Logue

This chapter first gives an introduction to the emotion representation models and then discusses the characteristics that datasets should have to incorporate the annotator's emotional subjectivity in machine learning systems. Later, we review the measurements that are used for calculating the agreement and disagreement between different annotators in affect labeling. Finally, we introduce previous state-of-the-art approaches that deal with annotators' emotional subjectivity.

2.1 Emotion Representation Models

Over the years, numerous representation schemes for emotions have been proposed, each one ranging from the most simplistic to the modern theories that refine and expand older models [60, 61]. The literature on this topic is too vast and it is out of the scope of this thesis to cover all these models. Cambria et al [62] presented a detailed overview of different emotion representation models and their use in affective computing. Overall, the two most popular representation models for identifying emotions are the Categorical and Dimensional models. In upcoming subsections, we will explain these two emotion representation models in detail.

2.1.1 Categorical Models

In categorical models, emotions are represented with one or more categories or labels (for example anger, happiness). The most popular categorical model is Ekman's model [63] of six basic emotions: anger, fear, surprise, joy or happiness, disgust, and sadness. The pictorial representation of six basic emotions is presented in Fig. 2.1. Another vastly used categorical model was proposed by Robert Plutchik [64] which provides eight basic emotion categories (anger, fear, sadness, disgust, surprise, anticipation, trust, and joy) as compared to 6 basic emotions.

Chapter 2: Literature Review

Plutchik's model is referred to as the wheel of emotions because it organizes these 8 basic emotions based on the physiological purpose of each. The model is actually the little "ice cream cone" which unfolds to the emotions wheel (see Fig. 2.2). Other categorical models [62] cover affects in general, which indicate emotion as part of affects such as WordNet Affect which was proposed by Strapparave and Valitutti [65]. It comprises more than 300 different affects, many of which are classified as emotions. This categorization provides a taxonomy of emotions as it gives information about the relationship between emotions and makes it possible to decide the level of granularity of the emotion expressed.



FIGURE 2.1: Ekman's model of six basic emotions [66].

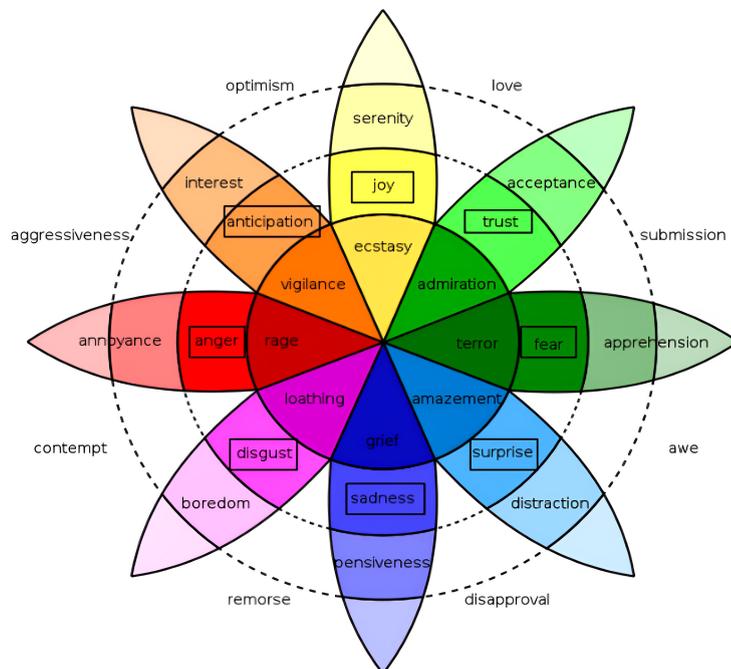


FIGURE 2.2: Plutchik's wheel of emotions [67].

2.1.2 Dimensional Models

On the dimensional side, emotions are represented in a set of dimensions that link the various emotional states. The first-dimensional model called PAD (Pleasure, Arousal, Dominance) or sometimes also called Valence-Arousal-Dominance (VAD) was proposed by Osgood et al. in 1957 [68]. In this model, Pleasure/-Valence represents whether the individual perceives the environment as enjoyable or not (i.e. positive or negative), while arousal represents the extent to which the environment stimulates the individual, and Dominance represents whether the individual feels in control or not in the environment. Fig 2.3 shows the graphical representation of the PAD model. Later, other researchers [69–72] considered that emotion or any other emotionally charged event are states experienced as simply feeling good or bad (i.e. pleasure/valence), active or passive (i.e. arousal). This is why they omit the dominance dimension and only work with pleasure and arousal for emotional representation. This model is also referred to 2D emotion representation model (see Fig. 2.4). More recently, Fontaine et al. [73] identified a fourth dimension, i.e. unpredictability. This new dimension refers to the appraisal of expectedness or familiarity.

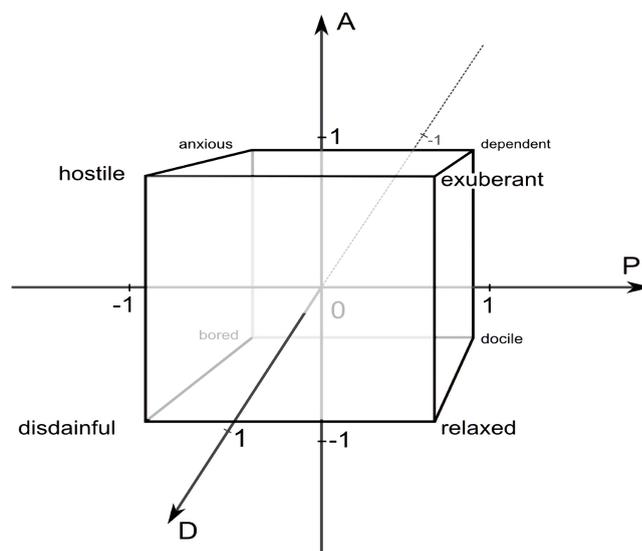


FIGURE 2.3: PAD (Pleasure, Arousal, Dominance) emotion representation model [66].

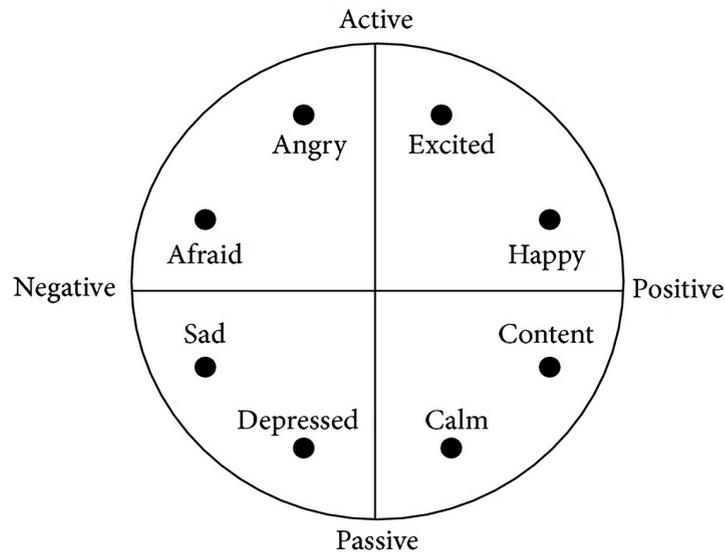


FIGURE 2.4: 2D (Pleasure and Arousal) emotion representation model [70].

2.2 Affect-related Datasets

In this section, we will take an overview of popular affect-related datasets. Most importantly, we studied the following aspects: *(i)* what type of data is available (either Single-Modality, such as image, audio, and text, or Multi-Modality), *(ii)* the number of annotators that were involved in the annotation process, *(iii)* the machine representation model used for the annotation, *(iv)* the process that was used to get the aggregated annotations, and *(v)* whether the subjective annotations are provided or not.

SEMAINE [74] is a multi-modality audiovisual interactive based dataset between humans and virtual agents. The interactive sessions were originally taken from TV chat shows. Often, in TV shows, hosts use a simple strategy: Invite guests to talk about topics that are emotionally significant for them and encourage (or provoke) them to express their emotion strongly, by inserting suitably chosen stock phrases at key points. The interactions involve two parties, a “user” (who is always human) and an “operator” (either a machine or a person simulating a machine). The main intention to develop this dataset is to build SAL (Sensitive Artificial Listener) agents. These SAL agents can engage humans in

Chapter 2: Literature Review

a controlled, emotional conversation. A total of 150 participants took part in recording 959 conversations. Each section is divided into clips and each clip lasts approximately 5 minutes. The clip was annotated by 6-8 raters into five affective dimensions, i.e. valence, anticipation, power, intensity, and activation. Unfortunately, the final annotations are aggregated such as average or majority voting of all annotations and the annotator-level annotations have not been released. This is why this dataset is not useful for subject-level emotional modeling. Fig. 2.5 shows an example where a clip is annotated into five dimensions.



FIGURE 2.5: SEMAINE: An interaction between a person and Sensitive Artificial Listener (SAL), annotated into five emotional dimensions. The presented annotations are aggregated (average) annotations from multiple annotators.

The **RECOLA** [75] dataset was created to understand the affective interaction between people when they are remotely connected, for example, during online meetings. The dataset consists of 3.9 hours of recording (audio-visual data) from 46 different individuals. The dataset was annotated by 6 different



FIGURE 2.6: RECOLA: An individual is doing an online meeting in a clip and each clip is annotated into Valence and Arousal dimensions. The annotations are aggregated annotations.

Chapter 2: Literature Review

annotators into two affective dimensions (Valence and Arousal). Later, a normalization has been done on annotations including local minimization for each annotation sequence for each annotator, and then used Zero-Mean (ZM) to get a single ground-truth label. This post-processing normalization technique loses the subjective perception of each annotator and hence the single ground truth annotations are not useful for modeling the subjective perception. Fig. 2.6 shows an example of a data sample annotated into arousal and valence dimensions.

HUMAINE [76] is a multimodality dataset including speech, language, gesture, face, and physiological aspects. The aim of the dataset is to understand the emotions that occur in our interactions and daily life actions. The dataset consists of 50 video clips, having 1.5 - 3 minutes in length. Every frame of a clip is annotated into intensity, arousal, valence, and dominance dimensions by 6 different annotators. Later, the percentage agreement between all 6 annotations is used to get a single ground truth for each frame. Again, in this dataset, the post-processing step to get the aggregated annotations lost the annotator-level perception, and the unavailability of the annotator-level annotations made this dataset not useful for subjective modeling. Fig. 2.7 shows the aggregated annotations of a video clip as an example.



FIGURE 2.7: HUMAINE: A video clip annotated into arousal, valence, dominance, and intensity dimensions. The given ground truth is the aggregated annotations.

FilmStim [77] is a movie dataset that has 70 movies with a length ranging from 1 to 7 minutes. The dataset was created to infer the emotional information

Chapter 2: Literature Review

that was induced in movie clips. It uses positive and negative scores, six discrete emotions (anger, disgust, sadness, fear, amusement, and tenderness), and 15 ‘mixed feelings’ scores assessing the effectiveness of each film excerpt to produce blends of specific emotions. A total of 364 participants annotated the clips. The dataset introduced the subject-level emotional analysis but later only released the normalized classification scores of each emotion category which lost the subject-level emotional information. Fig. 2.8 shows the aggregated ground truth of a movie clip into arousal, positive and negative affect, and the emotional category. The dataset does not provide annotator-level annotations.



FIGURE 2.8: FilmStim: A movie clip annotated into arousal, positive affect, negative affect, and a discrete emotional dimension.

The **LIRIS-ACCEDE** [78] is a multi-modality dataset and was a good attempt to create another dataset that has more number of movie clips with a diverse range of contextual emotions as compared to the FilmStim [77] dataset. The dataset contains 9,800 segmented video clips and each having a range from 8 to 12 seconds. These video clips are taken from 160 different movies that belong to different genres and languages. Each clip was annotated into 2D emotional dimensions: Arousal and Valence (see Fig. 2.9). A total of 1517 annotators from 85 different countries took part in the annotation process. Later, rating-by-comparison is used for getting gold standard labels. Each pair is displayed to annotators until the same answer has been given three times. The dataset is very comprehensive in terms of emotional content but due to the unavailability of annotator-level annotation, the dataset is not useful for modeling subjective emotional perception.



FIGURE 2.9: LIRIS-ACCEDE: A movie clip annotated into arousal and valence emotional dimensions. The annotations are aggregated annotations from multiple annotators.

MediaEval 2015 [79] This video dataset is an extension of the LIRIS-ACCEDE database. A total of 1,100 new video clips from 39 different movies had been added. For the annotation process, 16 annotators from different countries annotated each clip into valence (negative-neutral-positive) and a binary value that indicates the presence of violence. Lastly, the majority voting technique was used to get a single label for both categories (see Fig. 2.10). The dataset does not provide annotator-level annotations.



FIGURE 2.10: MediaEval 2015: A movie clip annotated into valence and a binary label for indicating the presence of violence in a clip. The annotations are aggregated annotations.

IEMOCAP [80] is an acted, multi-speaker dyadic dataset. The aim of this dataset is to understand the emotions in human conversations. The conversations are scenarios based which evoke different emotions in speakers. A total of 10 different actors (5 male, 5 female) took part in recording their facial motions, head movements, speech, and visual data. The actors played their roles in two

Chapter 2: Literature Review

different settings: scripted and spontaneous. The recordings were done in sessions and there are a total of 5 sessions in the dataset. After recording the sessions, the dialogues in each session were segmented into utterances. Each utterance was annotated into 9 emotion categories (anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state) and 3-dimensional labels (valence, arousal, dominance). Six annotators took part to classify the emotional content of utterances into discrete categories and the emotion dimensions. Each utterance was annotated by at least three annotators. On the other hand, for the continuous dimension, two annotators annotated the whole dataset. This dataset has a big advantage; the authors released the annotator-level annotations along with the aggregated annotations, which are obtained by majority voting. With the availability of annotator-level annotations, this dataset allows the study of modeling subjective emotional perception. Fig. 2.11 shows a single utterance and the annotations: Annotator-level and aggregated annotations.

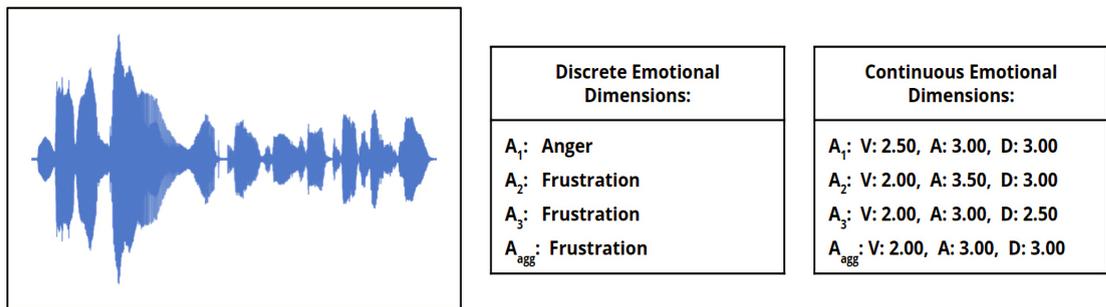


FIGURE 2.11: IEMOCAP: An utterance is annotated into Discrete and Continuous emotional categories. A_1 to A_3 represent the annotator’s ids, A_{agg} represents the aggregated annotation, and V, A, and D represent the Valence, Arousal, and Dominance emotional dimensions respectively.

The **COGNIMUSE** [81] is another dataset in the direction of understanding videos including movies and travel documentaries. In order to understand each movie scene, consecutive frames were annotated from the start of the movie. The dataset is generated for multiple tasks such as audio-visual and semantic saliency, audio-visual events and action detection, cross-media relations, and emotion recognition. The standard benchmark for emotional understanding [81] includes 7 Hollywood movies: “A Beautiful Mind” (BMI), “Chicago” (CHI), “Crash” (CRA),

Chapter 2: Literature Review

“Finding Nemo” (FNE), “Gladiator” (GLA), “The Departed” (DEP), and “Lord of the Rings: the Return of the King” (LOR). The total length of an annotated video is 30 minutes per movie. A total of 7 different annotators annotated each frame of a single movie in continuous values from -1 to +1 of arousal and valence domains. In terms of our research about subjectivity, the main advantage of this dataset over others is that this dataset also releases annotator-level annotations as compared to other datasets that only released aggregated annotations. These annotator-level annotations make the dataset suitable for subjective emotional modeling. Fig. 2.12 shows an annotated example.

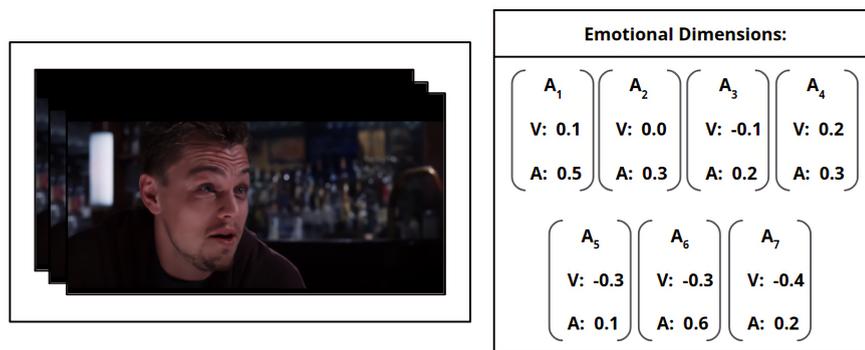


FIGURE 2.12: COGNIMUSE: A movie clip annotated by 7 different annotators into arousal and valence emotional dimensions. A_1 to A_7 represents the annotator’s ids. V and A represent the Valence and Arousal emotional dimensions respectively.

The **SemEval_2007** [82] is a text-based dataset that was developed to evaluate the participating systems in order to classify the emotions in news headlines. The dataset consists of 1,000 news headlines that were taken from different news channels including CNN, BBC and Google News, and the New York Times newspaper. The news headlines were annotated in 6 categorical emotions (Anger, Disgust, Fear, Joy, Sadness, Surprise) and 1 continuous dimension (Valence). The range of the categorical emotion was set between 0 to 100, where 0 represents No emotion and 100 represents the maximum intensity of the emotion category. On the other hand, the Valence dimension was set between -100 to 100, where -100 represents Very Negative, 0 represents Neutral, and 100 represents Very Positive

Chapter 2: Literature Review

(see Fig. 2.13). The dataset is annotated by 5 different annotators. Only aggregated annotations are provided with the dataset, which is getting by using the mean of all the 5 subjective annotations. Fortunately, the creator of the dataset agreed to provide us with annotator-level annotations for our experiments. We are thankful to the creator for this.

Books on science: The problems in modeling nature, with its unruly natural tendencies.	Discrete Emotional Dimensions:	Continuous Emotional Dimensions:
	A ₁ : Fear	A ₁ : Valence:20
	A ₂ : Surprise	A ₂ : Valence:40
	A ₃ : Surprise	A ₃ : Valence:-30
	A ₄ : Joy	A ₄ : Valence:50
	A ₅ : Surprise	A ₅ : Valence:-10

FIGURE 2.13: SemEval.2007: A news headline is annotated into discrete and continuous emotion categories by 5 different annotators. A_1 to A_5 represent annotator's ids.

GoEmotions [83] is a human-annotated dataset consisting of 58k Reddit tweets and annotated into 27 different emotion categories. The dataset was created to understand the textual conversation emotionally on social media platforms. The taxonomy of this dataset is different as compared to the big six emotions: it includes 12 positive emotions, 11 negatives, 4 ambiguous, and 1 neutral emotion. As compared to single positive and negative emotion classes, this taxonomy gives a better understanding of the emotion expressed in a text conversation. Each text in the dataset was annotated by 3 or 5 annotators and to get the aggregated annotations, the majority voting technique is used (see Fig. 2.14). The dataset does not provide annotator-level annotations. This means this dataset can not be used for subjective emotion modeling.

The **MovieGraph** [84] dataset was created to understand the movie scenes in more detail. It includes characters' attributes, actions, scene description, scene caption, etc. The dataset set consists of 7637 movie clips from 51 different movies. A graph is associated with each movie clip that captures the scene, situation label, interaction with time stamps, the relationship between different characters,

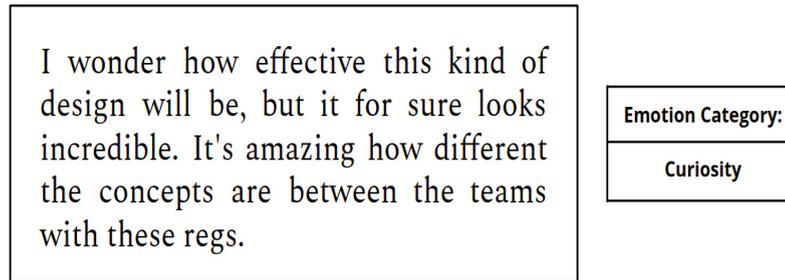


FIGURE 2.14: GoEmotions: A Reddit tweet is annotated in a single emotion category out of 27 different emotion categories by the majority voting on multiple annotations.

the emotion expressed by the characters, and a detailed description of the scene. A group of annotators was hired to annotate each graph and later, a majority voting technique was used to get the aggregated annotations for a graph. The dataset provides very rich annotations but due to the unavailability of annotator-level annotations, this dataset cannot be used to model subjective emotions. Fig. 2.15 shows a clip and its associated graph.

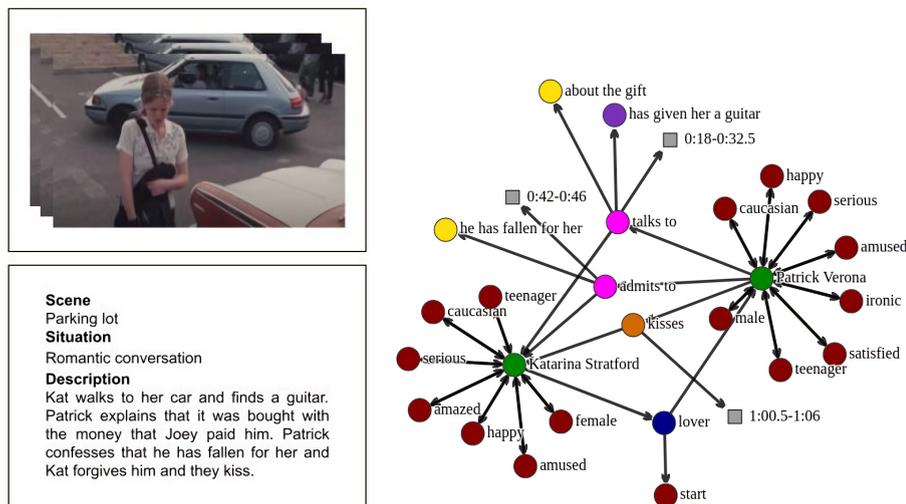


FIGURE 2.15: Left: A movie clip annotated into a scene, situation, and scene description in natural language. Right: An associated graph that captures the characters present in the clip, their interactions, the emotion expressed, and other attributes such as gender, action, timestamps, etc. [84]

The above literature presents the description of the most famous and commonly use affect-related datasets. Furthermore, the main characteristics are also presented in Table 2.1.

Chapter 2: Literature Review

TABLE 2.1: “Modality”: A means Audio/Speech, V means Visual/images, and T means Text/Verbal Sentences, “Single Annotation”: aggregated annotation, “Individual Annotation”: raw annotations provided by each separate annotator, “Annotation Details”: annotate emotions into which dimension: discrete or continuous or both.

Dataset	Size	Modality	Single Annotation	Individual Annotation	Annotation Details
SEMAINE [74]	959 characters interactions	AV	Yes	No	5 discrete emotional dimensions
RECOLA [75]	3.9 hours of recording	AV	Yes	No	2 continuous emotional dimensions
HUMAINE [76]	50 video clips	AV	Yes	No	4 continuous emotional dimensions
IEMOCAP [80]	45 hours of recording	AVT	Yes	Yes	9 discrete and 3 continuous emotional dimensions
MediaEval 2015 [79]	10,900 movie clips	AV	Yes	No	2 discrete emotional dimensions
LIRIS-ACCEDE [78]	9,800 movie clips	AV	Yes	No	2 continuous emotional dimensions
FilmStim [77]	70 movies	AV	Yes	No	3 continuous and 6 discrete emotional dimensions
COGNIMUSE [81]	7 movies	AVT	No	Yes	2 continuous emotional dimensions
MovieGraph [84]	51 movies	AVT	Yes	No	20 discrete emotional dimensions
SemEval_2007 [82]	1000 News headlines	T	Yes	Yes	6 discrete and 1 continuous emotional dimensions
GoEmotions [83]	58k tweets	T	Yes	No	28 discrete emotional dimensions

2.2.1 Subjective Annotations in Affect-related Datasets

As observed in the previous section, the most common practice is only to release the aggregated annotation [82, 83, 85–93]. Due to this, not all affect-related datasets can be considered for the study of subjective emotional modeling. Among the reviewed datasets, only COGNIMUSE [81], IEMOCAP [80], and SemEval_2007 [82] datasets provide the annotator-level annotations. With these annotator-level annotations, only these datasets can be used to study subjective annotations. For this reason, the experiments in Part I of this thesis are done with these three datasets.

2.3 Measuring Annotator Agreement

It is a common practice in the annotation process to compare annotations of a data sample (image, audio, text, etc.) by multiple annotators. This measurement may be useful for qualitative examine the annotations, the degree of

agreement between annotations, or the statistical modeling of annotator disagreements. Human annotation is basically an interpretation process [94] which is why it is very unlikely that the response of multiple annotators is identical. The common wisdom holds that the more annotators agree on a given annotation, the chances are higher the given annotation is correct. This is called raw agreement or observed agreement and it is still the most common way to represent the agreement [95]. This raw agreement is easy to measure and understand but it does not imply that the annotation process is reliable because some agreements may be accidental [96]. To quantify the agreement, researchers have proposed different agreement measures. We will discuss a few of them in this section.

2.3.1 Pearson's Correlation Coefficient

Pearson's Correlation Coefficient [97] is a standard measure of correlation. It measures the linear relationship between two random variables (in the annotation process, these are the two equal-size arrays that hold the annotated class of n samples from two different annotators). It is the ratio between the covariance of two variables and the product of their standard deviations (see Eq. 2.1). The coefficient (ρ) values range from -1 to +1, since the coefficient defines the relationship between the relative movements of two variables. This means a value equal to 0 indicates that there is no correlation between these two variables. A value greater than 0 indicates a positive correlation; that is, as the value of one variable increases, so does the value of the other variable, and a coefficient less than 0 indicates a negative correlation; that is, as the value of one variable increases, the value of the other variable decreases. Thus, the Pearson correlation can be used to quantify the linear relationship between the annotation provided by two annotators. The general equation to calculate Pearson's correlation coefficient is given:

$$\rho(Y_1, Y_2) = \frac{cov(Y_1, Y_2)}{\sigma(Y_1)\sigma(Y_2)} \quad (2.1)$$

Where,

$cov(Y_1, Y_2)$ = is the covariance of Y_1 and Y_2

$\sigma(Y_1)$ = is the standard deviation Y_1

$\sigma(Y_2)$ = is the standard deviation Y_2

In the COGNIMUSE dataset [81], the authors used Pearson’s correlation coefficient to show the inter-annotator statistics for the annotation of expressed emotions in given stimuli. The annotators annotated each stimulus into two emotion dimensions valence and arousal. The inter-annotator agreement for the valence is 0.293 and for the arousal emotion, it is 0.409.

2.3.2 Spearman’s Correlation Coefficient

The Spearman’s Correlation Coefficient [98] is a non-parametric measure of rank correlation. It measures the strength and direction of association between two ranked annotations. A rank is a scalar value assigned to every single item of annotations. It indicates the position of the item in the annotation. For example, order the values of annotations from greatest to smallest; assign the rank 1 to the highest score, 2 to the next highest, and so on. The Spearman’s correlation is equal to Pearson’s correlation of the two ranked annotations (See Eq. 2.2). Where the Pearson correlation represents the linear relationship, Spearman’s correlation represents the monotonic relationship (whether linear or not). The interpretation is similar to that of Pearson: 0 indicates there is no correlation, values closer to +1 indicate a very positive correlation and values closer to -1 means a very negative correlation. This measure is appropriate for continuous and discrete ordinal categories and is only limited to two variables, meaning that it can be used to quantify the relation of annotations provided by two annotators. However, one major drawback is their sensitivity to prolonged errors in ranking. The coefficient is represented by r_s and the equation is:

$$r_s = \frac{cov(R(Y_1), R(Y_2))}{\sigma_R(Y_1)\sigma_R(Y_2)} \quad (2.2)$$

Where,

$R(Y_1)$	= rank annotation from one annotator
$R(Y_2)$	= rank annotation from another annotator
$cov(R(Y_1), R(Y_2))$	= is the covariance of two rank annotations
$\sigma_R(Y_1)$	= standard deviation of $R(Y_1)$
$\sigma_R(Y_2)$	= standard deviation of $R(Y_2)$

Zhang et al. [99] proposed a human-machine annotation framework for getting more reliable single-ground truth annotations. They proposed a Dynamic Cooperative Learning (DCL) algorithm to decide which instance can be automatically labeled by the machine and which one needs human inspection. During the human inspection, multiple annotators annotate a single instance and used Spearman's correlation coefficient to get the level of agreement. DCL algorithm enables early stopping if Spearman's correlation coefficient level has reached a certain level.

2.3.3 Cohen's Kappa

Cohen's Kappa [100] is also often used to know the degree of inter-annotator agreement between two annotators quantitatively when they classify N samples into C mutually exclusive classes. Rather than just computing the percentage of items that the raters agree on, Cohen's Kappa accounts for the fact that the raters may happen to agree on some items purely by chance (See Eq. 2.3). The value of kappa (κ) indicates the agreement between two annotators: If the value of kappa is less than 0 means there is no agreement between raters, a value of 0 means there is an agreement equivalent to random chance, and the values greater than 0 represent the different levels of agreement. For example, a kappa value of 0.1 represents a slight agreement, whereas a kappa value of 0.5 represents a moderate agreement, and a kappa value of 1.0 represents the perfect agreement between two raters. The kappa is represented by a letter κ and is calculated as follows:

$$\kappa = \frac{P_{observed} - P_{expected}}{1 - P_{expected}} \quad (2.3)$$

Where,

$P_{observed}$ = number of agreements among annotators

$P_{expected}$ = number of agreements among annotators due to chance

In [101], the authors experimented with polysemy judgment of multiple words for the evaluation of the WSD (Word Sense Disambiguation) systems. For the experiments, the author first asked multiple annotators for POS (Part-of-Speech) tagging. After the annotation process, the author used Cohen's kappa to get the agreement level between different annotators.

2.3.4 Krippendorff's Alpha

Krippendorff's Alpha [102] is an alternative to Cohen's kappa for determining inter-annotator agreement. Compared to Cohen's kappa, it can handle multiple annotators, not only two. Krippendorff's Alpha (α) calculates the disagreement of the annotators instead of the agreement (see Eq. 2.4). It applies to multiple categories and different levels of measures including nominal, ordinal, interval, etc. It can also handle missing data. The alpha ranges are from 0 to 1, where 0 is perfect disagreement and 1 is perfect agreement. The following equation is used to calculate the alpha (α):

$$\alpha = 1 - \frac{D_o}{D_e} \quad (2.4)$$

Where,

D_o = number of disagreements among annotators

D_e = number of disagreements among annotators due to chance

Kralj et al. [103] proposed lexicons-based emoticons to analyze the effect of emojis in online social media conversation. In this paper, the authors used Krippendorff's Alpha to know the inter-annotator agreement between two annotators. The disagreement between these two annotators represents subjectivity in general.

2.3.5 Scott's Pi

Scott's Pi [104] is another statistical measure that can be used to quantify the inter-annotator agreement. It works by comparing the amount of agreement observed between the two annotators with how much agreement is expected if both annotators chose randomly (see Eq. 2.5). This is very similar to Cohen's kappa. The only difference is the way to calculate the chance agreement. Cohen's kappa uses geometric means whereas Scott's Pi uses the squared arithmetic means of the marginal proportions, which makes it less informative as compared to Cohen's kappa. This measure is best suited for nominal data with two annotators. Scott's Pi ranges between 0 and 1, with 1 indicating perfect agreement. The formula to calculate the π value is:

$$\pi = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (2.5)$$

Where,

$Pr(a)$ = agreement that was observed between the two annotators

$Pr(e)$ = agreement that is expected between the two annotators

Grouin et al. [105] proposed the guidelines for the name entity tagging for annotators. They proposed a tagging hierarchy in which the annotators will take advantage of a global corpus as well as a mini-reference corpus. In this paper, the authors used Scott's Pi to indicate the impact of proposed guidelines in the name entity annotation process. They got a high inter-annotator agreement using the proposed guidelines.

2.3.6 Gwet AC1

In 2001, Gwet [106] proposed a new agreement coefficient for inter-rater reliability (see Eq. 2.6). As compared to Cohen's kappa, it assumes that the agreement between observers is not totally at random. The Gwet adjusted chance

agreement by a conditional probability that two randomly selected raters will agree. It can handle multiple annotators and is more appropriate when an ordered categorical rating system is used. The range of the AC1 is from -1 to 1, for complete disagreement and perfect agreement, respectively. The equation to calculate AC1 is:

$$AC1 = \frac{P(a) - P(e\gamma)}{1 - P(e\gamma)} \quad (2.6)$$

Where,

$P(a)$ = overall percent agreement

$P(e\gamma)$ = probability of chance agreement

Hoek et al. [107] used CCR (Cognitive approach to Coherence Relations) to annotate the implicit and explicit relations in linguistics. The authors compare the inter-annotator agreement of two annotators using the Gwet AC1 measure and the values show that the agreement is low for implicit relation and high for explicit relation.

2.3.7 Fleiss Kappa

Fleiss Kappa is another variant of Cohen's kappa to measure the inter-annotator agreement (See Eq. 2.7). In contrast with Cohen's kappa, it works for any constant number of annotators giving categorical annotations to a fixed number of data samples. It is important to note that whereas other measures assume the same annotators annotated a set of data samples, Fleiss's kappa allows the annotators to annotate different subsets of the data samples. For example data sample A can be annotated by annotators 1, 2, and 3, and maybe the data sample B is annotated by annotators 5, 4, and 7. It is used to measure the inter-annotator agreement of the whole dataset. Kappa value ≤ 0 means no agreement (or agreement that you would expect to find by chance) and 1 means a perfect agreement. The equation is:

$$\kappa = \frac{\bar{P} - \bar{P}e}{1 - \bar{P}e} \quad (2.7)$$

Where,

$\bar{P} - \bar{P}e$ = the degree of agreement actually achieved above chance

$1 - \bar{P}e$ = the degree of agreement that is attainable above chance

Baveya et al. [78] created a movie dataset named LIRIS-ACCEDE for emotion understanding. The dataset consists of movie clips that are derived from multiple movies and each clip is annotated into valence and Arousal dimensions. More than 1,000 annotators took part in annotating 9,800 movie clips. The authors used Fleiss's kappa value to show the inter-annotator agreement for representing the reliability of the annotations.

2.3.8 Rosenberg and Binkowski Kappa

Rosenberg and Binkowski [108] proposed a new variant of Cohen's kappa for inter-annotator agreement. In Cohen's kappa, the inter-annotator agreement is measured based on a single label assigned to each data sample. In contrast, Rosenberg and Binkowski's kappa measures the inter-annotator agreement when there are two labels assigned to each data sample. One is called the primary label and the other is called the secondary label. Basically, when an annotator selected a single label, that single label has a weight equal to 1.0. In primary and secondary labeling, a weight (ρ) is assigned to a primary label and $(1 - \rho)$ is assigned to a secondary label. In a typical setting of Cohen's kappa (see Eq. 2.3), they used relative frequencies of each annotator's labeling preference instead of assuming an even distribution of labels to calculate the expected probability of agreement. The range of kappa ranges from 0 to 1, where 0 means no agreement and 1 means strong agreement. The equation for the measure is:

$$K' = \frac{P(A) - P(E)}{1 - P(E)} \quad (2.8)$$

Where,

$P(A)$ = observed probability between two annotators

$P(E)$ = expected probability between two annotators

Ignatova et al. [109] proposed an annotation scheme for online Question and Answering. The proposed scheme has additional attributes to identify unclear and opinion questions. The authors used the inter-annotator agreement measure as proposed in [108] and showed that the proposed annotation scheme has got a high inter-annotator agreement.

2.4 Approaching Agreement/Disagreement from a Machine Learning Perspective

In the literature, there are just a few relevant works that attempt to highlight the weakness of conventional aggregated annotation-based approaches for addressing the inter-annotator affective perception. As discussed in Section 2.2, one of the reasons why researchers focus on the aggregated annotations in subjective tasks is because most of the public datasets just provide the aggregated annotations [110–112].

In general, researchers consider the subjective behavior of annotators in two ways: (i) subjectivity as noise, where researchers assume that the annotator's subjective behavior is noise, and (ii) subjectivity as information, where researchers consider that all the available annotations associated with a data sample are correct. On the one hand, the main objective of all the developed approaches under the assumption of subjectivity as noise is to validate each annotation associated with a data sample. The validation process is based on some defined measures. If the annotation satisfies this measure then it is considered a true annotation; if not then the next step is to correct or remove this annotation. The majority of the research work developed that considers annotators' subjective behavior is under

this assumption. In contrast, the approaches developed under the assumption of subjectivity as information take advantage of all the annotations (also known as soft labels) during processing, and the output is based on all these annotations. In the work presented in this thesis, we followed this assumption, where all the available annotations for a data sample are correct and represent the subjective behavior of annotators. Our proposed model has the capability to process multiple annotations for a single data sample. The rest of this section reviews the literature on both considerations, i.e. subjectivity as noise and subjectivity as information, regardless of their underlying task.

2.4.1 Subjectivity as Noise

Raykar et al. [113] proposed an approach that uses multiple annotations in their experiments. The main objective is to find the true gold annotation using multiple noisy annotations. The authors assume that the disagreeing annotations hold the noise. The authors use prior information for each annotator to capture the skills such as expert or novice. The work is related to finding the quality of the data annotation. This work assumes that there is a single gold standard annotation that exists behind multiple noisy annotations.

Another technique in the series of findings on the reliability of the data annotations is presented by Yan et al. [42]. The previously mentioned approach assumed that the annotations are dependent on the annotator's expertise. Instead, in this paper, the authors assume that the expertise level is dependent on the data sample that each annotator observes. With this assumption, the authors model the error rate of each annotator as dependent on the data sample. The proposed algorithm produces a single outcome for each data sample.

Morales-Álvarez et al. [114] proposed a Gaussian Process based approach that handles multiple annotations without knowing their expertise level. The objective of this approach is similar to what we discussed in previously presented approaches, i.e. use multiple annotations to get a single true annotation. After

fully converging, the proposed model gives less importance to those annotators who produce noisy annotations.

Cohn et al. [115] proposed a Gaussian Process base multitask learning to model the subjectivity of different annotators based on their level of expertise and reliability. This approach set a prior for each annotator which is correlated to other annotators and all annotators are dependent on each other.

Rodrigues et al. work [116] also considers the disagreeing behavior of the annotators as a noise, i.e. each annotator has a different level of expertise. The authors proposed a crowd layer on top of the output layer called the bottleneck layer. This crowd layer has multiple outputs, where each output belongs to each individual annotator. The crowd layer learns the annotators' behavior and adjusts the annotator bias according to the labels assigned to each annotator. The adjusted gradient from each annotator is then passed to the previous bottleneck layer. This layer aggregates the multiple gradients that are coming from N annotators and backpropagates down to the network. This crowd layer is only used for the training time, i.e. once the network is fully trained then it is removed. Therefore, in inference time, it is not clear that the predicted output belongs to each individual annotator.

In summary, most of the studies that consider subjectivity as noise have been done in the domain of crowdsourcing, where we have hundreds or thousands of annotators for annotating a single data sample. Researchers use subjective labels to find the reliability of the annotators and based on the reliability score they give weight to a particular annotation in order to find the underlying gold label. The proposed approaches were evaluated on tasks that are less subjective in nature than affect-related tasks, such as Breast malignancy [42], Quality Estimated of translated sentences [43], Image classification (cats and dogs [44] /multi-category like highways, streets, forests, etc. [45]), and human activity recognition (like walking, standing, etc.) [46]. To our best knowledge, there is no publicly available affect dataset that has annotations taken from hundred or thousand of annotators per data sample.

2.4.2 Subjectivity as Information

There are few works that consider subjectivity as information. For example, Fornaciari et al. [117] proposed a multi-task model that learns from multiple annotations. The multi-task model is built on top of the single-task, where the single-task predicts the standard single output, and this single output is treated as the distribution of labels. Then an additional task is added on top of it called an auxiliary task. The aim of this additional task is to predict the soft label distributions with respect to each annotator. The training of both tasks is done in a joint manner. Two different losses are computed: one for the main task (predicting gold labels) and one for the auxiliary task (predicting soft labels). The proposed multi-task was evaluated on two different tasks: POS (Part-Of-Speech) tagging and morphological stemming.

Fayek et al. [118] proposed an algorithm to address emotion subjectivity. The approach is based on the training of multiple Deep Neural Networks (DNNs). The multiple DNNs were trained against two different types of labels: (i) hard labels, where the authors treated each individual annotator's annotations as hard labels, and (ii) soft labels, where the authors used an encoding scheme proposed in [119] to get soft labels from each annotator's hard labels. For the hard labels, a separate model was trained for each individual annotation. On the other hand, only a single DNN is trained for the soft labels. Later, the proposed approach used two different techniques to combine the outputs of all DNNs trained on hard and soft labels: the Geometric Mean and the Unweighted Majority. The model was tested on the IEMOCAP dataset. The results showed that introducing subjectivity in the model using soft labels increases the model performance for emotion recognition.

To address emotional subjectivity, Chou et al. [120] proposed an approach that uses the aggregated annotations (hard label) and the distribution of annotations (soft label) simultaneously in a joint manner. The learning consists of multiple models, i.e. 5 models for each individual annotator, and a combination of 2 previous models ([121],[122]). Later they concatenated all the outputs and used

Softmax for the final predictions. The approach is based on the concatenation of two different types of subjective perception: *(i)* modeling subjective perception using original soft labels, i.e labels that are annotated by each individual annotator, and *(ii)* generating soft labels using the approach proposed by Ando et al. [121] and then modeling the subjective perception. The authors did not provide clarity about what type of soft labels is considered to address subjectivity as a general. For example, original soft labels that are directly coming from annotators or soft labels that are getting from any encoding scheme or a combination of both. This is why it is not fair to consider this approach as a general approach. Another limitation is the evaluation protocol in which only a single dataset is used with the consideration of a single modality (Audio). This approach has achieved state-of-the-art results in modeling emotions for individual and aggregated annotations using the IEMOCAP dataset.

Chapter 3

A Multi-Task (MT) Learning Approach to Model Subjectivity in Affect-related Tasks

*“I believe everyone should be
treated as an individual”*

Doug Stanhope

In this chapter, we introduced our proposed Single-Task (ST) and Multi-Task (MT) learning approaches to annotators' emotional subjectivity. The chapter explains the building blocks of each individual approach as well as the working of both approaches in detail.

The proposed approach is inspired by the following scenario: Given an input sample, there are multiple annotators $\{A_1, \dots, A_N\}$ which provide a label for the input sample according to an affect-related task. As previously discussed, the affect-related tasks are highly subjective, which means that the annotations provided by multiple annotators are not identical. Each annotation represents the annotator's affective response to a given input sample. These affective responses are the subjective perception of multiple annotators. In order to analyze how each subjective perception can be leveraged by the aggregated perception, we compare two types of CNN architectures: (i) a Single-Task (ST) architecture, where we modeled each subjective perception independently, and (ii) a Multi-Task (MT) architecture, where all subjective perception along with the aggregated perception are jointly modeled.

Our Hypothesis

The common patterns between all individuals and the aggregated annotations help to improve the generalization capabilities of individual and aggregated annotator perception.

3.1 An Introduction to Multi-Task (MT) Learning

Multi-Task learning was first proposed by Rich Caruana in 1997 [123]. The goal is to improve generalization capabilities by leveraging the domain-specific information contained in the training data of related tasks. It is basically a machine-learning approach in which the objective is to learn multiple tasks simultaneously,

Chapter 3: Multi-Task (MT) Learning Approach

optimizing multiple loss functions at once. Rather than training independent models for each task, multi-task allows a single model to learn all tasks at once. In this process, the model uses all available data across the different tasks to learn generalized representations of the data that are useful in multiple contexts. For example, when training a model on some task A , our aim is to learn a good representation of the task A that ideally ignores the data-dependent noise and generalizes well. As different tasks have different noise patterns, a model that learns two tasks simultaneously is able to learn a more general representation. Learning only task A bears the risk of overfitting related to task A , while learning A and B jointly enables the model to obtain a better representation F by averaging the noise patterns.

In the context of deep learning, multi-task learning is typically done with either soft or hard parameter sharing of hidden layers. In soft parameter sharing, each task has its own model with its own parameters. The distance between the parameters of the model is then regularized in order to encourage the parameters to be similar. For example, Duong et al. [124] used the l_2 norm for regularization, while Yang et al. [125] used the trace norm. Fig. 3.1 represents the graphical structure of soft parameter sharing multi-task learning.

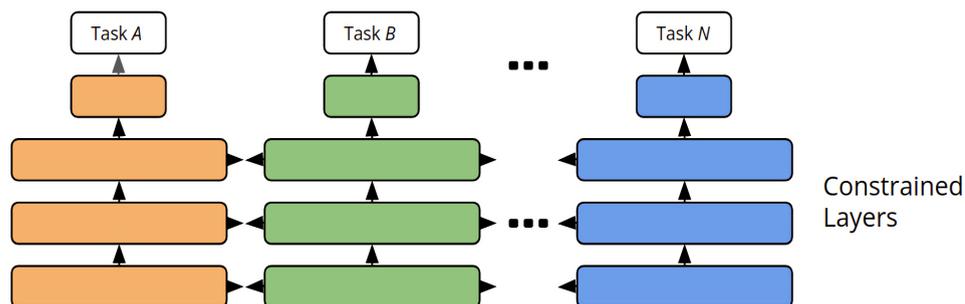


FIGURE 3.1: Soft parameter sharing for multi-task learning in deep neural networks.

On the other hand, hard parameter sharing is the most commonly used approach to multi-task learning in neural networks. Hard parameter sharing greatly reduces the risk of overfitting. In fact, Jonathan Baxter [126] showed that the risk of overfitting the shared parameters is an order N . That means the more tasks we

Chapter 3: Multi-Task (MT) Learning Approach

are learning simultaneously, the more our model has to find a representation that captures all the tasks and the less chance of overfitting on our original task. Fig. 3.2 represents the graphical structure of hard parameter sharing multi-task learning. In this thesis, we followed hard parameters sharing multi-task learning.

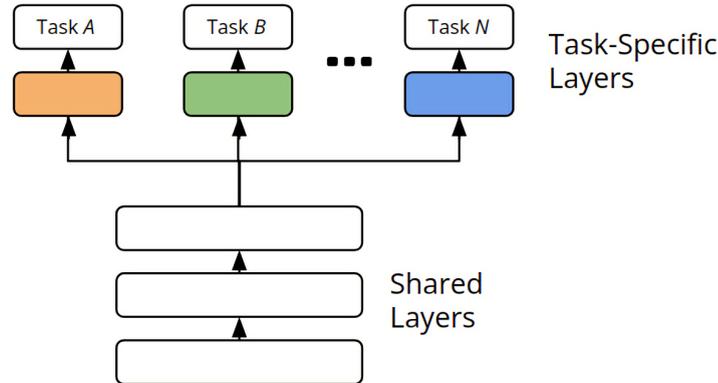


FIGURE 3.2: Hard parameter sharing for multi-task learning in deep neural networks.

Taylor et al. [127] proposed multi-task learning (MTL) techniques to train personalized machine learning models which are customized to the needs of each individual, but still leverage data from across the population. Three different variations of multi-task learning were proposed and compared: i) MTL deep neural networks, which share several hidden layers but have final layers unique to each task; ii) Multi-task Multi-Kernel learning, which feeds information across tasks through kernel weights on feature types; and iii) a Hierarchical Bayesian model in which tasks share a common Dirichlet Process prior. These techniques were investigated in the context of predicting future mood, stress, and health using data collected from surveys, wearable sensors, smartphone logs, and the weather.

Similarly, Lopez-Martinez and Picard [128] proposed a pain intensity measurement method based on physiological signals. Specifically, a multi-task learning approach based on neural networks that account for individual differences in pain responses while still leveraging data from across the population. The authors used multi-modal data (skin conductance and heart rate) for personalized nociceptive pain recognition in healthy subjects.

The mentioned research works used multi-task learning to model subjectivity and the data belongs to just a single study. There is no further exploration of this idea in the context of subjective annotations, which is the contribution of our research work.

3.2 Modeling Subjective Annotations with Multi-Task (MT) Learning

3.2.1 Single-Task (ST) Architecture

The Single-Task (ST) architecture is a traditional supervised deep learning model, i.e. a model that is trained, validated, and tested on X_{train} , X_{val} , and X_{test} data sets respectively. Typically in supervised learning, it is usually assumed that only a single label is considered for each input data but in our scenario, we are considering that multiple labels are available for each input data. Therefore, instead of using the classic notation (x_i, y_i) , we will use the notation $(x_i, y_{i,j})$ to also refer to one of the multiple labels available for x_i , specifically the one provided by the annotator A_j . The ST architecture is illustrated in Fig. 3.3. The architecture consists of three different blocks:

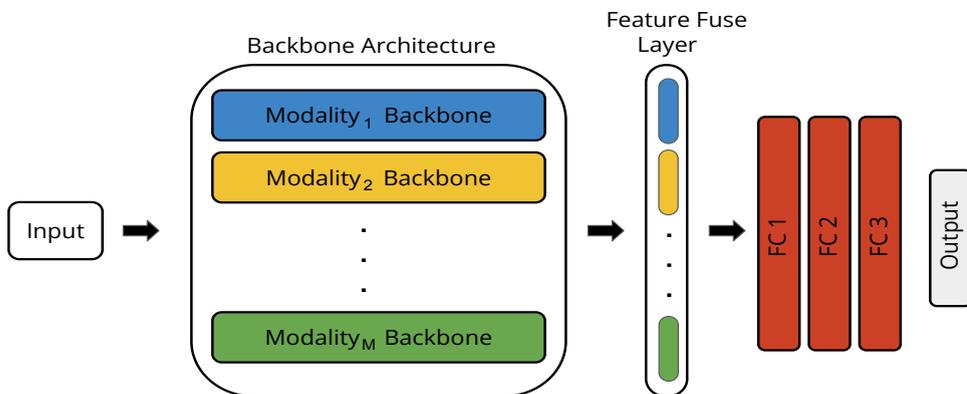


FIGURE 3.3: Architecture of the Single-Task (ST) model. It consists of feature extractors for each modality, a feature fusion layer, and a fully-connected block.

Backbone Architecture: The backbone architecture consists of as many branches as the number of modalities of the input data which will be considered for the modeling of affective perception. Examples of modalities are visual modality, text modality, and audio modality. Each branch represents the feature extraction for a particular modality (see Eq. 3.1). We are free to use any available on-the-shelf algorithm or to develop our own.

$$F_m = O_m(I) \quad (3.1)$$

Where,

I = Input sample

$O_m(I)$ = Feature extractor of modality m

Feature Fusion: Once obtained the features of each modality from the backbone architecture, the next step is to convert these multi-modality data into single multi-modality data. For this purpose, we concatenate all the features of multiple modalities into one big tensor (see Eq. 3.2). This multi-modal feature vector is ready to feed for the fully-connected layers.

$$F_c = [F_1, F_2, \dots, F_M] \quad (3.2)$$

Where,

F_c = Single dimensional feature vector

F_M = Feature of each modality

Fully-Connected Block: The multi-modality data is then fed into the block of fully-connected layers. Finally, the loss function used in the architecture would depend on the inference task (e.g. classification, regression, etc.). Each hidden layer has a following equation:

$$Layer^{in} = F_c * W + Bias$$

$$Layer^{out} = ReLU(Layer^{in})$$

Where,

F_c = Single dimensional feature vector

W = Weights of the neuron

$Layer^{in}$ = output of the multiplication

$ReLU$ = applying non-linearity on the outputs

$Layer^{out}$ = output of the fully-connected layer

3.2.2 Multi-Task (MT) Architecture

Unlike the Single-Task (ST), where ST always predicted a single output (y_i) per input sample (x_i), the Multi-Task (MT) architecture has capabilities to produce multiple outputs (y_1, y_2, \dots, y_N) per input sample (x_i), where N is the total number of annotators. Each single output in Multi-Task (MT) architecture refers to an individual annotator. The MT architecture is illustrated in Fig. 3.4. The MT architecture consists of the following blocks:

- Backbone architecture that consists of feature extractor for each modality
- Feature fusion blocks behave in a similar manner as in Single-task (ST) architecture.
- After the feature fusion block, we introduced an addition block called “Shared fully-connected”. It consists of two fully-connected layers. These layers allow the network to learn the common patterns found in the data.
- After that, a separate fully-connected block is considered for each annotator to learn their specific perception. Each separate block consists of 3 fully connected layers. Each block has its own loss function and optimizer which models each subjective perception.

Another advantage of Multi-Task architecture is that the model just does not learn the subjective perception of each annotator. It has a separate block for the aggregate annotator. Therefore, along with specific patterns for each subjective perception the model also learns the aggregated perception.

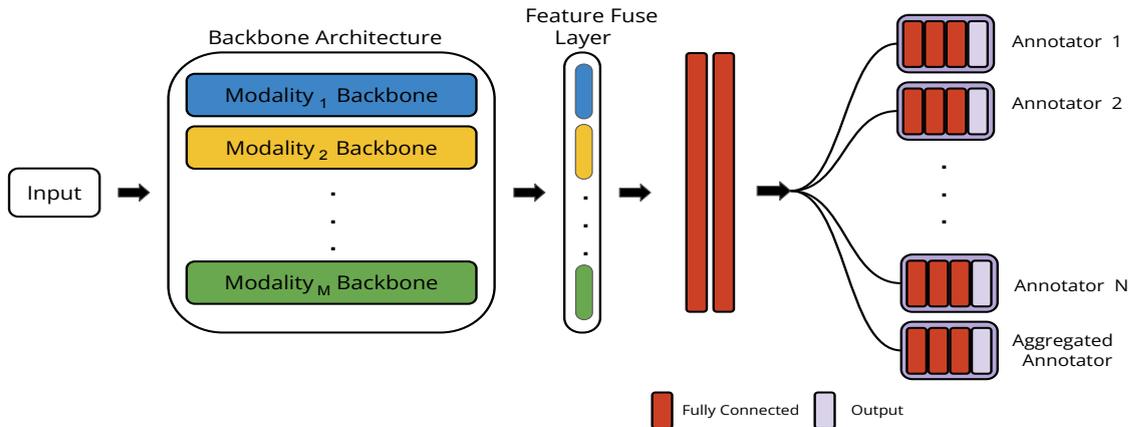


FIGURE 3.4: Architecture of the Multi-Task (MT) model. This architecture is applied to all annotators in a joint manner by sharing a common two fully-connected layers and a dedicated fully-connected block for each annotator.

3.3 The MT Approach in the Context of Related Work

In this section, we compare our proposed Multi-Task (MT) approach with previous state-of-the-art approaches which considered subjectivity as information (see Section 2.4.2).

Cohn et al. [115]: They addressed the annotators' bias and proposed a multi-task learning technique to learn annotators' specific behavior. The authors targeted the problem of predicting the quality of sentence translation using datasets that have been annotated by several judges with different levels of expertise and reliability. The proposed multi-task approach is based on Gaussian Processes (GPs) [129], which is a kernelized Bayesian non-parametric learning framework.

The results of the proposed multi-task approach [115] show that it is very rare that the annotators are independent of one another, i.e. often annotations are dependent on each other. In contrast, our Multi-Task (MT) approach was tested on the dataset where in some cases the pair-wise inter-annotator agreement is 0. Secondly, the work from [115] sets a prior for each annotator’s function, and this prior must be correlated to each other annotators, whereas in our approach, there is no such type limitation on setting the priors of each annotator’s function.

Fornaciari et al. [117]: The authors proposed a multi-task approach that is able to capture the disagreement of multiple annotators. The proposed approach used two different losses, in addition to the standard error computation (i.e. Single-Task), the authors used soft labels (i.e. probability distributions over the annotator labels) as an auxiliary task. The proposed network measures the divergence between the predictions and the target soft labels with several loss functions. In comparison with this approach [117], our Multi-Task (MT) is not dependent on the outcome of the single task. In our approach, we consider aggregated annotations as an additional annotator on top of N annotators. To get leverage from the aggregated and the individual annotator, we added a novel block called Shared Fully-Connected before predicting soft labels.

Since our MT approach learns from individual and aggregated annotations simultaneously, this limits our MT approach to consider between a range of 5 to 10 annotations including the aggregated annotations. In contrast, the datasets used in the approach [117] for POS and Stemming tasks have 177 and 26 soft annotations respectively. This is the reason behind not testing our Multi-Task (MT) with the approach proposed by Fornaciari et al. [117]. Here, one considerable point is related to the level of subjectivity, i.e. we evaluated our approach on tasks that are more subjective in nature, i.e. affect-related tasks, as compared to [117].

Fayek et al. [118]: In this approach, the authors model the subjectiveness of emotions by incorporating inter-annotator variability, with soft labels and model ensembling, where each model represents an annotator. This approach [118] has a major limitation: the ensemble approach was only tested using three annotators

without mentioning their ids. The categorical label was annotated by 6 different annotators for the IEMOCAP dataset and it is unclear which three annotators were used out of 6 annotators [80]. Secondly, the author’s implementation is not publicly available. Third, the authors did not provide each ensemble DNN performance. Instead, our model is providing the performance of subjective and aggregated emotional perception.

Chou et al. [120]: We compared our Multi-task (MT) approach with this state-of-the-art [120] using the IEMOCAP dataset in Chapter 4. Unfortunately, we can not consider this approach as a general approach for other datasets. The reason behind this is that the Chou et al. approach depends on two previous approaches that have been designed explicitly for the IEMOCAP dataset.

More concretely, the proposed approach uses the hard label and soft emotion distribution which provides complementary affect modeling information, and finally, joint learning of subjective emotion perception and individual rater model provides the best discriminative power.

Chapter 4

Experiments and Results

*“Every advance in knowledge
brings us face to face with the
mystery of my own being”*

Max Planck

This chapter presents the results of Single-Task (ST) and Multi-Task (MT) approaches. Both approaches were first tested with a synthetic dataset and later with human-annotated datasets. The chapter investigates the following questions:

- **Q1:** *How does Multi-Task (MT) approach behave in learning subjective and aggregated affect patterns jointly?*
- **Q2:** *Does Multi-Task (MT) follow the same learning pattern from a less complex (binary class) to a more complex (multi-class) classification problem?*
- **Q3:** *How does Multi-Task (MT) behave in learning subjective and aggregated patterns from different levels of data complexities, i.e. from linear uni-model synthetic data to non-linear multi-model human-annotated data?*

This chapter first presents the experimental details that we used to train our proposed Single-Task (ST) and Multi-Task (MT) models, then the evaluation metrics, and finally the experimental results. For the results, we tested our proposed Single-Task (ST) and Multi-Task (MT) approaches using synthetic datasets and later moved to the human-annotated datasets. The reason behind this is to understand the learning behavior of proposed Multi-Task (MT) learning from less complex to more complex data representation.

4.1 Implementation Details

In this section, we give the experimental details that have been used for the training of our proposed Single-Task (ST) and Multi-Task (MT) architectures in our experiments.

4.1.1 ST Architecture

(1) The fully connected block is composed of three fully connected layers with 1024, 512, and 256 units, respectively. The weights of the neurons in all layers are initialized with the random distribution and the bias is set to 0.

(2) We use Adam/Gradient Descent optimizer [130, 131] for training, with a learning rate 10^{-3} .

(3) Regarding the loss function, all the tasks tested in this work are classification tasks. Thus, we use cross-entropy loss. For datasets that have a big difference between positive and negative affect classes, we use weighted cross-entropy loss:

$$J(w) = -\frac{1}{N} \sum_{n=1}^N [y \log(\hat{y}) * (\alpha) + (1 - y) \log(1 - \hat{y})]$$

Where,

N = Number of total samples

y = true label

\hat{y} = predicted outcome

α = adjust the recall and precision of the model

(4) To encounter overfitting, Lasso Regularization [132] was introduced in the loss (see Equation 4.1). The reason behind $L1$ is that it helps in feature selection by eliminating those features that are not important, i.e. features that do not play a significant role in reducing the gradient.

$$Loss = J(w) + \lambda \sum_{k=1}^M |\gamma_k| \quad (4.1)$$

Where,

M = total number of trainable parameters

$|\gamma_k|$ = an absolute value of a single one

λ = scalar weight given to the regularization term (0.2 is used in the experiments)

(5) Lastly, to get the best model performance we implemented early stopping. It saves the model weights when it has peak generalization capability. If the validation loss does not decrease for $K = 20$ evaluation steps, the training process is stopped. The comparing threshold between the current and the previous loss is 0.01.

(6) The saved model is then reloaded and tested on the test set. The accuracies reported in Chapter 4 are the accuracies that we got on the test set.

4.1.2 MT Architecture

We use the following implementation details to train our Multi-Task (MT) architecture:

(1) The two fully-connected layers in the Shared block have 2048 and 1024 hidden units, respectively.

(2) In each separate perception block, the three fully-connected layers have 512, 256, and 128 units, respectively.

(3) The random initializer is used to initialize the weights of all the neurons in the network, with the bias set to 0.

(4) As in the ST architecture, the sigmoid/softmax function is applied to the output layer of each separate block to get the final prediction of each subjective perception.

(5) The weights of the shared backbone block (in case of developing our own feature extractor) and the two fully-connected layers are updated in every training batch whereas the weights of the fully-connected block from a separate branch are only updated when the batch contains annotations of that specific annotator.

(6) As in the ST architecture, we use Adam/Gradient Descent optimizer with a learning rate 10^{-3} and we also use Lasso regularization to avoid overfitting.

(7) The early stopping technique is applied to each individual block. The main purpose of using this is to save the model weights at the peak performance of each individual and the aggregated perception. We used the same number of evaluation steps and the threshold value as used in Single-Task.

4.2 Evaluation Metrics

In the experiments, we followed the same evaluation metrics that were adopted in the state-of-the-art approach with respect to each dataset. In general, we used two types of evaluation metrics to compare Single-Task (ST) and Multi-Task (MT) approaches: classification accuracy and unweighted average recall.

4.2.1 Classification Accuracy

Classification accuracy is a metric that summarizes the performance of a classification model as the number of correct predictions with respect to all the possible classes divided by the total number of predictions (see Eq. 4.2). One of the problems with this metric is that it does not take into account the imbalanced number of samples per category.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (4.2)$$

4.2.2 Unweighted Average Recall (UAR)

In the Unweighted Average Recall, all the categories have the same importance regardless of the number of samples. This evaluation metric is more appropriate for problems with an imbalanced representation of their categories. UAR is the average recall with respect to each category. The recall (R) is the ratio of the total correctly predicted to the total number of predictions with respect

to a single category (see Eq. 4.3). The classification accuracy gives the correct predictions for all possible categories whereas UAR gives the correct predictions for each class. UAR is calculated by using the Eq. 4.4.

$$\text{Recall}(R) = \frac{\text{Correctly predictions of single class}}{\text{Total predictions of single class}} \quad (4.3)$$

$$\text{UAR} = \frac{1}{N} \sum_{i=1}^N R_i \quad (4.4)$$

Where,

R_i = Recall of the i^{th} category

N = Total number of categories

4.3 Experiments with Synthetic Dataset

Synthetic data generation is a common practice in the machine learning community. The reason behind this is to accelerate the methodical development [133–135]. Our reason behind generating synthetic data is to use a controlled setup to compare Single-Task (ST) and to empirically explore the capacity of Multi-Task (MT) to outperform Single-Task (ST).

In the control setup, we considered the following scenario: we generated (X_1, \dots, X_k) number of examples. The generation process is presented in the upcoming section. Each X_i example has m number of features i.e. $X_i = (X_{i1}, \dots, X_{im})$ and annotated by (A_1, \dots, A_n) annotators in positive and negative emotional dimensions. We assumed that annotators are parameterized by $P_j = P_1, \dots, P_n$, where P_j is the performance that annotator j has on the presence of features presented in example X_i .

4.3.1 Data generation

In our experiments, we used a distribution-based technique for generating features and labels. The following steps are used in data generation:

- Uniform distribution is used to generate $n - dimensional$ feature vector $X = (x_1, \dots, x_m)$. Where x_i is the sampled from a uniform distribution X .

$$X \sim U(0.0, 1.0] \quad (4.5)$$

- Uniform distribution is also used to generate an n -dimensional vector that represents the preference vector P . This vector carries information about the importance that an individual annotator gives to each component in the feature vector. Notice here that $P \neq X$.
- Eq. 4.6 is used to generate the output label for each feature vector X .

$$y = \sum_{i=1}^n (X_i * P_i) \quad (4.6)$$

- All the labels are normalized between 0 and 1.
- All the labels are binarized using 0.5 as a threshold.

Table 4.1 shows the multiple configurations that have been used to generate the synthetic data. The data was generated based on the combination of three different variables which are the number of annotators, number of samples, and feature dimension.

4.3.2 Results

Each configuration of the synthetic data was treated in the following way. Firstly, the data is divided into 10 equal data chunks. Secondly, cross-validation folds were used, i.e. out of 10 data folds, 7 were used for training, 2 for validation,

Chapter 4: Experiments and Results

TABLE 4.1: The different configurations that were used to generate synthetic data.

Configuration_ID	No_of_Annotators	No_of_Samples	Feature_Dimension
Config_1	5	2,000	10
Config_2	5	2,000	100
Config_3	5	2,000	1000
Config_4	10	2,000	10
Config_5	10	2,000	100
Config_6	10	2,000	1000
Config_7	10	10,000	10
Config_8	10	10,000	100
Config_9	10	10,000	1000
Config_10	10	20,000	10
Config_11	10	20,000	100
Config_12	10	20,000	1000

and 1 for testing. We ensured that each data fold must be in the validation and the test set (see Table 4.2). Thirdly, we added aggregated annotator with the individual annotator. Here aggregated annotator is the one that represents aggregated annotations. The arithmetic mean of all individual annotations is used to generate aggregated annotations. Lastly, tested all synthetic data configurations for Single-Task (ST) and Multi-Task (MT). The accuracies represented are the mean of all the cross-validation testing folds. The results show that the common patterns between all individuals and the aggregated annotations help to improve the generalization capabilities of individual and aggregated annotator perception. For simplicity, only the results of Config_1 (see Table 4.3), Config_2 (see Table 4.4), Config_11 (see Table 4.5), and Config_12 (see Table 4.6) are presented in the thesis. These mentioned configurations represent extreme scenarios (less complex to more complex). For example, Config_1 has 5 annotators with 2,000 samples per annotator and the feature dimension of each sample is 10. On the other hand, Config_12 consists of 10 annotators, each annotator has 20,000 samples and each sample is a vector of dimension 1,000. For visual analysis, we present a bar plot between Single-Task (ST) and Multi-Task (MT) approaches for all 12 configurations. In this graph, we only considered aggregated annotator (see Fig.

Chapter 4: Experiments and Results

TABLE 4.2: Cross-validation folds for Single-Task (ST) and Multi-Task (MT) using the synthetic data.

Fold_id	Training	Validation	Testing
Fold_1	2, 3, 4, 5, 6, 7, 9	1, 8	0
Fold_2	3, 4, 5, 6, 7, 8, 9	2, 0	1
Fold_3	0, 3, 4, 5, 7, 8, 9	6, 1	2
Fold_4	0, 1, 2, 4, 7, 8, 9	6, 5	3
Fold_5	0, 1, 2, 3, 5, 6, 8	9, 7	4
Fold_6	0, 1, 2, 3, 4, 7, 8	9, 6	5
Fold_7	0, 1, 2, 3, 5, 7, 8	4, 9	6
Fold_8	0, 2, 4, 5, 6, 8, 9	3, 1	7
Fold_9	0, 1, 2, 3, 4, 5, 7	9, 6	8
Fold_10	0, 1, 2, 3, 6, 7, 8	5, 4	9

TABLE 4.3: Single-Task (ST) and Multi-Task (MT) results for synthetic data, generated using Config_1.

Annotator_id	Single-Task (ST)	Multi-Task (MT)
Annotator_1	81.85	83.50
Annotator_2	80.10	81.15
Annotator_3	77.59	79.60
Annotator_4	78.01	81.50
Annotator_5	79.27	82.50
Annotator_agg	74.68	77.75

4.1).

4.4 Experiments with Human-Annotated Datasets

We consider three public affect-related datasets that provide the individual annotations of multiple annotators: COGNIMUSE [81], IEMOCAP [80], and SemEval-2007 [82]. For each dataset, we give first a short description of the data. Then we analyze the annotator agreement, which empirically illustrates the subjectivity of the tasks. For this, we compute the Cohen Kappa statistic (see Section 2.3) to measure the agreement between every pair of annotators, including the aggregated annotator. After that, we describe the details of the backbone

Chapter 4: Experiments and Results

TABLE 4.4: Single-Task (ST) and Multi-Task (MT) results for synthetic data, generated using Config_2.

Annotator_id	Single-Task (ST)	Multi-Task (MT)
Annotator_1	67.34	70.20
Annotator_2	60.62	62.80
Annotator_3	65.83	66.70
Annotator_4	56.30	63.60
Annotator_5	64.61	66.90
Annotator_agg	59.37	61.20

TABLE 4.5: Single-Task (ST) and Multi-Task (MT) results for synthetic data, generated using Config_11.

Annotator_id	Single-Task (ST)	Multi-Task (MT)
Annotator_1	63.50	65.71
Annotator_2	65.44	67.55
Annotator_3	64.65	66.50
Annotator_4	62.55	65.40
Annotator_5	61.10	64.60
Annotator_6	60.40	64.67
Annotator_7	63.88	68.70
Annotator_8	60.75	65.45
Annotator_9	58.90	62.48
Annotator_10	62.15	65.27
Annotator_agg	65.47	69.85

TABLE 4.6: Single-Task (ST) and Multi-Task (MT) results for synthetic data, generated using Config_12.

Annotator_id	Single-Task (ST)	Multi-Task (MT)
Annotator_1	55.84	58.50
Annotator_2	57.46	60.56
Annotator_3	52.25	55.67
Annotator_4	56.10	58.41
Annotator_5	58.85	61.71
Annotator_6	54.35	57.44
Annotator_7	55.25	58.35
Annotator_8	57.30	59.55
Annotator_9	54.50	56.10
Annotator_10	56.28	58.65
Annotator_agg	60.97	63.20

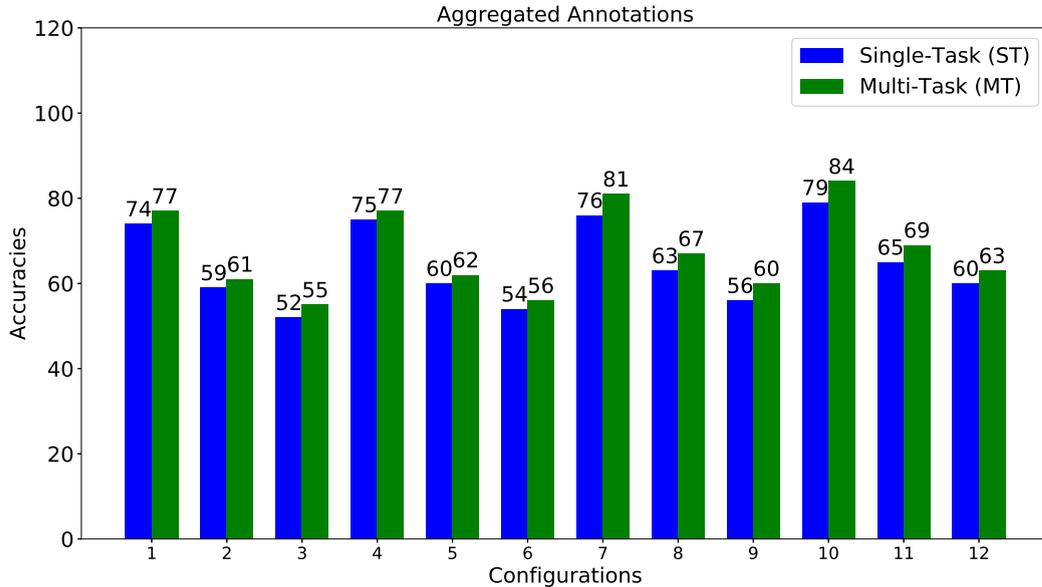


FIGURE 4.1: Comparison between Single-Task (ST) and Multi-Task (MT) approaches for aggregated annotator, considering all 12 configurations (see Table 4.1). To understand the performance difference easily, we only present the integer part of the accuracies.

architecture used for each dataset. Concretely we specify the implementation details used per each of the corresponding data modalities of the dataset (visual, text, and/or audio). Finally, we perform a quantitative analysis and provide some qualitative results for each dataset.

4.4.1 COGNIMUSE Dataset

The COGNIMUSE dataset is a multimodal video dataset [81]. The dataset is generated for multiple tasks such as audio-visual and semantic saliency, audio-visual events and action detection, cross-media relations, and emotion recognition. The dataset includes movies and travel documentaries with human annotations. The standard benchmark for emotional understanding [81] includes 7 Hollywood movies: “A Beautiful Mind” (BMI), “Chicago” (CHI), “Crash” (CRA), “Finding Nemo” (FNE), “Gladiator” (GLA), “The Departed” (DEP), and “Lord of the Rings - the Return of the King” (LOR) and each movie has 30 minutes in length. The emotions evoked by the movies are represented in the valence and arousal space. Valence represents how positive or negative the emotion evoked by the clip

is, while Arousal encodes the viewer’s excitement, agitation, or readiness to act. The frame rate of each movie is 25 fps and 7 different viewers provided annotations for each frame in continuous values from -1 to $+1$ of arousal and valence domains.

4.4.1.1 Data Distribution and Annotator Agreement Analysis

The annotations are binarized into Negative and Positive emotions and each annotator annotates Positive and Negative emotions in a different pattern. Figure 4.2 shows the histogram (total counts) of positive and negative labels per annotator for all movies. For better understanding, Figure 4.3 shows another distribution of positive and negative examples per movie with respect to each annotator. For both figures, it is clear that there is a notable difference among the annotators.

The authors of the COGNIMUSE dataset [81] provide an analysis of the inter-annotator agreement, obtaining a Pearson correlation value of 0.29 for the valence values. This low value for agreement shows that the individual emotional experience is highly subjective. We extended the inter-annotator agreement analysis and measured the pair-wise inter-annotator Cohen’s kappa values (see Figure. 4.4). This is done to understand the emotional subjectivity between each individual as well as with the aggregated emotions. We found that there is a low correlation between each pair of annotators, but a higher correlation between every single annotator with respect to the aggregated annotations (average of all annotations). This higher correlation between every single annotator and the aggregated annotator motivates the idea that a multitask learning approach could leverage the patterns learned from each individual annotator for the prediction of the aggregated annotator.

4.4.1.2 Backbone Architecture

We used Visual, Text, and Audio modalities for the recognition of evoked emotions using the COGNIMUSE dataset. Below we describe the feature extraction backbone for each modality.

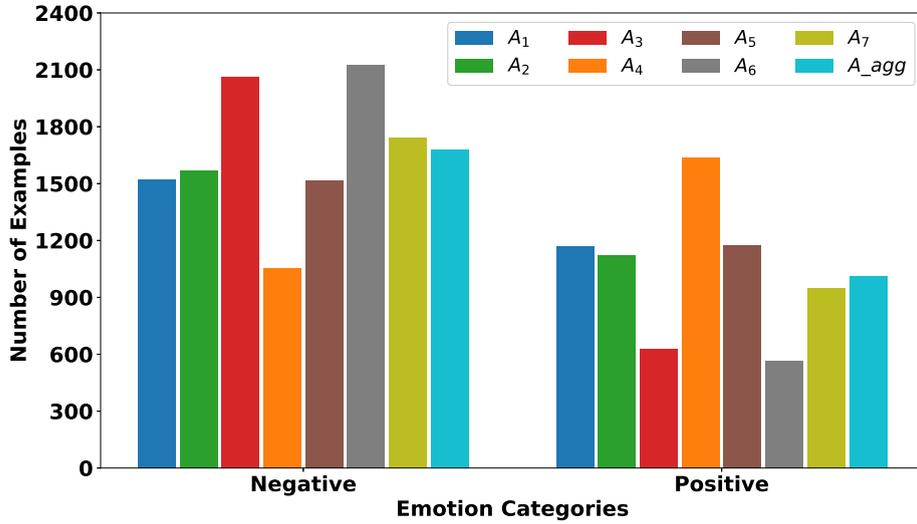


FIGURE 4.2: COGNIMUSE: Distributions of negative and positive examples with respect to each individual and aggregated annotations.

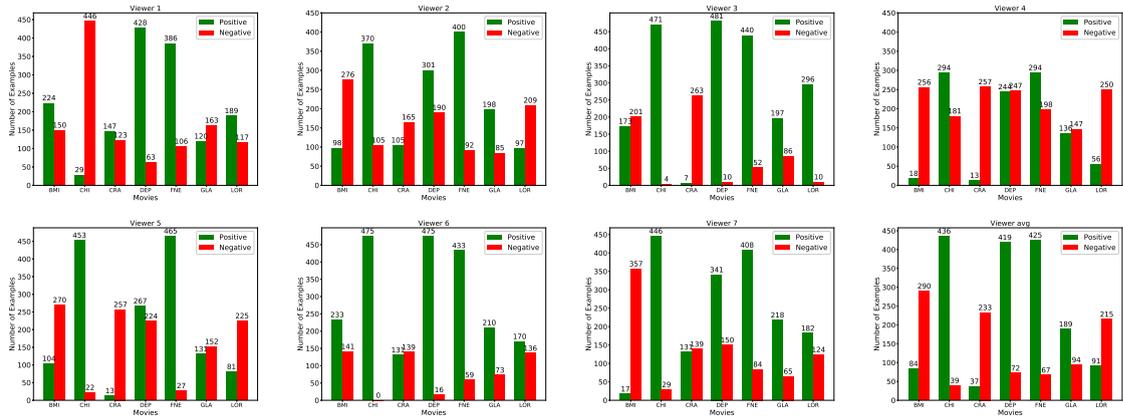


FIGURE 4.3: Annotator/Viewer annotation distribution for Valence (Positive vs. Negative) on each movie, including the average annotator (last plot). Notice that, for all the movies, the distribution of positive and negative labels significantly varies across the different annotator.

Visual Modality- For the visual modality, we use a fixed pre-trained RGB-I3D model [136] on the Kinetics-400 Dataset. The architecture uses 3D convolutions and max-pooling operations to learn seamless spatio-temporal features from video. The I3D model is known as the Inflated 3D, which is based on the state-of-the-art Inception-V1 [137] model. All the convolutional and pooling filters of Inception-V1 were converted from 2D into 3D. This additional dimension is known as the temporal dimension and helps the model in learning temporal

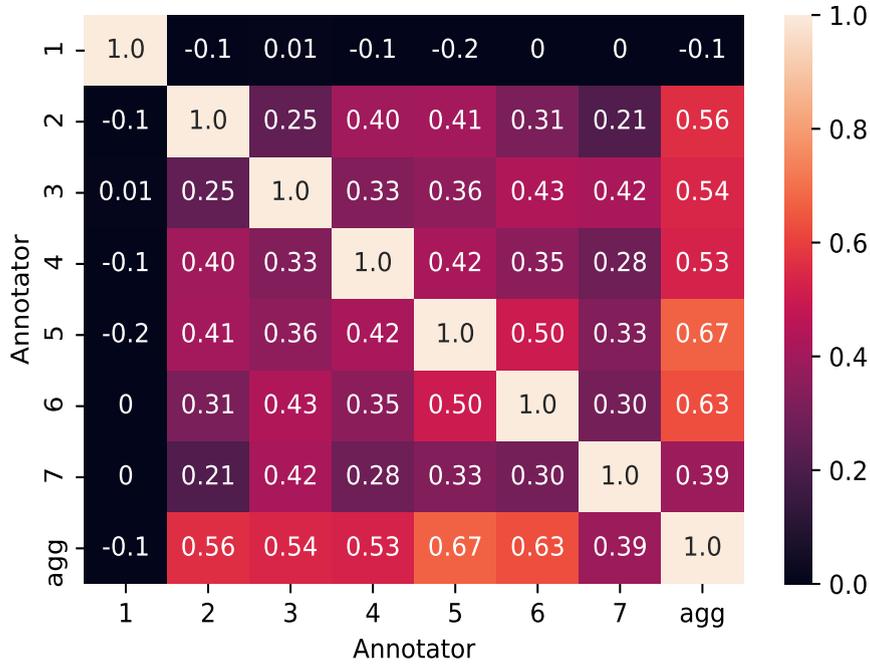


FIGURE 4.4: Pair-wise inter-annotator Cohen’s kappa of COGNIMUSE dataset.

patterns of the video. Each convolutional layer is followed by batch normalization [137] and a ReLU activation function. The I3D model has multiple end-points to collect the features of the given input video. In our experiments, we used the features of the last endpoint of the model, which is “Mixed-5c”. We processed a batch of 16 consecutive frames with a stride of 8 frames of a single clip and the features are then global average-pooled. To get the most important segments of the clip we then max-pooled across the temporal domain.

Text Modality- For the text modality we learn a word-embedding matrix to map every word in a sentence to a d-dimensional vector. This way, sentences can be represented as vectors of numerical values. Since the length of the sentences is variable, a maximum length of 18 words is considered and a particular word is padded at the end of shorter sentences. After the word-embedding we use a sequence of two pairs of convolutional layers plus max pooling. We initialize this text feature extraction branch randomly and we train it with the labeled data.

Audio Modality- For the audio modality, we use the pre-trained VGGish

[138] to get audio features. VGGish is a modified version of VGG (configuration E) [139]. It is trained on the large-scale audio events dataset called AudioSet [140] having 632 audio classes. In our experiments, we use the features of the last convolution layers, which has 512 kernels. Before feeding raw audio into the model, some preprocessing steps are done. The audio sequence of each corresponding clip is first divided into n number of frames having the length of 960 ms, where n represents the length of the clip in seconds. After getting the frames, the next step is to extract the spectral information of each frame, i.e. how much energy the windowed signal contains at different frequency bands. The Short-Time Fourier Transform (STFT) is used for extracting the spectral information with 25 ms window length and 10 ms window shift. STFT transforms each 960 ms frame into a 64 Mel-spaced frequencies vector, and the magnitude of each bin is log-transformed. This gives the 2D log-Mel-spectrogram patch of $96 * 64$ bins. Each audio clip results in a $n * 96 * 64$ tensor. To get the most important frequencies of the clip we then max-pooled across the temporal domain.

4.4.1.3 Results

Firstly, we trained our Single-Task (ST) and Multi-Task (MT) models on cross-validation folds: 5 movies for training, 1 movie for validation, and 1 movie for testing. We ensure that each movie should be in the test set. In comparison with our previous study [141], the audio modality has been also added to the backbone architecture besides the visual and text modalities. The Multi-Task approach showed better generalization for each annotator. On average MT has achieved 4.6 points higher accuracy as compared to the ST for all individual annotators. Furthermore, for the aggregated annotator the improvement is even more significant, i.e. 8.78 points (see Table 4.7). This is strong evidence that MT learning takes advantage of all the annotators in generalization. These experiments could not be reproduced with the Nguyen et al. [142] approach since the code is not publicly available.

Chapter 4: Experiments and Results

TABLE 4.7: ST vs. MT comparison on the COGNIMUSE dataset. Results of ST and MT when modeling single and aggregated annotators with cross-validation evaluation.

Annotator_ID	Methods	
	Single-Task (ST)	Multi-Task (MT)
A_1	65.86	70.09
A_2	68.50	72.13
A_3	66.00	71.38
A_4	71.08	74.29
A_5	78.47	83.07
A_6	71.15	76.22
A_7	67.29	74.05
A_{agg}	71.24	80.02
Mean	69.94	75.15

TABLE 4.8: MT comparison with state-of-the-art on the COGNIMUSE dataset. Result of the Nguyen et al. [142] and MT when modeling aggregated annotator (A_{agg}) with the data split from [142].

Annotator_ID	Methods	
	Nguyen et al. [142]	Multi-Task (MT)
A_{agg}	83.2	90.40

Secondly, the original split of the dataset used in [142] (BMI, CHI, FNE, GLA and LOR in training, CRA and DEP in test) was also considered in order to be able to compare our proposed MT architecture with the Nguyen et al. [142]. Since the Nguyen et al. [142] approach was only trained on aggregated annotations, it is not possible to make a comparison for each individual annotator. Therefore, we only present the results obtained by aggregated annotators when using our MT approach (see Table 4.8). The MT learning performed significantly better and got 7.2 points higher accuracy than the Nguyen et al. [142] approach.

4.4.1.4 Qualitative Analysis

In this section, we present the qualitative analysis of our Single-Task and Multi-Task (MT) approaches. We randomly selected a few samples from the

Chapter 4: Experiments and Results

dataset and compare their ground truth labels with the predictions of our ST and MT models. The main idea is to visualize the effect of the Multi-Task approach with respect to the Single-Task with multiple datasets. The datasets have different amounts of variances for each emotion category with respect to each individual annotator and the aggregated annotator. The results show how the Multi-task approach balance this variance and benefits in predicting the individual and aggregated emotional perception jointly.

Figure 4.5 shows the randomly selected movie clips from each movie of the COGNIMUSE dataset. The ground truth (GT) of each movie with respect to each individual annotator is compared with the predictions of our Single-Task (ST) and Multi-Task (MT) models.

Movie BMI									
Movie Clips	Labels Type	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_{agg}
	GT	Neg	Neg	Neg	Neg	Pos	Neg	Pos	Neg
	ST	Neg	Pos	Neg	Pos	Pos	Pos	Neg	Neg
	MT	Neg	Neg	Pos	Neg	Pos	Neg	Neg	Neg
	GT	Pos	Pos	Pos	Neg	Pos	Neg	Pos	Pos
	ST	Pos	Pos	Neg	Pos	Neg	Pos	Pos	Neg
	MT	Pos	Pos	Pos	Pos	Neg	Neg	Pos	Pos
	GT	Pos	Pos	Neg	Neg	Pos	Pos	Pos	Pos
	ST	Neg	Neg	Pos	Neg	Neg	Pos	Pos	Neg
	MT	Pos							
	GT	Neg	Neg	Neg	Pos	Neg	Neg	Neg	Neg
	ST	Neg	Pos	Pos	Neg	Neg	Pos	Pos	Pos
	MT	Neg	Neg	Neg	Pos	Neg	Neg	Pos	Pos

Movie DEP									
Movie Clips	Labels Type	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_{agg}
	GT	Pos	Pos	Neg	Neg	Neg	Pos	Pos	Neg
	ST	Pos	Neg	Neg	Neg	Pos	Pos	Neg	Neg
	MT	Pos	Pos	Neg	Neg	Pos	Pos	Pos	Pos
	GT	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Pos
	ST	Neg	Pos	Neg	Pos	Pos	Pos	Neg	Neg
	MT	Pos	Pos	Pos	Pos	Pos	Neg	Pos	Pos
	GT	Neg	Pos	Neg	Neg	Pos	Pos	Pos	Pos
	ST	Pos	Pos	Pos	Neg	Neg	Pos	Pos	Neg
	MT	Neg	Pos	Pos	Pos	Pos	Neg	Neg	Pos
	GT	Pos	Pos	Pos	Pos	Pos	Neg	Pos	Pos
	ST	Pos	Neg	Pos	Neg	Pos	Pos	Pos	Neg
	MT	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Pos

Movie GLA									
Movie Clips	Labels Type	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_{agg}
	GT	Neg	Pos	Neg	Neg	Pos	Pos	Neg	Neg
	ST	Pos	Pos	Pos	Neg	Neg	Neg	Pos	Neg
	MT	Neg	Neg	Neg	Neg	Pos	Neg	Neg	Neg
	GT	Pos	Neg	Neg	Neg	Neg	Pos	Pos	Neg
	ST	Neg	Neg	Pos	Pos	Neg	Pos	Pos	Neg
	MT	Pos	Pos	Neg	Neg	Neg	Neg	Pos	Neg
	GT	Neg							
	ST	Pos	Pos	Neg	Neg	Neg	Pos	Pos	Pos
	MT	Neg	Neg	Pos	Pos	Neg	Neg	Neg	Neg
	GT	Pos	Pos	Neg	Pos	Pos	Pos	Neg	Pos
	ST	Neg	Pos	Pos	Pos	Neg	Pos	Neg	Neg
	MT	Neg	Pos	Pos	Pos	Pos	Neg	Neg	Pos

Movie LOR									
Movie Clips	Labels Type	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_{agg}
	GT	Neg	Neg	Neg	Pos	Pos	Neg	Neg	Neg
	ST	Pos	Pos	Neg	Neg	Neg	Neg	Neg	Pos
	MT	Neg	Neg	Neg	Neg	Pos	Neg	Neg	Pos
	GT	Neg	Pos	Pos	Neg	Pos	Neg	Pos	Pos
	ST	Pos	Pos	Neg	Pos	Pos	Pos	Pos	Neg
	MT	Neg	Pos	Pos	Neg	Neg	Pos	Pos	Pos
	GT	Pos	Neg						
	ST	Pos	Pos	Neg	Neg	Pos	Neg	Neg	Pos
	MT	Pos	Neg	Neg	Neg	Pos	Neg	Neg	Neg
	GT	Neg	Neg	Neg	Pos	Pos	Pos	Neg	Pos
	ST	Pos	Pos	Pos	Pos	Neg	Pos	Neg	Pos
	MT	Neg	Neg	Pos	Pos	Neg	Pos	Neg	Pos

FIGURE 4.5: Qualitative results for some randomly selected segments of 4 different movies. The figure shows the ground truth annotation per each annotator, including aggregated annotator, and compares the predictions of Single-Task (ST) and Multi-Task (MT) models with the GT (Ground Truth). Blue is for GT (Ground Truth), Green is for a correct prediction, and Red is for an incorrect prediction.)

4.4.2 IEMOCAP Dataset

The IEMOCAP [80] is an acted, multi-speaker dyadic dataset. A total of 10 different actors (5 male, 5 female) took a part in recording their face motion, head movement, speech, and visual data. The actors played their roles in two different settings: scripted and spontaneous. The recordings were done in sessions and there are a total of 5 sessions in the dataset. After recording the sessions, the dialogues in each session were segmented into utterances. Each utterance was annotated into 9 categorical (anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state) and 3 dimensional (valence, activation, dominance) labels. Six annotators took part to classify the emotional content of utterances into categorical emotion dimensions. Each utterance was annotated by at least three annotators. Instead, for the continuous dimension, two annotators annotated the whole dataset. Only annotator A_3 was discarded because of the low number of annotations provided. The rest of all other annotators and aggregated A_{agg} (majority voting) hold enough annotations which can be used for modeling their emotional perception. The emotions classes used in our experiments are the same as the ones considered in [120]: Anger, Happiness, Neutral, and Sadness. Other emotions such as Surprise, Fear, Disgust, and Frustration are very low in number. This is why these emotional categories did not take into account for modeling. Furthermore, as done in [120], all samples originally annotated as Excitement are also included in the Happiness category.

4.4.2.1 Data Distribution and Annotator Agreement Analysis

Each annotator annotates emotion into 4 different categories. Fig. 4.6 shows the histogram (total counts) of anger, happiness, neutral, and sadness labels per annotator for all utterances.

Whereas the provided inter-annotator agreement in [80] is for the entire dataset (Fleiss' kappa= 0.48), we calculated pair-wise inter-annotator agreement (see Fig. 4.7). The reason behind this is that we are modeling each individual

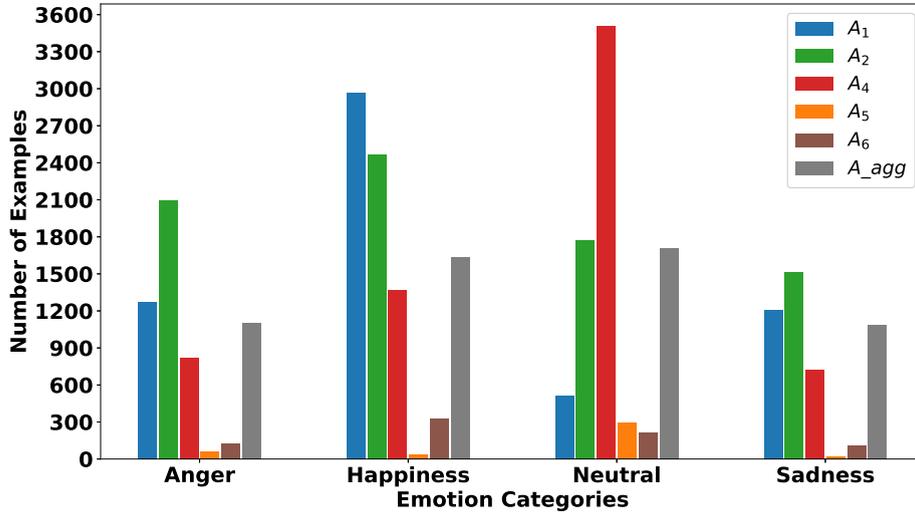


FIGURE 4.6: IEMOCAP: Represents the number of samples in each category annotated by each annotator also with aggregated annotations.

and the aggregated annotator together and pair-wise agreement is a useful measure to understand the agreement and the disagreement between each annotator with respect to others. The “N/A” means that there is no sample annotated by these pairs of annotators. The low inter-annotator Cohen’s kappa shows that emotions are highly subjective. On the other hand, the high Cohen’s kappa between each individual and the aggregated annotators can be interpreted as a loss of emotional subjectivity.

4.4.2.2 Backbone Architecture

To compare with Chou et al. [120] we only used audio modality in our experiments. Therefore, other possible cues such as the face motion or the head movement are discarded. For the audio feature of each utterance, we followed the same approach as we used in the COGNIMUSE dataset (see Section 4.4.1). In the audio preprocessing step, the audio sequence is first divided into n number of frames having the length of $960ms$. In the IEMOCAP dataset few utterances were shorter than $960ms$ so to adjust this we used silence padding after the utterance.

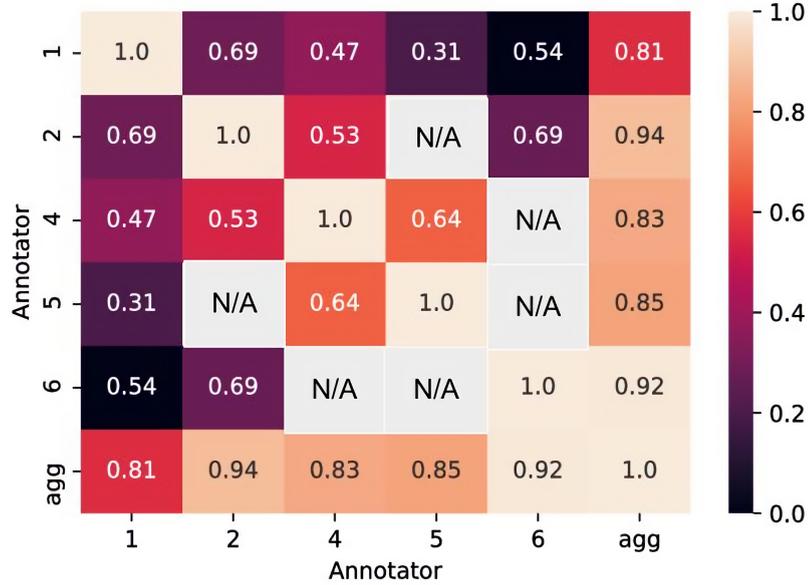


FIGURE 4.7: Pair-wise inter-annotator Cohen’s kappa of IEMOCAP dataset.

4.4.2.3 Results

We tested our Single-Task (ST) and Multi-Task (MT) learning approaches using the leave-one-session-out cross-validation fold. As compared to the Baseline in which 4 sessions were used for training and 1 for testing, we used 3 sessions for training, 1 for validation, and 1 for testing. We ensure each session should appear in the validation and the test set. Whereas we were addressing a binary classification problem in COGNIMUSE dataset, a multi-class classification problem with different emotion categories is tackled on IEMOCAP. The accuracies are reported for each individual and the aggregated annotators with respect to all the possible classes.

As it can be seen in Table 4.9, we found the same pattern for each single and the aggregated annotators as we observed when using the COGNIMUSE dataset. On average, the MT has achieved 8.27 points higher UAR (Unweighted Average Recall) for each individual annotator and 4.76 points higher UAR for the aggregated annotator as compared to the ST. When we compared the MT results with the Chou et al. approach [120], the MT learning got significant improvement in modeling annotator A_1 , A_4 , A_5 , and A_6 , where the point differences are 6.09, 10.93, 17.88, and 3.55 respectively. For the aggregated annotator the MT

Chapter 4: Experiments and Results

TABLE 4.9: Result comparison of ST, MT, and the state-of-the-art using the IEMOCAP dataset. Mean accuracies were obtained with cross-validation. The accuracies are obtained by considering four emotion categories (i.e. 4-class classification).

Annotator_ID	Methods		
	Chou et al. [120]	Single-Task (ST)	Multi-Task (MT)
A_1	50.98	51.70	57.07
A_2	59.68	51.63	58.11
A_4	48.59	53.19	59.52
A_5	37.62	41.04	55.50
A_6	45.82	40.62	49.37
A_{agg}	61.48	56.75	61.51
Mean	50.69	49.15	56.84

performed slightly better than Chou et al. [120]. As compared to Chou et al. [120], on average, our MT approach has increased the classification performance with an increment of 6.41 including the individual and the aggregated emotional perception.

The reason behind this significant improvement in Annotator A_4 and Annotator A_5 is due to the Shared Fully-Connected Block in Multi-Task (MT) learning, which helps to improve the performance of each individual and also the aggregated annotator. In Table 4.10, we can clearly observe the differences in the number of samples per emotion category with respect to each individual annotator and the aggregated annotator. Annotator A_5 has the lowest number of samples per emotion category. However, thanks to the Shared Fully-Connected Block, the separate block of annotator A_5 improves the performance. The results prove that the Shared Fully-Connected Block learns the patterns that are common in each individual and the aggregated annotator and helps in the classification performance of each individual and the aggregated annotator. Our results also show that the Shared Fully-Connected Block in our proposed Multi-Task (MT) approach can deal with imbalanced datasets and helps in improving the classification performance (see Table 4.11). Overall the proposed MT approach again showed that it is best suited to subjective learning from single and aggregated annotators.

Chapter 4: Experiments and Results

TABLE 4.10: The number of training samples per emotion category with respect to each individual and aggregated annotator of IEMOCAP dataset.

Annotator_ID	Emotion Categories			
	Angry	Happiness	Neutral	Sadness
A_1	1271	2970	509	1203
A_2	2095	2467	1769	1514
A_4	821	1369	3511	724
A_5	61	37	297	22
A_6	125	324	212	112
A_{agg}	1103	1636	1708	1084

TABLE 4.11: ST and MT Results per each emotion category using IEMOCAP.

Annotator_ID	Angry		Happiness		Neutral		Sadness	
	ST	MT	ST	MT	ST	MT	ST	MT
A_1	52.12	55.88	54.63	61.50	51.38	56.46	48.68	54.45
A_2	51.42	57.23	53.27	62.48	51.88	56.58	50.26	56.15
A_4	51.54	58.57	52.64	58.87	55.82	63.31	52.79	57.34
A_5	48.20	62.17	32.56	53.99	49.30	57.83	34.13	48.07
A_6	36.54	48.20	42.85	49.20	40.66	46.77	42.46	52.89
A_{agg}	56.44	60.00	57.59	61.62	57.79	62.36	55.20	58.30

4.4.2.4 Qualitative Analysis

In Table 4.12, we present a few utterances that were randomly selected from 5 different sessions of the IEMOCAP dataset. Each utterance was annotated by 3 different annotators. The annotations from all 3 annotators and the aggregate annotators are compared with the predictions of our ST and MT approaches. We can see clearly that the Multi-Task (MT) approach benefits in most cases. Another observation is that there are examples where the output of the aggregated annotator branch does not correspond to the majority voting of the predictions made by individual annotators (see the second example from Table 4.12).

Chapter 4: Experiments and Results

TABLE 4.12: Qualitative results: randomly selected utterances from 5 sessions of the IEMOCAP dataset. The table has the ground truth annotations from each individual annotator, including aggregated annotators, and predictions of Single-Task (ST) and Multi-Task (MT) models. In the table, "-" means not available. "happ" stands for happiness. Blue is for GT (Ground Truth), Green is for a correct prediction, and Red is for an incorrect prediction.

Utterances	Labels Type	A_1	A_2	A_4	A_5	A_6	A_{agg}
I don't understand you, do I?	GT	anger	sadness	anger	-	-	anger
	ST	anger	neutral	neutral	-	-	happ
	MT	anger	sadness	neutral	-	-	anger
What of it?	GT	happ	neutral	neutral	-	-	neutral
	ST	happ	happ	happ	-	-	happ
	MT	happ	neutral	happ	-	-	neutral
Thank you dear. The same applies to you...	GT	happ	neutral	neutral	-	-	neutral
	ST	angry	happ	neutral	-	-	angry
	MT	neutral	neutral	neutral	-	-	neutral
Well there has to be something you haven't tried.	GT	neutral	-	neutral	neutral	-	neutral
	ST	happ	-	anger	anger	-	neutral
	MT	neutral	-	anger	neutral	-	neutral
What?	GT	anger	anger	-	-	anger	anger
	ST	sadness	neutral	-	-	neutral	anger
	MT	anger	anger	-	-	anger	anger
It's just so much.	GT	sadness	-	sadness	sadness	-	sadness
	ST	anger	-	anger	sadness	-	anger
	MT	sadness	-	sadness	sadness	-	sadness
I was coming from um the Midwest, like Iowa.	GT	happ	-	neutral	neutral	-	neutral
	ST	happ	-	happ	happ	-	neutral
	MT	happ	-	neutral	angry	-	happ

4.4.3 SemEval 2007 Dataset

The dataset SemEval_2007 [82] was developed to evaluate the participating systems in order to classify the emotions in news headlines. The dataset consists of 1000 news headlines that were taken from different news channels including CNN, BBC and Google News, and the New York Times newspaper. The news headlines were annotated in 6 categorical emotions (Anger, Disgust, Fear, Joy, Sadness, Surprise) and 1 continuous dimension (Valence). The intensity of the categorical emotion was set between 0 to 100, where 0 represents No emotion and 100 represents the maximum intensity of the emotion category. On the other hand, the Valence dimension was set between -100 to 100, where -100 represents Very Negative, 0 represents Neutral, and 100 represents Very Positive. The dataset is annotated by 5 different annotators. To follow our previous experiments protocol, i.e. modeling individual and aggregated annotations, we added aggregated annotations (average of all annotations) in the dataset. We only considered the Valence dimension for coherence with the experiments done on COGNIMUSE.

4.4.3.1 Data Distribution and Annotator Agreement Analysis

Each annotator annotates valence emotion into different categories. Fig. 4.8 shows the histogram (total counts) of positive, negative, and neutral labels per annotator for all news headlines. We observe notable differences among the annotators.

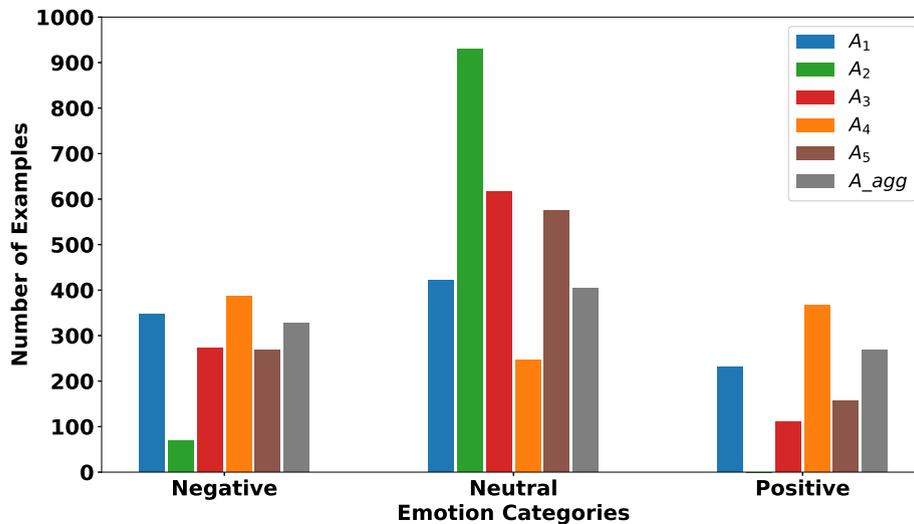


FIGURE 4.8: SemEval_2007: It represents the number of samples in each category annotated by each annotator also with aggregated annotations.

In SemEval_2007 [82], Pearson correlation was measured to understand the inter-annotator agreement for the entire dataset. The inter-annotator agreement for the Valence dimension is 0.78. The agreement measure is a bit high but still holds a subjective nature. Since we are modeling each single and aggregated emotion; pair-wise inter-annotator agreement analysis gives a better understanding (see Figure 4.9). We found the same pattern as we found in COGNIMUSE and IEMOCAP datasets, i.e. the low agreement between every single annotator and a high agreement between each annotator and the aggregated annotator, which supports our multi-task learning hypothesis.

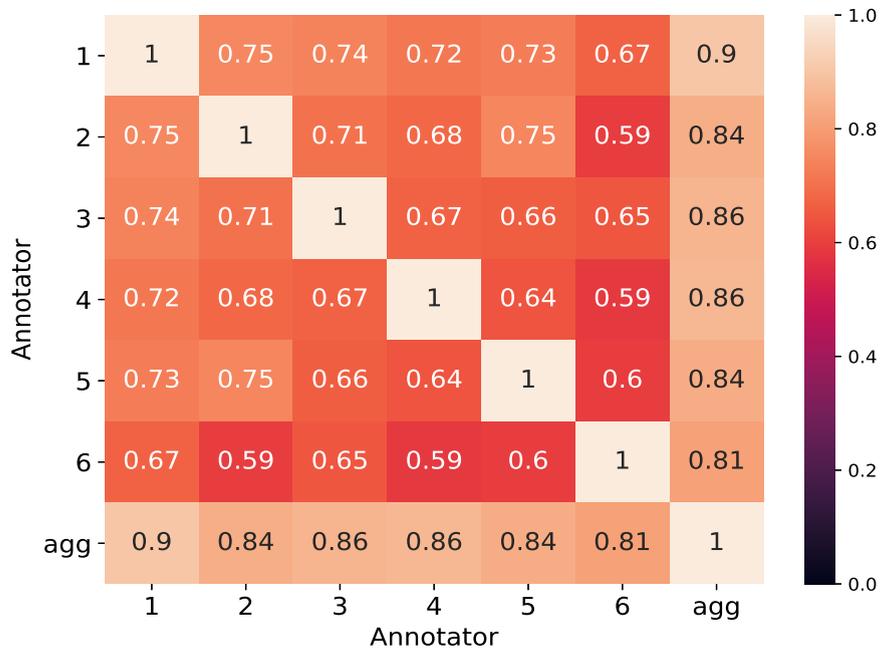


FIGURE 4.9: Pair-wise inter-annotator Cohen's kappa of SemEval_2007 dataset.

4.4.3.2 Backbone Architecture

All the news headlines from SemEval_2007 dataset are textual. There are no other modalities (visual or audio) available. Therefore, only textual modality is used to extract the features. For text feature representation, we used a pre-trained language representation model called BERT [143]. It is based on the Transformers [144], which obtained state-of-the-art results on a wide range of Natural Language Processing (NLP) tasks. BERT is available in many different configurations. From those, we selected Uncased-BERT-Base for our experiments. We considered the output of the *PooledOutput* layer which generates a 768-dimensional vector for each news headline for further classification.

4.4.3.3 Results

We tested our ST and MT approaches using 10-fold cross-validation. We divided the 1000 news headlines into 10 different folds having the same number of news headlines in them. We selected 7 folds for training, 2 for validation, and 1

Chapter 4: Experiments and Results

for testing. We ensured that each fold must be in a test set. The model accuracy is calculated with respect to all 3 classes considered for the Valence dimension: Negative, Neutral and Positive. We observed the same learning pattern of MT as we found for COGNIMUSE and IEMOCAP. As it can be seen in Table 4.13, the experiment results show that the MT learning outperformed the ST approach. On average, MT achieved 1.96 points higher accuracy than ST for every single annotator. For the aggregated annotator, MT is also 1.70 points better than ST in terms of accuracy. Table 4.14 compares the results with the baseline approach. It is worth noting that our models were trained and validated on the same dataset using a cross-validation approach. Instead, the baseline was trained on another dataset. Therefore, our approaches may have some advantage over the baseline since the samples from the same dataset may have more common patterns than samples belonging to other datasets.

TABLE 4.13: ST vs. MT comparison on the SemEval_2007 dataset. Mean accuracies obtained with cross-validation.

Annotator_ID	Methods	
	Single-Task (ST)	Multi-Task (MT)
A_1	58.30	60.60
A_2	80.50	82.30
A_3	66.60	68.40
A_4	57.70	59.80
A_5	58.40	60.20
A_{agg}	57.40	59.10
Mean	63.15	65.06

TABLE 4.14: Comparison of state-of-the-art [82], ST and MT on *SemEval_2007* dataset. Accuracies on the aggregated annotator obtained with the data partition of [82].

Annotator_ID	Methods		
	Strapparava et al. [82]	Single-Task	Multi-Task (MT)
A_{agg}	55.10	57.40	59.10

Chapter 4: Experiments and Results

TABLE 4.15: Qualitative results: randomly selected sentences from SemEval_2007 dataset. The table has the ground truth annotations from each individual annotator, including aggregated annotators, and predictions of Single-Task (ST) and Multi-Task (MT) models. In Table, "neg" stands for negative, "neu" stands for neutral, and "pos" stands for positive. Blue is for GT (Ground Truth), Green is for a correct prediction, and Red is for incorrect prediction.

Sentences	Labels Type	A_1	A_2	A_3	A_4	A_5	A_{agg}
Cases: when the simple solution is the right one	GT	pos	pos	neu	neu	pos	pos
	ST	neu	neu	neu	neg	pos	neg
	MT	neu	pos	neu	pos	pos	pos
Rio De Janeiro jmyal: drawing lines across the sand, between classes	GT	neu	neg	neg	neu	neu	neu
	ST	neu	pos	neg	pos	neg	pos
	MT	neu	neg	neg	pos	neu	neu
Passing exchange becomes political flashpoint	GT	neu	neu	neu	neu	neu	neu
	ST	pos	pos	neu	neu	neu	pos
	MT	neu	pos	neu	pos	pos	neu
Memo from Frankfurt: Germany relives 1970s terror as 2 seek release from jail	GT	neu	neg	neg	neg	neg	neg
	ST	neg	neg	pos	neu	neg	neg
	MT	neu	neg	pos	neu	neg	neg
Luxury digs in South Carolina's Lowcountry	GT	neu	neu	pos	neu	neu	neu
	ST	neg	pos	pos	neu	neu	neg
	MT	neu	pos	pos	neu	neu	neu

4.4.3.4 Qualitative Analysis

For SemEval_2007 data, we randomly selected 5 sentences and compare their ground truth (GT) with the predictions of Single-Task (ST) and Multi-Task (MT) models with respect to each individual annotator. We can see how our proposed Multi-Task (MT) approach gives benefits in modeling subjective and aggregated emotional perception (see Table 4.15).

Chapter 5

Discussion

*“If you can’t summarize an issue
on one page, you don’t
understand the issue well
enough.”*

Ronald Reagan

In this study, we addressed the problem of subjectivity in supervised machine learning, particularly in the context of affect recognition. The problem of subjectivity has been addressed broadly in two broad dimensions: Subjectivity as noise and Subjectivity as information. Researchers have proposed different approaches to tackle subjectivity under both perspectives. The proposed approaches considered subjectivity from different tasks. Some tasks are less subjective such as Quality Estimated of translated sentences or Image classification, but some are highly subjective such as emotions. Our target is emotional subjectivity and we considered subjectivity as information.

Researchers who considered subjectivity as information proposed techniques for modeling emotional perceptions in a joint manner, i.e. modeled subjective and aggregated emotional perception together. These approaches have a major limitation: they are modeling subjective perception but the final output is a single emotional label that loses the true subjective perception with respect to each individual annotator.

Our Multi-Task (MT) approach also followed the joint modeling of subjective and the aggregated emotional perception. In contrast with others, in our Multi-Task (MT) we treated aggregated perception as an individual annotator rather than modeling the subjective perception first and then use any mathematical technique such as mean or majority voting to get a single emotional representation. This single emotional representation basically loses the subjective perception and hence represents the aggregated perception. Modeling aggregated perception separately gives more benefits as compared to other approaches where aggregated perception is dependent upon the concatenation of subjective perception. Our aim is to model inter-annotator disagreement and the inter-annotator agreement in such a way that each subjective perception remains separate but there are some common patterns that can be learned and benefit to each individual and the aggregated perception. Secondly, we considered multiple modalities in our experiments as compared to previous approaches, where they just considered a single modality.

At the beginning of Chapter 4, we were looking at three research questions. To find out the answers to these questions, firstly, we evaluated our proposed Single-Task (ST) and Multi-Task (MT) approaches using synthetic datasets. After that, we moved to use three affect-related datasets: COGNIMUSE, IEMO-CAP, and SemEval_2007. The results show that our proposed MT performed well in modeling subjective and aggregated emotional perception. The novel Shared FC Block has performed well in learning common patterns from all annotators, including individual and aggregated ones. This addresses the first and the last research questions of Chapter 4. Furthermore, the experiments performed to analyze the performance of proposed Single-Task and Multi-Task (MT) are ranging from binary to multi-class emotion classification, which satisfies the research question second of Chapter 4. The answers to these three questions prove our hypothesis, i.e. joint modeling of all emotional perceptions in a multi-task manner has more generalization capabilities compared to modeling all emotional perceptions available in the dataset in a single-task manner.

The proposed Multi-Task (MT) approach performed well when the number of annotators is low, i.e. from 5 to 10 annotators. To our best knowledge, there is no publicly available affect-related dataset that has hundreds or thousands of annotations per data sample. This is why we are unable to evaluate this approach in a scenario with a large number of annotators. One possible future direction of this work is to find a way to learn common patterns when we have hundred or thousand of annotators while keeping the disagreement in contact independently for affect-related tasks.

Publications

- Hassan Hayat, Carles Ventura, Agata Lapedriza, ‘Modeling Subjective Affect Annotations with Multi-Task Learning’, **Sensors**. 2022; 22(14):5245. DOI: <https://doi.org/10.3390/s22145245>.
 - Source code available at: [GitHub Repository Link](#)
- Hassan Hayat, Carles Ventura, Agata Lapedriza, ‘Recognizing Emotions evoked by Movies using Multitask Learning’, International Conference on **Affective Computing & Intelligent Interaction (ACII 2021)**. DOI:<https://arxiv.org/abs/2107.14529>
 - Source code available at: [GitHub Repository Link](#)

Part II

Emotion Subjectivity and Personality Traits

Chapter 6

Introduction

*“Personality has power to uplift,
power to depress, power to curse,
and power to bless”*

Paul Harris

This chapter presents a short review of works in psychology related to human personality from two perspectives: (i) how to characterize human personality, and (ii) how personality influences emotional experiences.

Personality is defined as a psychological construct aimed at explaining the wide variety of human behaviors in terms of a few, stable and measurable individual characteristics. In other words, “individual differences in characteristic patterns of thinking, feeling, and behaving” [145]. Personality predicts the patterns of thought, emotion, and behavior [146] as well as important life aspects, including happiness, physical and psychological health, quality of relationships with peers, family, lovers others occupational choice, satisfaction, and performance, community involvement, criminal activity, and political ideology [147]. Furthermore, attitude and social behavior towards a given individual depend, to a significant extent, on the personality impression others develop about him/her [148].

A personality trait is a pattern of behaviors that are related to the person who is showing the consistency of such pattern from situation to situation [149]. Currently, various personality trait theories have been developed to categorize, interpret and understand the human personality. For example, Helen Palmer proposed a model named Enneagram [150] which categorizes human personality into 9 different traits (perfectionist, helper, achiever, individualist, investigator, loyalist, enthusiast, challenger, and peacemaker). Similarly, Catell et al. [151] categorize human personality into 16 different traits. Another researcher named Hans Eysenck [152] proposed a model which consists of 3 personality traits (psychoticism, extraversion, and neuroticism). Another model named Myers-Briggs Type Indicator (MBTI) was proposed by Katharine Briggs and her daughter Isabel Briggs [153]. MBTI model has 8 different traits (introversion, extraversion, sensing, intuition, thinking, feeling, judging, and perceiving). Lastly, the most famous and widely used model for human personality is Big-Five [154]. The Big Five assess human personality in five dimensions, which are known as the Big Five Personality Traits, sometimes also called OCEAN (Openness, Conscientiousness,

Extraversion, Agreeableness, Neuroticism). In this thesis, we followed the Big-Five personality trait model. This is why the details of other models are out of the scope of this thesis.

The Big-Five (OCEAN) personality traits are the following: *(i)* **Openness** to experience, which encodes how imaginative versus practical someone is, *(ii)* **Conscientiousness**, which encodes whether the person is organized or sloppy, *(iii)* **Extraversion**, which measures whether the person is friendly or reserved, *(iv)* **Agreeableness**, which measures how authentic or self-interested a person is, and *(v)* **Neuroticism**, which measures whether someone is comfortable or uneasy. These personality traits have been repeatedly obtained by applying factor analyses to various lists of trait adjectives used in personality description questionnaires [155–157]. The basis for such factor analyses is the Lexical Hypothesis [158], i.e. that the most relevant individual differences are encoded into the language, and the more important the difference, the more likely it is to be expressed as a single word.

Over the past few decades, the Big Five model has become a standard in psychology, and experiments using the Big Five have shown that personality traits influence many aspects of task-related individual behavior. For example, the success of most interpersonal tasks depends on the personalities of the participants, and personality traits influence leadership ability [159], general job performance [160], attitude toward machines [161], teacher effectiveness [162], academic ability and motivation [163, 164], and personality traits predict the desired affective state of people [165, 166]. In contrast with other affective dimensions such as emotions or mood, which may be relatively contextualized or short-lived, human personality is usually considered to be a longer-term and more stable aspect of life [167].

Researchers in psychology found that personality has an influence on emotional responses [168–171]. To understand this influence, psychologists established a link between emotional responses and Emotional Intelligence (EI) [172–174]. Emotional Intelligence (EI) is first defined by Salovey et al. [175] in 1990 as “the subset of social intelligence that involves the ability to monitor one’s own and

other’s feelings and emotions, to discriminate among them, and to use this information to guide one’s thinking and actions”. Later, there have been strong controversy about the definition and nature of emotional intelligence [176]. Over the years, experts have come to an agreement that there are two types of emotional intelligence: the ability of emotional intelligence and the trait of emotional intelligence. On the one hand, the ability of emotional intelligence entails a particularly high ability to process emotional information that is related to, but distinct from cognitive ability. On the other hand, the trait of emotional intelligence, which is also known as Trait Emotional Intelligence (trait EI), is a construct first proposed by Petrides et al. [177]. Trait Emotional Intelligence relates to personality and represents a combination of personality traits, particularly effective in situations with emotional and social implications. Further criticism against measuring EI as an ability is that emotional experiences are very subjective [178]. As a result of these criticisms, it is widely accepted by researchers that EI is best conceptualized as a trait [179, 180]. Further research findings have shown strong relationships between trait emotional intelligence and the Big five personality traits [181–184].

In this part of this thesis, we present the research work that emphasizes the use of the Multi-Task (MT) approach in modeling subjective emotional perception of speakers using their personality patterns, i.e. Big five personality traits. In particular, we proposed a Multi-Task (MT) model that efficiently models subjective emotional expression with respect to each individual speaker in dialogues systems compared to the Single-Task.

Chapter 7

A Study on Modeling Subjective Emotion Expression using Personality Traits in the Context of Dialogue Systems

*“The social brain includes
circuitry designed to attune to
and interact with another
person’s brain”*

Daniel Goleman

This chapter presents the personalized emotion experienced by each individual speaker in the context of dialogue systems. The study shows that the personality information, when used in Single-Task (ST) settings, is not so much effective in generating subjective emotional responses. The results show that personality traits help to boost the Multi-Task (MT) learning capabilities in order to predict the subjective emotional response of each speaker in the dialogue system.

As discussed in Chapter 6, the automatic recognition of emotions has been widely studied in computer science. A wide variety of research has been conducted in detecting emotions from images, body movements, speech signals, text, and physiological signals, which include the electroencephalogram (EEG), temperature (T), electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), respiration (RSP), etc. The established research work in the field of speech signal processing and textual understanding revolutionized Human-Computer Interaction (HCI) and developed the advanced form of natural language interface called conversational agents/chat-bots/dialogue systems/virtual assistants [185, 186].

Dialogue systems are an important tool to achieve intelligent user interaction. These systems have the capability to process natural language data and simulate a smart conversational process with humans [187]. These conversational mechanisms are built and driven by a wide variety of techniques of different complexity, from traditional, pre-coded algorithms to emerging adaptive machine learning algorithms [188]. Dialogue systems have gotten attention in recent years [189] but the study of natural language communication between human beings and machines is indeed not a novel concept. ELIZA [190], which has been historically considered the first chatbot, was designed and developed by the Massachusetts Institute of Technology (MIT) more than half a century ago, between 1964 and 1966. Alongside successive chatbots like PARRY [191], these innovative systems laid the groundwork for specialized research in the field of human-computer interaction (HCI), focusing on the social and communicative perspectives and their impact on the design and development of these systems. Recent advances in the

field of artificial intelligence have brought back attention to the potential of conversational agents, especially with the emergence of machine and deep learning techniques [192]. Furthermore, specialized research fields such as natural language understanding (NLU), natural language generation (NLG), and dialogue stage tracking (DST) have become disruptive areas by introducing innovative, efficient, and accurate solutions to machine cognitive problems [193].

In the quest for generating more human-like conversations, one of the major challenges is to generate emotional responses. Recent development in artificial intelligence approaches pushes researchers to develop dialogue systems that are more understandable not only in contextual meaning but also emotionally in Human-Computer Interaction (HCI) [194–198]. Emotional understanding is an essential feature for many conversation scenarios such as social interaction and mental health support [199, 200]. Research findings support that incorporating emotions is advantageous to the dialogue system by allowing the dialogue system to emulate the conversational behavior of human beings and, at the same time, to increase the user’s engagement in the conversation [201–203].

Currently, research on dialogue systems focuses on task-oriented dialogue (TOD) systems [204–206], and chit-chat dialogue systems also called open-domain dialogue systems (ODD) [194, 207–209]. The aim of this study is to infer the subjective emotional responses using Big Five personality traits in any type of dialogue system. Concretely, given a few conversation utterances, we approach the problem of predicting the emotion category (positive, neutral, negative) of the upcoming utterance with respect to each individual speaker (see Fig. 7.1).

Our research work is based on the previous state-of-the-art [210] approach where the authors showed that personality traits help in predicting subjective emotional responses with respect to speakers. In contrast, our objective is to show that Multi-Task (MT) approach is more efficient in predicting the subjective emotional responses of speakers with the help of their personality traits patterns.

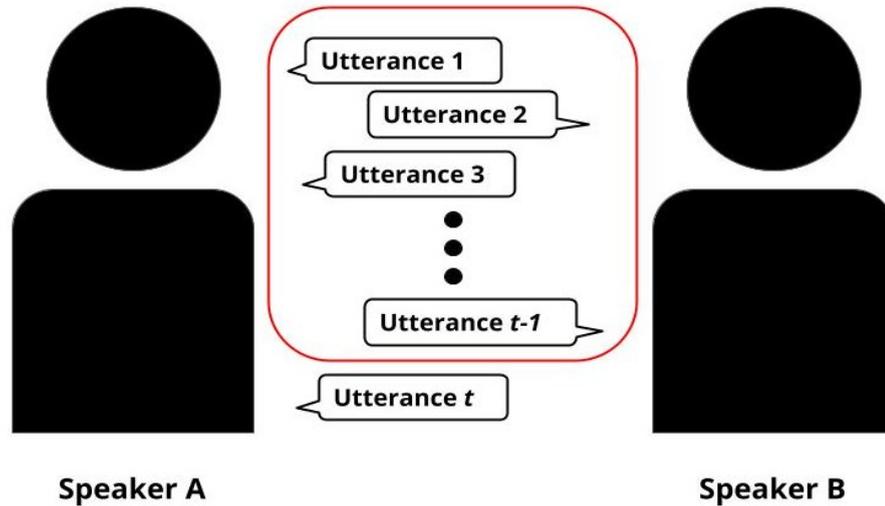


FIGURE 7.1: This figure replicates the scenario of a typical conversation between two speakers. The first utterance is from Speaker A, the second is from Speaker B, and so on. In this example, our aim is to predict the emotion category of the upcoming utterance (i.e. utterance t) with respect to Speaker A. To do this, the system will use all the preceding $t - 1$ utterances (represented in red outline), preceding emotions E_i , and the personality traits of Speaker A.

Application Domain Areas

Affective Healthcare/E-health: Researchers in psychology already established the strong association between Big Five personality traits with different indications of subjective well-being such as positive and negative mood, general health concern, chronic illness, serious illness, and psychological distress [211–214]. Developing computational systems that are emotionally intelligent to understand human psychological health using two-way communication between the patient and system is an emerging research area.

Social Robotics: Social robots are designed to interact with people in a natural and interpersonal manner. The aim is to achieve social-emotional understanding in diverse applications such as education, health, quality of life, entertainment, communication, and collaboration. The long-term goal of creating social robots that are competent and capable partners for people is quite a challenging task. This drove the researchers not only to develop a cognitive level but also an emotional level of understanding as well for robots.

Research Question

Is the addition of personality information in affect recognition systems useful to model subjective emotional responses?

Hypothesis

The Big Five (OCEAN) personality traits of any human may help in predicting subjective emotional responses in a conversation.

7.1 Related work

Most of the existing work focuses on generating the specified emotional response. For example, Zhou et al. [194] explored the emotional factor in large-scale conversation generation. The authors proposed Emotional-Chatting Machine (ECM) that is capable to generate emotional responses. Colombo et al. [215] designed an affect-driven dialogue system that is used to generate emotional responses using the continuous representation of emotions. Zandie et al. [216] proposed an EmpTransfo (a multi-head Transformer) model for generating the emotional response. The proposed architecture understands the emotion of the user and then generates a response empathetically. Another study on empathic response was proposed by Zhong et al. [217]. They developed a model called CoBERT, which efficiently generates the empathic response in conversation based on the human persona. Liu et al. [218] approach to generating empathic response is not based on the given certain emotions. Instead, their idea is to understand the user's emotions first and then reply appropriately. Asghar et al. [219] proposed three different ways to incorporate affective responses in conversation: (a) affective embeddings, (b) affective-base loss function, and (c) affective beam search for decoding. The results show that all three proposed methods improved emotional responses in conversation. Li et al. [220] developed a network that uses reinforcement learning to generate more meaningful and customized emotional responses. To generate the emotional responses the network first predicts an emotional keyword for an input dialogue at the initial stage. This emotional keyword

behaves as a piece of prior knowledge throughout the processing and helps to predict the final emotional response. Sun et al. [221] designed a reinforcement learning-based architecture that uses emotional tags with the input dialogue and the generated response dialogue. These emotional tags partially reward the model for generating satisfactory emotional responses.

In contrast, research work that automatically selects the emotion for a response without any initial consideration of emotion is seldom discussed. Wei et al. [222] proposed a dialogue system that can respond at semantic and emotional levels. To learn emotional responses, the network was trained on online dialogues. These dialogues or conversations belong to different speakers. This is why the proposed dialogue system is unable to understand individual differences in expressing emotions. Overall, there is not much research work on automatically selecting emotional responses with respect to the individual speaker. Zhou et al. [194] highlighted the issue of emotional subjectivity in their research work and left this for future work. Recently, Wen et al. [210] work is the first that considers the subjective emotional response automatically. The authors proposed a dialogue system that takes into account individual personality traits. To automatically select the subjective emotional responses, the dialogue system first simulates the transition of emotions in the conversation. Then this transitioned emotion is triggered by two factors: (i) the preceding dialogue context, and (ii) the specified individual personality traits. Finally, the response emotion is the sum of the preceding emotion and the transitioned emotion. This work [210] is state-of-the-art in automatically generating subjective emotional responses in dialogue systems. The idea to use the transition of emotions for predicting the future emotions of any individual was inspired by the research work done by Thornton et al. [223].

In [210], the authors first model the preceding emotions with the help of the preceding dialogue context in the Valence-Arousal-Dominance (VAD) emotional space. Then they used the preceding dialogue context and the specified personality traits to encode the transition of emotions. Finally, the emotional response is selected from the sum of the preceding emotions and the transition of emotions. However, the proposed approach [210] uses the personality information of

each speaker in order to generate the emotional response but the model has a single output layer for each speaker. As a result, the model loses the true sense of emotional subjectivity with respect to each speaker. The model is also biased toward the majority emotion class and there is a considerable difference between the classification of each emotion class in predicting 7-class emotion categories.

7.2 Dialogue Datasets with Emotion Labels

Most of the available datasets in dialogues that are labeled with emotional information only have emotional annotations with no subjective information such as the speaker’s personality traits or other metadata (see Table 7.1 for a summary of datasets).

TABLE 7.1: “Speaker Affect Annotations”: dialogues annotated into emotional dimensions, “Speaker’s Subjective Affect Annotations”: speaker’s subjective information such as personality traits or mood.

Dataset	Speaker Affect Annotations	Speaker’s Subjective Affect Annotations
DailyDialog [195]	Yes	No
MOJITALK [196]	Yes	No
CBET [224]	Yes	No
EMPATHETICDIALOGUES [225]	Yes	No
PERSONA-CHAT [226]	Yes	No
TED-LIUM [227]	No	Yes
AIT-2018 [228]	Yes	No
PELD [210]	Yes	Yes

Li et al. [195] developed a dialogue dataset named DailyDialog. The dialogues in the dataset reflect daily communication. Each utterance in the dataset was annotated into two categories: (i) act classes (inform, question, directives, and commissive), and (ii) emotion classes (anger, disgust, fear, happiness, sadness, and surprise). Fig. 7.2 shows an example of the conversation in the dataset.

A: I'm worried about something.
B: What's that?
A: Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.
B: That's annoying, but nothing to worry about. *Just breathe deeply when you feel yourself getting upset.*
A: Ok, I'll try that.
B: Is there anything else bothering you?
A: Just one more thing. A school called me this morning to see if I could teach a few classes this weekend and I don't know what to do.
B: Do you have any other plans this weekend?
A: I'm supposed to work on a paper that'd due on Monday.
B: *Try not to take on more than you can handle.*
A: You're right. I probably should just work on my paper. Thanks!

FIGURE 7.2: An example in **DailyDialog** [195] dataset. The underlined words in red indicate the emotions in the utterances.

Zhou et al. [196] developed a large-scale Twitter conversation dataset named MOJITALK that includes emojis in the response and consider that these emojis represent the underlying emotions category. A sample of a dataset is shown in Fig. 7.3.



FIGURE 7.3: Twitter conversation with emoji (top) in MOJITALK [196] dataset

Shahraki et al. [224] also developed a Twitter-based dataset named Cleaned Balanced Emotional Tweets (CBET) to understand the emotional content in the conversation. The authors consider only those tweets that have emotional hashtags

Chapter 7: Subjective Emotions using Personality Traits

and use those hashtags as tweet labels. Fig. 7.4 shows few tweets with emotional hashtags.

Tweet	Hashtag	Label
Say No Fur! animal rights #animalrights #vegan #compassion	#disgusting	disgust
@user We send super surprise gift to japan #Fiverr #gift #japan	#surprise	surprise
Thank Lord Thank blessings guiding everyday You never fail #blessed	#thankful	thankfulness
#jewelry #sets Vintage shell necklace matching earrings real gold marked	#love	love
#rain #hail #thunder #storm	#fear	fear

FIGURE 7.4: Tweets with emotion hashtags in CBET [224] dataset

Rashkin et al. [225] created a dataset named EMPATHETICDIALOGUES to recognize feelings when a conversation is happening with a machine partner and reply accordingly in an empathetic manner. Each conversation in the dataset is annotated with a situation label and the emotion label. Two conversation samples are shown in Fig. 7.5.

<p>Label: Afraid Situation: Speaker felt this when... "I've been hearing noises around the house at night" Conversation: Speaker: I've been hearing some strange noises around the house at night. Listener: oh no! That's scary! What do you think it is? Speaker: I don't know, that's what's making me anxious. Listener: I'm sorry to hear that. I wish I could help you figure it out</p>	<p>Label: Proud Situation: Speaker felt this when... "I finally got that promotion at work! I have tried so hard for so long to get it!" Conversation: Speaker: I finally got promoted today at work! Listener: Congrats! That's great! Speaker: Thank you! I've been trying to get it for a while now! Listener: That is quite an accomplishment and you should be proud!</p>
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

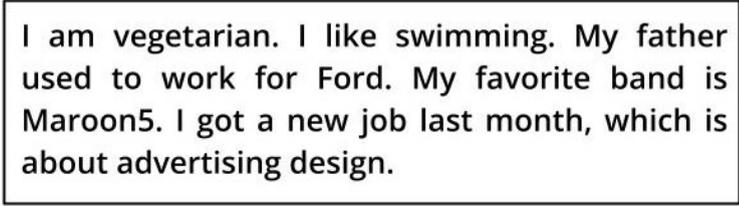
FIGURE 7.5: Two examples from EMPATHETICDIALOGUES [225] dataset

Fung et al. [227] developed a virtual robot called 'Zara the Supergirl'. This virtual robot can empathize while interacting with a user. It also has the ability after 5 to 10 minutes of conversation. The robot was trained 207 hours of speech extracted from 1495 TED Talks and annotated into six (criticism, anxiety, anger, loneliness, happiness, and sadness) emotional categories. The authors also included the personality information of the user that interacts with the robot.

Chapter 7: Subjective Emotions using Personality Traits

This dataset has the personality information of the users but does not have the user-associated dialogues to get the subjective emotional responses.

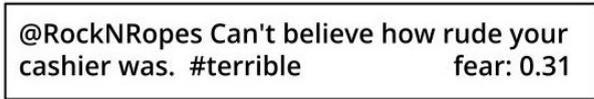
Zhang et al. [226] proposed a new idea for knowing more about speakers while interacting with the machines. They developed a dataset called PERSONA-CHAT. The dataset consists of the profile information of the speakers. The results showed that this improved the engaging behavior of chit-chat systems. Fig. 7.6 shows a sample of profile information in the PERSONA-CHAT dataset.



I am vegetarian. I like swimming. My father used to work for Ford. My favorite band is Maroon5. I got a new job last month, which is about advertising design.

FIGURE 7.6: Profile information of a speaker in PERSONA-CHAT [226] dataset

Mohammad et al. [228] developed a Twitter-based dataset to infer the emotional state of a person from their tweets. The dataset was annotated into fear, joy, sadness, and anger emotional categories. For each of the four emotions, the 0 to 1 range is partitioned into the classes: no emotion can be inferred, low emotion can be inferred, moderate emotion can be inferred, and high emotion can be inferred. These are referred to as emotion intensities. This dataset does not have any person's subjective information for modeling subjective emotional responses. An example of a tweet with annotated emotional category is shown in Fig. 7.7.



@RockNRopes Can't believe how rude your cashier was. #terrible fear: 0.31

FIGURE 7.7: A single tweet is annotated with an emotional category in SemEval_2018 [228] dataset

All of the above datasets were built to understand the speaker/user's affective state during conversation. Since emotions are highly subjective in nature

it is essential to understand the affective response in consideration of subjectivity. None of the above-discussed datasets annotated speaker/user’s emotional responses with their subjective information. Recently, Wen et al. [210] constructed a dataset called Personality EmotionLines dataset (PELD). The dialogues in the dataset are annotated with the speaker’s personality traits. It consists of dialogue scripts taken from a famous TV series named Friends. The authors only consider the conversation from six speakers: Chandler, Joey, Monica, Phoebe, Rachel, and Ross. Each dyadic conversation represents a triplet. Fig. 7.8 shows a triple example of the PELD dataset. The utterances and their emotional labels are taken from the dialogues in the MELD [229] and the EmoryNLP dataset [230], the two famous datasets analyzing emotional responses in Friends.

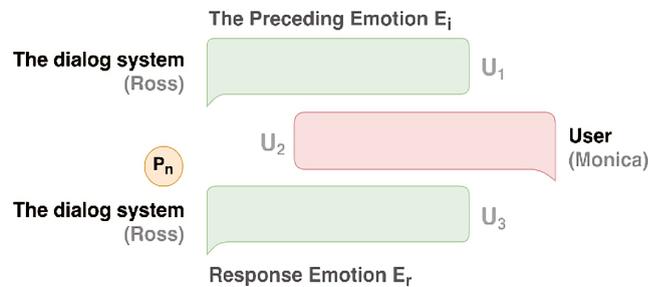


FIGURE 7.8: A triple example in PELD [210]. The dyadic conversation between Ross and Monica (two main roles in the Friends TV show). P_n is the personality of Ross. The dialogue system is set as Ross and talks with the user which is set as Monica in this example. The response emotion E_r corresponds to the utterance U_3 and the preceding emotion E_i corresponds to the utterance U_1 both are from the system, i.e. Ross. The emotions of the user’s (Monica) utterance is unknown. The system will predict the response emotion E_r of utterance U_3 using preceding emotion E_i and all preceding dialogues, i.e. U_1 and U_2 in this example.

7.3 Proposed Approach

In any conversation, every speaker generates emotional responses differently. Generally, these emotional responses are dependent upon the context of the conversation and the speaker itself, where the personality, culture, and internal state of the speaker modulate this response. In this study, we have the

personality information and we will use this information during inference. Concretely, in our experiments, we compare two types of machine learning models: (i) a Single-Task (ST) architecture, where we can model each subjective emotional response independently, and (ii) a Multi-Task (MT) architecture, where all multiple subjective emotional responses can be modeled jointly. Both models use a parameterization of personality traits as a source of information.

7.3.1 Single-Task (ST) Architecture

In Single-Task (ST) architecture, at a time the model is trained with a single personality vector. Therefore the model has the capability to generate emotional responses that are specific to that specific personality. The Single-Task architecture is illustrated in Fig. 7.9. The response emotions E_r of upcoming utterance U_t are mainly dependent on the dialogue context and the given personality information (see Eq. 7.1).

$$E_r = ST(E_i|P, C) \tag{7.1}$$

Where E_r is the response emotion, E_i is the preceding emotion, P is the personality vector of a speaker, and C is the dialogue context.

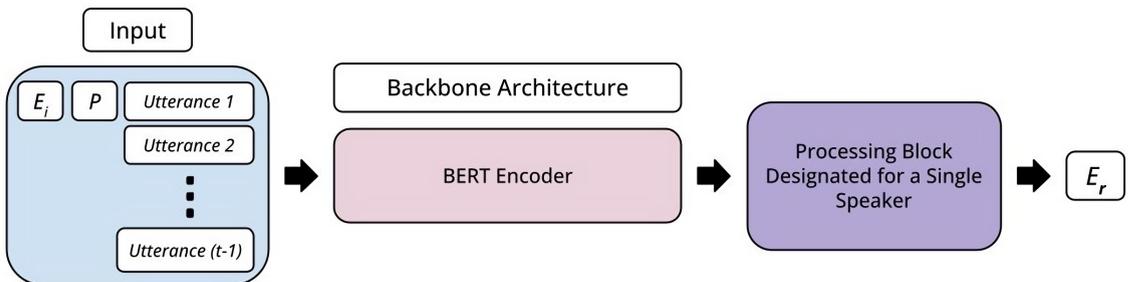


FIGURE 7.9: Personality-based Emotional Responses in Conversation using Single-Task (ST) Learning. E_i is the preceding emotions, P is the personality vector of the designated speaker, and E_r is the predicted emotional response for the upcoming utterance t with respect to the designated speaker.

7.3.2 Multi-Task (MT) Architecture

Unlike Single-Task (ST) architecture, where ST always predicted single subjective emotional responses, the Multi-Task (MT) predicts multiple subjective emotional responses (E_{r1}, \dots, E_{rn}), where n is the number of annotators, characterized by their unique personality information. The Multi-Task (MT) has multiple branches; one for each specific personality vector. Every branch refers to a single subjective emotional response (see Eq. 7.2). The MT architecture is illustrated in Fig. 7.10. The Multi-Task (MT) architecture consists of a single BERT encoder connected with multiple separate branches. Each branch is only designated to learn single subjective emotional responses. In a single training loop, the model only considers the dialogues that are associated with the same personality information. After getting the semantics representation of the dialogues, only a single branch that is associated with that personality information is active for further processing. The objective function is calculated using dialogue labels associated with that personality information and at a backpropagation step, only the selected branch and the BERT encoder hyperparameters are updated. Similarly, for a second training loop, dialogues that are associated with another personality information are selected and processed accordingly. This is iteratively done until all the available dialogues having unique personality information available in the dataset are processed. After every iteration, the hyperparameters of the BERT encoder are always updated. This is why the BERT encoder is shared with all the speakers.

$$\begin{aligned} E_{r1} &= MT(E_i|P_1, C) \\ &\vdots \\ E_{rn} &= MT(E_i|P_n, C) \end{aligned} \tag{7.2}$$

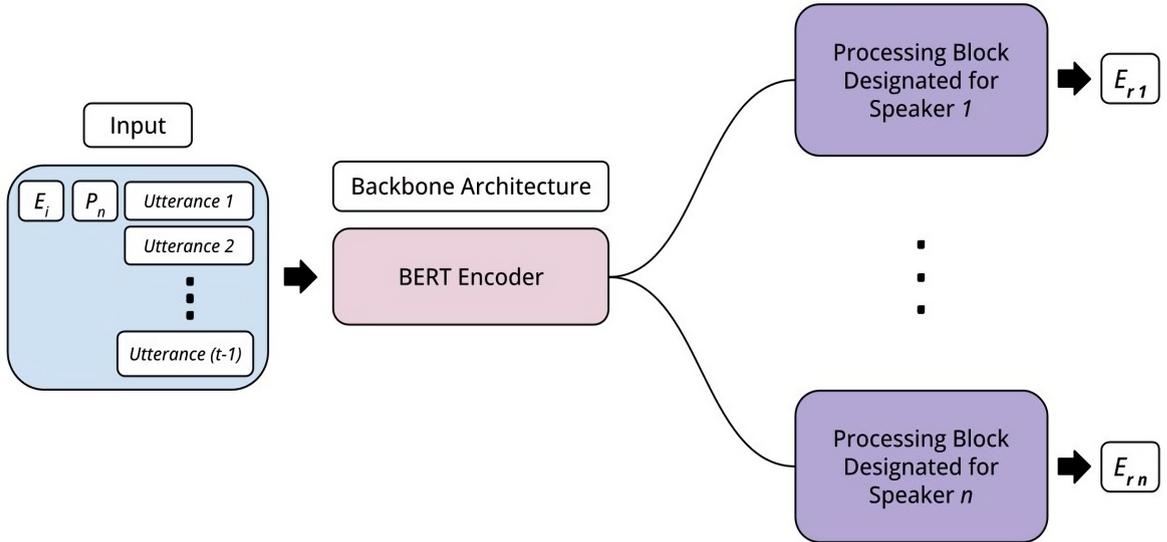


FIGURE 7.10: Personality-based Subjective Emotional Responses in Conversation using Multi-Task (MT) Learning. E_i is the preceding emotions, P_n is the personality vector of the n^{th} speaker, and E_{rn} is the predicted emotional response for the upcoming utterance t with respect to the n^{th} speaker

7.3.3 Implementation Details

Our Single-Task (ST) architecture is inspired by the state-of-the-art [210] approach for the PELD dataset. There are two changes between ours and the architecture from [210]. First, we fine-tuned the BERT-Base [143]) language processing model as compared to the RoBERTa [231]. Secondly, we applied the *Tanh* activation function on the personality-based emotional variations $T_{(VAD)}$ instead of the *Sigmoid* used in [210] for getting the emotional responses E_r . Both ST and MT architectures have the following main processing steps.

Preliminary Steps: Before feeding the dialogues and the personality information into the model, two preliminary steps need to be done. First, convert the categorical emotions into continuous emotional space known as VAD (Valence, Arousal, and Dominance) [232]. Each utterance in the dialogue is categorized into six basic emotions [233]: Anger, Disgust, Fear, Joy, Sadness, and Surprise. Russell et al. [234] proposed an analysis that is used to convert categorical emotions into the VAD emotional space (see Table 7.2). The VAD space indicates emotion

TABLE 7.2: Categorical emotions into VAD emotional space [234]

Emotion Category	Corresponding VAD Vector
Anger	(-0.51, 0.59, 0.25)
Disgust	(-0.60, 0.35, 0.11)
Fear	(-0.62, 0.82, -0.43)
Joy	(0.81, 0.51, 0.46)
Neutral	(0.00, 0.00, 0.00)
Sadness	(-0.63, -0.27, -0.33)
Surprise	(0.40, 0.67, -0.13)

intensity in three different dimensions, where Valence measures the positivity/negativity, Arousal the agitation/calmness, and Dominance the control/no-control.

Secondly, estimate the valence, arousal, and dominance expression of a speaker using personality information, i.e. big five (OCEAN) traits. For this, we use a temperament model that was developed by Mehrabian et al. [235]. The model is derived through a linear regression to show the VAD emotional scale of the personality traits as specified in the following equation:

$$\begin{aligned}
 P_V &= 0.21(E) + 0.59(A) + 0.19(N) \\
 P_A &= 0.15(O) + 0.30(A) - 0.57(N) \\
 P_D &= 0.25(O) + 0.17(C) + 0.60(E) - 0.32(A)
 \end{aligned} \tag{7.3}$$

Where, P_V is the personality-influenced valence, P_A is the personality-influenced arousal, and P_D represents the personality-influenced dominance emotional vectors. O for Openness, C for Conscientiousness, E for Extraversion, A for Agreeableness, and N for Neuroticism.

The dialogues in the PELD dataset belong to six different speakers named Chandler, Joey, Monica, Phoebe, Rachel, and Ross. Table 7.3 shows the personality traits of each speaker in the OCEAN format.

Contextual Understanding. In order to predict the emotional expression of the upcoming utterance t , it is necessary to understand the context of all

Chapter 7: Subjective Emotions using Personality Traits

TABLE 7.3: Personalities traits of speakers in PELD dataset

Speaker	Personality Traits (O,C,E,A,N) Vector
Chandler	[0.648, 0.375, 0.386, 0.58, 0.477]
Joey	[0.574, 0.614, 0.297, 0.545, 0.455]
Monica	[0.713, 0.457, 0.457, 0.66, 0.511]
Phoebe	[0.6, 0.48, 0.31, 0.46, 0.56]
Rachel	[0.635, 0.354, 0.521, 0.552, 0.469]
Ross	[0.722, 0.489, 0.6, 0.533, 0.356]

the preceding $t - 1$ utterances. For this, the BERT Base [143] model is fine-tuned to get the textual embeddings of $t - 1$ utterances. BERT is a famous pre-trained language model whose performance is widely validated in many natural language tasks. BERT encodes each utterance into a 768 – *dimensional* vector.

Learning Personality-based VAD Vector. After the conversion of the big five personality traits vector into the VAD emotional vector, still, the VAD values are not the true representation of emotions with respect to the data. The reason behind this is that the temperament model [235] was based on the analysis of 72 participants and hence represents the weights related to the data generated by these 72 participants. In order to use this VAD emotional vector in our experiments, it is necessary to learn the appropriate weights with respect to the underlying data. This is why a linear layer that transformed $P(VAD)$ to $P^l(VAD)$ is applied.

Contextual-based Emotional Variations. The dialogue context is one of the main factors in order to generate a certain emotion in the speaker while speaking an utterance [229]. Since the dialogues consist of multiple utterances, the emotional representation also changes with respect to each utterance. It means that the dialogue as a whole represents the transition of emotions from the first utterance to the last utterance. Similarly, in order to generate emotional responses for the next utterance, the preceding variations of emotions should be known. To compute this emotional variation, the dialogue context $C \in \{U_1, U_2, \dots, U_{(t-1)}\}$ is encoded into emotional space, i.e. $C_{(VAD)}$ (see Eq. 7.4).

$$\begin{aligned}
 C_{(t-1)} &= [B_r(U_1), B_r(U_2), \dots, B_r(U_{(t-1)})] \\
 C_{(VAD)} &= \text{LinearLayer}(C_{(t-1)})
 \end{aligned}
 \tag{7.4}$$

Where B_r is the BERT-Base encoder, $C_{(t-1)}$ is the contextual semantics of preceding utterances, and $C_{(VAD)}$ is the context-based emotional variations presented in the preceding utterances.

Personality-based Emotional Variations. After obtaining the weighting parameters of the personality $P^l(VAD)$ and the contextual-based emotional variations in preceding utterances, i.e. $C_{(VAD)}$, the personality-influenced emotional variations are generated by the sum of two different VAD vectors: the first VAD vector represents the initial emotions ($E_{i(VAD)}$) and the second VAD vector is the contextual-based emotional variation $C_{(VAD)}$ affected by the personality $P^l(VAD)$ vectors (see Eq. 7.5).

$$T_{(VAD)} = E_{i(VAD)} + P^l_{(VAD)} * C_{(VAD)}
 \tag{7.5}$$

Where $T_{(VAD)}$ is the personality-based emotion variations, $E_{i(VAD)}$ is the initial emotions, and $C_{(VAD)}$ is the emotional variations due to the context.

Response Emotions: To generate the subjective emotional responses, the personality-based emotional transition $T_{(VAD)}$ is combined with the personality vector $P^l(VAD)$ of a speaker and the preceding $C_{(t-1)}$ utterances. Lastly, we feed this concatenated vector to a linear layer to transform it into a probability distribution on the discrete emotion category. The output E_r is the response emotion that has the largest probability (see Eq. 7.6).

$$\begin{aligned}
 L_1 &= [T_{(VAD)}, P^l_{(VAD)}, C_{(t-1)}] \\
 E_r &= \text{OutputLayer}(L_1)
 \end{aligned}
 \tag{7.6}$$

Where L_1 is the concatenation of personality-based emotional variations, personality-based emotions, and contextual semantics of preceding utterances. E_r is the response emotions.

7.4 Experiments and Results

7.4.1 PELD Dataset

The dataset consists of dialogues between two speakers. The dialogues were taken from a famous TV series named Friends. The dataset only considered the conversation from six speakers: Chandler, Joey, Monica, Phoebe, Rachel, and Ross. Each dyadic conversation has three utterances. Utterance 1 belongs to the first speaker, utterance 2 is from the second speaker, and utterance 3 is again from the first speaker. Table 7.4 shows the number of dialogues in Train, Val, and Test sets with respect to each individual speaker. The number of dialogues per speaker is approximately the same except for speaker Phoebe, which has fewer dialogues than the others. To get more insight into the data, Fig. 7.11 mapped the number of dialogues per emotion class with respect to each speaker. This Fig. 7.11 shows two important statistics of the data. Firstly, the data is highly biased toward the Neutral emotion class for each speaker. Secondly, each speaker has approximately the same number of dialogues per emotion class except Phoebe, which has slightly fewer dialogues for Anger and Fear emotions. Overall, the data supports modeling each speaker’s emotional response separately.

TABLE 7.4: Number of dialogues per each speaker in PELD dataset

Speaker	Train	Val	Test	Total
Chandler	880	97	108	1085
Joey	912	109	102	1123
Monica	850	94	107	1051
Phoebe	782	87	103	972
Rachel	921	112	123	1156
Ross	928	87	108	1123

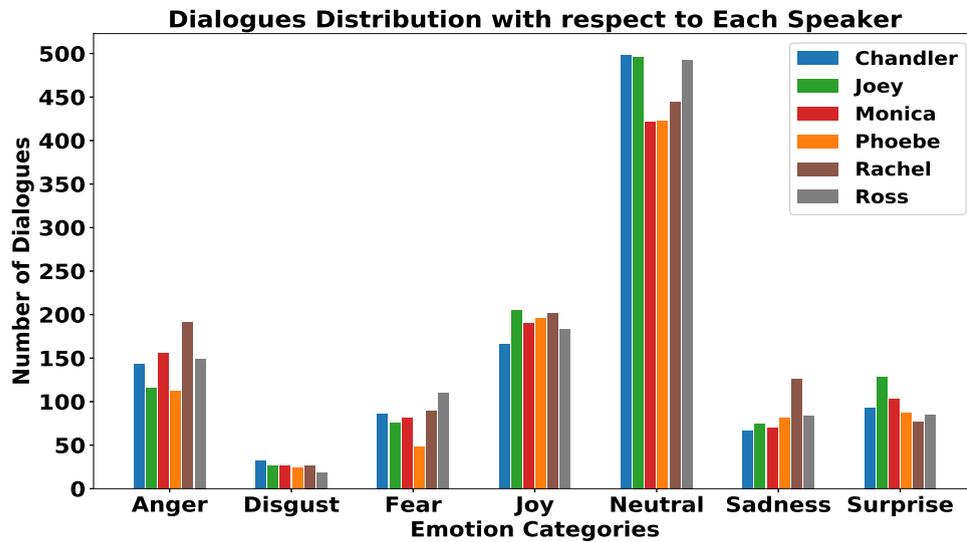


FIGURE 7.11: Number of dialogues per each emotion category with respect to each speaker

7.4.2 Results

In the experiments, we first follow the simplest approach of emotion classification, i.e. predicting emotional responses into 3 valence categories (negative, neutral, and positive). Later, we predict emotional responses in 7 categories (anger, disgust, fear, joy, neutral, sadness, and surprise). Both Single-Task (ST) and Multi-Task (MT) approaches were tested for predicting emotional responses in these 3 and 7 categories. The performance of each ST and MT was evaluated by the F1-score for each emotional category and two aggregated evaluation measures: the macro average (m-avg) and the weighted average (w-avg) of the F-score values. The macro-average F1 score (m-avg) is computed using the arithmetic mean (also known as unweighted mean) of all the per-category F1 scores (see Eq. 7.7). The weighted-average F1 score (w-avg) is calculated by taking the mean of all per-category F1 scores (see Eq. 7.8) using a weight that depends on the number of true labels of each category. The weight of the particular category is calculated by using Eq. 7.9.

$$macro \ average = \frac{1}{n} \sum_{i=1}^n F1_{category_i} \quad (7.7)$$

$$weighted \ average = \frac{1}{n} \sum_{i=1}^n F1_{category_i} * Weight_i \quad (7.8)$$

$$Weight_i = \frac{TP_i + FP_i}{\sum_{i=1}^n (TP_i + FP_i)} \quad (7.9)$$

Where TP_i and FP_i represent the True Positive and False Positive of a particular category respectively.

7.4.2.1 Predicting Response Emotions with 3 Categories

We first group 7 emotions into 3 valence categories: positive, negative, and neutral. Specifically, anger, disgust, fear, and sadness emotions are considered negative, whereas joy and surprise are considered positive. The neutral category only considers the neutral emotion (see Table 7.5). After the conversion, Table 7.6 shows the number of samples per emotion category with respect to each speaker.

For Single-Task (ST) approach, each time a single speaker’s personality traits are considered in predicting subjective emotional responses for the next utterance. Since the PELD dataset consists of conversations belonging to six different speakers, six ST models were trained independently, one for each personality traits vector (see Table 7.7).

On the other hand, for Multi-Task (MT), each separate branch is reserved for every single personality traits vector and these separate branches are jointly trained. Therefore, a single model needs to be trained for MT. The performance of

Chapter 7: Subjective Emotions using Personality Traits

TABLE 7.5: 7 emotions to 3 emotions categories conversion

7 Emotions	3 Emotions
anger	negative
sadness	negative
neutral	neutral
joy	positive
surprise	positive
fear	negative
disgust	negative

TABLE 7.6: Number of dialogues for the Train, Val, and Test set for 3 emotions categories with respect to each speaker

Speaker	Emotion	Train	Val	Test	Total
Chandler	Negative	265	31	32	328
	Neutral	409	39	50	498
	Positive	206	27	26	259
Joey	Negative	235	33	26	294
	Neutral	403	45	48	496
	Positive	274	31	28	334
Monica	Negative	267	38	31	336
	Neutral	338	38	46	422
	Positive	245	18	30	293
Phoebe	Negative	224	17	24	265
	Neutral	327	48	48	423
	Positive	231	22	30	283
Rachel	Negative	341	47	45	433
	Neutral	362	37	45	444
	Positive	218	28	33	279
Ross	Negative	291	29	34	354
	Neutral	420	30	43	493
	Positive	217	20	31	268

each branch using MT learning is presented in Table 7.8. For comparison with the state-of-the-art, we compute the mean values of each evaluation measure for all six speakers. This is because there is no prior work on separate modeling of each subjective emotional response using personality traits. The comparison of ST, MT, and state-of-the-art [210] approaches is presented in Table 7.9. The results show that the MT approach has got considerable improvement with respect to the ST approach and [210] for predicting the subjective emotional response in 3 valence categories.

Chapter 7: Subjective Emotions using Personality Traits

TABLE 7.7: Emotional responses for the next utterance in 3 categories using Single-Task (ST) learning. F1-score is measured to predict negative, neutral, and positive emotion categories with respect to each speaker. The macro ($m - avg$) and weighted average ($w - avg$) are also measured.

Speaker	Negative	Neutral	Positive	m-avg	w-avg
Chandler	0.422	0.531	0.280	0.434	0.485
Joey	0.291	0.505	0.550	0.449	0.478
Monica	0.426	0.340	0.470	0.412	0.396
Phoebe	0.470	0.518	0.250	0.413	0.429
Rachel	0.438	0.535	0.589	0.520	0.531
Ross	0.363	0.384	0.373	0.378	0.374
Mean	0.401	0.468	0.418	0.434	0.448

TABLE 7.8: Emotional responses for the next utterance in 3 categories using Multi-Task (MT) learning. F1-score is measured to predict negative, neutral, and positive emotion categories with respect to each speaker. The macro ($m - avg$) and weighted average ($w - avg$) are also measured.

Speaker	Negative	Neutral	Positive	m-avg	w-avg
Chandler	0.523	0.642	0.468	0.544	0.572
Joey	0.483	0.594	0.459	0.512	0.522
Monica	0.424	0.315	0.495	0.411	0.410
Phoebe	0.580	0.321	0.346	0.416	0.418
Rachel	0.505	0.466	0.461	0.477	0.477
Ross	0.541	0.510	0.390	0.480	0.493
Mean	0.509	0.475	0.436	0.473	0.482

TABLE 7.9: Comparison of MT and ST approaches with SOTA for predicting emotional responses in 3 categories.

Method	Negative	Neutral	Positive	m-avg	w-avg
Wen et al. [210]	0.492	0.474	0.327	0.431	0.445
Single-Task (ST)	0.401	0.468	0.418	0.434	0.448
Multi-Task (MT)	0.509	0.475	0.436	0.473	0.482

7.4.2.2 Predicting Response Emotions with 7 Categories

The same procedure for training ST and MT has been adopted for predicting the emotional responses into 3 categories. The only difference is now the

Chapter 7: Subjective Emotions using Personality Traits

emotional responses of the next utterance are classified into 7 categories: anger, disgust, fear, joy, neutral, sadness, and surprise. Firstly, Single-Task (ST) learning is used to train separate models for each speaker independently to predict speaker-specific emotional responses. Table 7.10 shows the performance of each individual ST model with respect to each speaker.

Secondly, Multi-Task (MT) approach is applied to predict each speaker's emotional responses separately in a single model. The results of the Multi-Task (MT) approach are shown in Table 7.11. The model performance is measured with respect to each individual speaker. The results show that few speakers have better results in the prediction of a few emotions and others have a bit lower in predicting the same emotion category. For example, Phoebe and Rachel have got better performance in predicting Anger emotions as compared to others. Similarly, Chandler has better performance in predicting Neutral emotion as compared to other speakers. The possible reasons are as follows. One reason might be the number of samples for that specific emotion class. Table 7.12 shows the number of dialogues taken for the Train, Val, and Test set per emotion category with respect to each speaker. In MT, when training a model with an individual sub-branch the model is actually learning the emotional responses that are specific to each individual speaker. This helps to converge the hyperparameters of the shared part as well as the individual part better as compared to the combined modeling of all emotional responses. Another possible reason is the dialogue context. Maybe one speaker has a relatively clear context (strong keywords for emotions) for inducing a particular emotion and the other speaker does not have a clear context (weak keywords for emotions) for that specific emotion class.

As done in the previous experiments with 3-categories, we compute the mean values to compare our approaches with the state-of-the-art [210]. Table 7.13 shows the comparison of Single-Task (ST) and Multi-Task (MT) with state-of-the-art [210].

The Multi-Task (MT) approach has surpassed the state-of-the-art [210] and Single-Task (ST) in the majority (see Table 7.13). Except for anger, neutral,

Chapter 7: Subjective Emotions using Personality Traits

TABLE 7.10: Predicting 7 categories emotional responses for the next utterance in a dialogue using Single-Task (ST) learning. F1-score is measured to predict each emotion category with respect to each speaker.

Speaker	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	m-avg	w-avg
Chandler	0.0	0.0	0.181	0.266	0.586	0.153	0.133	0.188	0.381
Joey	0.200	0.105	0.0	0.146	0.553	0.0	0.279	0.183	0.332
Monica	0.235	0.0	0.250	0.162	0.482	0.200	0.125	0.204	0.325
Phoebe	0.344	0.0	0.142	0.250	0.574	0.0	0.250	0.233	0.351
Rachel	0.444	0.0	0.0	0.456	0.568	0.296	0.347	0.301	0.418
Ross	0.133	0.0	0.206	0.321	0.435	0.0	0.320	0.202	0.308
Mean	0.226	0.017	0.129	0.266	0.533	0.108	0.242	0.218	0.352

TABLE 7.11: Predicting 7 categories emotional responses for the next utterance in a dialogue using Multi-Task (MT) learning. F1-score is measured to predict each emotion category with respect to each speaker.

Speaker	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	m-avg	w-avg
Chandler	0.375	0.0	0.125	0.388	0.560	0.153	0.250	0.305	0.422
Joey	0.250	0.333	0.166	0.294	0.520	0.200	0.272	0.324	0.410
Monica	0.250	0.0	0.125	0.235	0.521	0.250	0.153	0.304	0.410
Phoebe	0.300	0.0	0.166	0.260	0.520	0.200	0.142	0.304	0.412
Rachel	0.272	0.400	0.142	0.263	0.533	0.272	0.071	0.308	0.412
Ross	0.230	0.0	0.250	0.277	0.534	0.250	0.153	0.302	0.412
Mean	0.279	0.122	0.162	0.286	0.531	0.221	0.163	0.308	0.413

and surprise emotions, the Multi-Task (MT) has got significant gains in disgust, fear, joy, and sadness emotion categories with respect to ST and [210] approaches. With the performance increment in the individual categories, the macro-average of MT also obtained significant improvement (see Table 7.13). The state-of-the-art approach has a higher score only for anger emotion and macro average measure as compared to Single-Task (ST) and Multi-Task (MT). For the neutral emotional class, the ST and MT have got approximately the same score, the difference is so minor that it can even be neglected. Furthermore, the Multi-Task (MT) learning also helps to tackle the class imbalance problem. Table 7.12 shows that the emotion categories disgust, fear, sadness, and surprise have a lower number of samples as compared to others (anger, neutral, and joy) and the results show that the MT approach gives benefits in learning emotional patterns for those categories with respect to each individual speaker.

Chapter 7: Subjective Emotions using Personality Traits

TABLE 7.12: Number of dialogues for the Train, Val, and Test for emotions with respect to each speaker

Speaker	Emotion	Train	Val	Test	Total
Chandler	Anger	118	17	8	143
	Disgust	27	2	3	32
	Fear	72	6	8	86
	Joy	131	17	18	166
	Neutral	409	39	50	498
	Sadness	48	6	13	67
	Surprise	75	10	8	93
Joey	Anger	94	10	12	116
	Disgust	19	5	3	27
	Fear	57	13	6	76
	Joy	167	21	17	205
	Neutral	403	45	48	496
	Sadness	65	5	5	75
	Surprise	107	10	11	128
Monica	Anger	131	13	12	156
	Disgust	20	4	3	27
	Fear	62	13	8	82
	Joy	160	13	17	190
	Neutral	338	38	46	422
	Sadness	54	8	8	70
	Surprise	85	5	13	103
Phoebe	Anger	98	4	10	112
	Disgust	18	3	3	24
	Fear	38	4	6	48
	Joy	158	15	23	196
	Neutral	327	48	48	423
	Sadness	70	6	5	81
	Surprise	73	7	7	87
Rachel	Anger	148	21	22	191
	Disgust	17	4	5	27
	Fear	74	9	7	90
	Joy	167	16	19	202
	Neutral	362	37	45	444
	Sadness	102	13	11	126
	Surprise	51	12	14	77
Ross	Anger	122	14	13	149
	Disgust	16	2	1	19
	Fear	89	13	8	110
	Joy	151	14	18	183
	Neutral	420	30	43	493
	Sadness	64	8	12	84
	Surprise	66	6	13	85

Chapter 7: Subjective Emotions using Personality Traits

TABLE 7.13: Comparison of MT and ST approaches with [210] for predicting emotional responses in 7 categories.

Method	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	m-avg	w-avg
Wen et al. [210]	0.320	0.070	0.140	0.198	0.528	0.155	0.098	0.203	0.424
Single-Task (ST)	0.226	0.017	0.129	0.266	0.533	0.108	0.242	0.218	0.352
Multi-Task (MT)	0.279	0.122	0.162	0.286	0.531	0.221	0.163	0.308	0.413

7.4.2.3 Predicting Response Emotions without Personality

The results presented in Table 7.13 and Table 7.9 show clearly that the Multi-Task (MT) learning has a significant impact on generating subjective emotional responses of a speaker. However, the results do not reveal the stand-alone contribution of personality information in order to generate subjective emotional responses of a speaker. To figure this out, we decided to perform the Multi-Task (MT) approach without considering the personality information. We removed the influence of personality information in Eq. 7.5 (i.e. $T_{(VAD)} = E_{i(VAD)} + P_{(VAD)}^l * C_{(VAD)}$) and the new equation for predicting response emotions without personality is $T_{(VAD)} = E_{i(VAD)} * C_{(VAD)}$.

Table 7.14 and Table 7.15 show the results of the Multi-Task (MT) approach without considering the personality information in order to predict emotional responses with respect to each speaker for 3 and 7 categories respectively. Tables 7.14 and 7.15 have many 0 values (i.e. F1-score) in predicting emotional responses for both 3 and 7 categories. This is due to the very less number of dialogues for each emotional category with respect to each speaker (see Table 7.12). For example, speaker Chandler has only 3 test dialogues for disgust emotion. For the comparison between these two MT models (with and without personality traits), only the mean value of each emotion category is considered (see Table 7.16 and Table 7.17) for 3 and 7 categories respectively.

The results presented in Table 7.16 and Table 7.17 show that the personality information of speakers plays a significant role in predicting emotions responses for all categories. The results in Table 7.17 also reveal that in predicting a neutral response emotion, the model without personality information has got a slightly

Chapter 7: Subjective Emotions using Personality Traits

TABLE 7.14: Predicting 3 categories emotional responses without personality vector for the next utterance in a dialogue using Multi-Task (MT) learning. F1-score is measured to predict each emotion category with respect to each speaker.

Speaker	Negative	Neutral	Positive	m-avg	w-avg
Chandler	0.0	0.690	0.0	0.230	0.363
Joey	0.0	0.0	0.496	0.165	0.164
Monica	0.0	0.634	0.0	0.211	0.294
Phoebe	0.0	0.591	0.0	0.197	0.248
Rachel	0.0	0.0	0.496	0.165	0.164
Ross	0.373	0.534	0.0	0.302	0.332
Mean	0.062	0.408	0.165	0.212	0.261

TABLE 7.15: Predicting 7 categories emotional responses without personality for the next utterance in a dialogue using Multi-Task (MT) learning. F1-score is measured to predict each emotion category with respect to each speaker.

Speaker	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	m-avg	w-avg
Chandler	0.148	0.0	0.0	0.121	0.492	0.0	0.125	0.126	0.253
Joey	0.181	0.0	0.0	0.090	0.590	0.0	0.0	0.123	0.323
Monica	0.133	0.0	0.0	0.225	0.470	0.0	0.0	0.118	0.262
Phoebe	0.344	0.0	0.0	0.049	0.595	0.0	0.0	0.141	0.358
Rachel	0.266	0.0	0.0	0.264	0.529	0.0	0.0	0.151	0.264
Ross	0.322	0.0	0.0	0.136	0.566	0.0	0.0	0.146	0.368
Mean	0.232	0.0	0.0	0.148	0.540	0.0	0.020	0.134	0.304

higher F1-score compared to the model with personality information. The reason behind this is that the majority of psychologists consider that the neutral state is not an affective state because the neutral state does not evoke a strongly felt reaction [70, 236–238]. With this basis, it makes sense why the personality information did not contribute to predicting the subjective neutral emotional response of a speaker. This phenomenon can only be observed in predicting emotional responses in more diverse categories (anger, disgust, fear, joy, neutral, sadness, and surprise) as compared to the simple ones, i.e. negative, neutral, and positive categories.

Another observation from the results (Tables 7.16 and 7.17) about neutral emotion is that the MT without personality information model is more biased towards neutral emotion when predicting emotional responses in 7 categories as

Chapter 7: Subjective Emotions using Personality Traits

compared to 3 categories. This is why MT without the personality information model has got slightly higher performance, i.e. 0.540 in predicting Neutral emotion as compared to MT with the personality information model, i.e. 0.531 (see Table 7.17). But for predicting emotional responses in 3 categories, MT with personality information has got higher performance than MT without personality (see Table 7.16).

TABLE 7.16: Comparison of MT with and without considering personality information for predicting 3 categories emotional responses. F1-score is measured to predict each emotion category with respect to each speaker.

Method	Negative	Neutral	Positive	m-avg	w-avg
MT without personality	0.062	0.408	0.165	0.212	0.261
MT with personality	0.509	0.475	0.436	0.473	0.482

TABLE 7.17: Comparison of MT with and without considering personality information for predicting 7 categories emotional responses. F1-score is measured to predict each emotion category with respect to each speaker.

Method	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	m-avg	w-avg
MT without personality	0.232	0.0	0.0	0.148	0.540	0.0	0.020	0.134	0.304
MT with personality	0.279	0.122	0.162	0.286	0.531	0.221	0.163	0.308	0.413

7.5 Qualitative Analysis

This section presents a qualitative analysis of our Multi-Task (MT) approach from multiple perspectives. For this, we selected a single dialogue from each speaker that is wrongly predicted by the state-of-the-art approach.

(i) **Multi-Task (MT) versus State-of-the-art:** In this analysis, we get the predictions made by our Multi-Task (MT) approach for those dialogues that were wrongly predicted by the state-of-the-art approach with respect to each speaker. Tables 7.18 and 7.19 present the results for the 3 and 7 categories of emotion classification problems respectively. The results show that the Multi-Task (MT) learning performs better in predicting subjective emotional responses as compared to the state-of-the-art approach.

TABLE 7.18: A single dialogue from each individual speaker for 3-categories emotion classification. GT stands for Ground Truth, SOTA is for state-of-the-art, and MT is Multi-Task. The red color indicates the wrong prediction and the green indicates the correct prediction. *U1*, *U2*, and *U3* represent Utterances 1, Utterances 2, and Utterances 3 respectively.

Speaker	Dialogue	GT	SOTA [210]	MT
Chandler	<i>U1</i> : Hey, I've been honing! <i>U2</i> : What was with the dishes? <i>U3</i> : Oh, uh..	Negative	Positive	Negative
Joey	<i>U1</i> : I feel like I can't do stuff! <i>U2</i> : What kinda stuff? <i>U3</i> : Will you grow up? I'm not talking about sexy stuff, but, like, when I'm cooking naked.	Negative	Neutral	Negative
Monica	<i>U1</i> : How does she do that? <i>U2</i> : I can not sleep in public places. <i>U3</i> : Would you look at her? she is so peaceful.	Positive	Neutral	Neutral
Phoebe	<i>U1</i> : Left! Thank you. <i>U2</i> : You're welcome. <i>U3</i> : Ross!	Positive	Neutral	Positive
Rachel	<i>U1</i> : Yeah, it's a real shame you can't make it to that one-woman show tonight. <i>U2</i> : Oh, I'd love to, but I gotta get up so early the next day and so, you know me, work comes first <i>U3</i> : Oh, yeah, yeah, yeah...	Neutral	Positive	Neutral
Ross	<i>U1</i> : This is my wedding! <i>U2</i> : All right, you know what? We are really late! Let's go! Let's go! Let's go! <i>U3</i> : Fine, you will- you will watch it in a video when we get back.	Negative	Neutral	Positive

Chapter 7: Subjective Emotions using Personality Traits

TABLE 7.19: A single dialogue from each individual speaker for 7-categories emotion classification. GT stands for Ground Truth, SOTA is for state-of-the-art, and MT is Multi-Task. The red color indicates the wrong prediction and the green indicates the correct prediction. *U1*, *U2*, and *U3* represent Utterances 1, Utterances 2, and Utterances 3 respectively.

Speaker	Dialogue	GT	SOTA [210]	MT
Chandler	<i>U1</i> : Do I ever. <i>U2</i> : Chris says they're closing down the bar <i>U3</i> : No way!	Surprise	Neutral	Surprise
Joey	<i>U1</i> : Here, I need to borrow some moisturizer <i>U2</i> : For what? <i>U3</i> : What do you think? Today's the big day!	Joy	Neutral	Surprise
Monica	<i>U1</i> : Hey, did you pick a roommate? <i>U2</i> : You betcha! <i>U3</i> : Is it the Italian guy?	Neutral	Joy	Neutral
Phoebe	<i>U1</i> : Because I am dumping him today <i>U2</i> : What? you said he was sweet! <i>U3</i> : He	Surprise	Neutral	Sadness
Rachel	<i>U1</i> : But I told you, I didn't have the time! <i>U2</i> : Yeah, well you never have the time. I mean, I didn't feel like I even have a girlfriend anymore, Rachel. <i>U3</i> : Wh, Ross what do you want from me?	Anger	Fear	Anger
Ross	<i>U1</i> : I loved this place! To tell you the truth, I wish I didn't have to move. <i>U2</i> : Uhh, are you saying that you're not entirely happy about this? <i>U3</i> : Well, I mean if uh, if Emily gave me a choice	Sadness	Anger	Sadness

(ii) **Multi-Task (MT) without Personality Information:** In our previous analysis, it is clear that personality information when used in Multi-Task (MT) manner is more effective in predicting subjective emotional responses. To get more deep insight into the effectiveness of personality information in predicting subjective emotional responses, we decided to get the predictions of our Multi-Task (MT) approach without considering the personality information of each speaker. For this, the same dialogues (see Table 7.18) with respect to each speaker were used. We can see in Table 7.20 and Table 7.21 that personality information plays a significant role in predicting subjective emotional responses in most emotional categories. However, the personality information does not contribute much to predicting neutral emotions as a response.

TABLE 7.20: A single dialogue from each individual speaker for 3-categories emotion classification. GT stands for Ground Truth, MT-W/P means Multi-Task with personality and MT-WO/P means Multi-Task without personality information respectively. The red color indicates the wrong prediction and the green indicates the correct prediction. *U1*, *U2*, and *U3* represent Utterances 1, Utterances 2, and Utterances 3 respectively.

Speaker	Dialogue	GT	MT-W/P	MT-WO/P
Chandler	<i>U1</i> : Hey, I've been honing! <i>U2</i> : What was with the dishes? <i>U3</i> : Oh, uh..	Negative	Negative	Neutral
Joey	<i>U1</i> : I feel like I can't do stuff! <i>U2</i> : What kinda stuff? <i>U3</i> : Will you grow up? I'm not talking about sexy stuff, but, like, when I'm cooking naked.	Negative	Negative	Positive
Monica	<i>U1</i> : Phoebe why don't you just call her? You obviously want to. <i>U2</i> : You think you know me so well. <i>U3</i> : Well, don't wanna?	Neutral	Neutral	Neutral
Phoebe	<i>U1</i> : Left! Thank you. <i>U2</i> : You're welcome. <i>U3</i> : Ross!	Positive	Positive	Neutral
Rachel	<i>U1</i> : Yeah, it's a real shame you can't make it to that one-woman show tonight. <i>U2</i> : Oh, I'd love to, but I gotta get up so early the next day and so, you know me, work comes first <i>U3</i> : Oh, yeah, yeah, yeah...	Neutral	Neutral	Negative
Ross	<i>U1</i> : No you do know what, you're not gonna suck me into this. <i>U2</i> : Oh sure I am, because you always have to be right. <i>U3</i> : I do not always have to be okay, okay.	Negative	Negative	Neutral

TABLE 7.21: A single dialogue from each individual speaker for 7-categories emotion classification. GT stands for Ground Truth, MT-W/P means Multi-Task with personality and MT-WO/P means Multi-Task without personality information respectively. The red color indicates the wrong prediction and the green indicates the correct prediction. *U1*, *U2*, and *U3* represent Utterances 1, Utterances 2, and Utterances 3 respectively.

Speaker	Dialogue	GT	MT-W/P	MT-WO/P
Chandler	<i>U1</i> : Do I ever. <i>U2</i> : Chris says they're closing down the bar <i>U3</i> : No way!	Surprise	Surprise	Neutral
Joey	<i>U1</i> : You liked it? <i>U2</i> : You really liked it ? <i>U3</i> : Oh-oh-oh	Surprise	Surprise	Joy
Monica	<i>U1</i> : Hey, did you pick a roommate? <i>U2</i> : You betcha! <i>U3</i> : Is it the Italian guy?	Neutral	Neutral	Neutral
Phoebe	<i>U1</i> : Okay. Does it have to do with Ross and Rachel? <i>U2</i> : No <i>U3</i> : Does it have to do with Joey?	Neutral	Neutral	Surprise
Rachel	<i>U1</i> : But I told you, I didn't have the time! <i>U2</i> : Yeah, well you never have the time. I mean, I didn't feel like I even have a girlfriend anymore, Rachel. <i>U3</i> : Wh, Ross what do you want from me?	Anger	Anger	Fear
Ross	<i>U1</i> : I loved this place! To tell you the truth, I wish I didn't have to move. <i>U2</i> : Uhh, are you saying that you're not entirely happy about this? <i>U3</i> : Well, I mean if uh, if Emily gave me a choice	Sadness	Sadness	Fear

(iii) **Getting all Speaker’s Emotional Responses against a Single Dialogue in Multi-Task (MT) Approach:** In this analysis, we consider a single dialogue (belonging to a specific speaker) and get the predicted emotional responses from all speakers. The reason behind this analysis is to understand how each separate branch in Multi-Task (MT) behaves when we feed dialogues that do not belong to them. For this, a pre-trained Multi-Task (MT) model is used for testing. We considered the emotional responses of each dialogue from all the speakers, i.e. all separate branches of MT. The dialogues are the same as we considered in our previous analysis (see Table 7.18). The results (see Tables 7.22 and 7.23) show that when a dialogue (belongs to a specific speaker) passes through a different speaker’s branch produces an incorrect result. This is another indicator that personality information is more effective when used in Multi-Task (MT) approach for predicting subjective emotional responses. Furthermore, the results also show a loose connection between emotional responses and contextual information. For example, the first dialogue is between speakers Chandler and Ross, and when passing the same dialogue to all separate branches to get the emotional responses to the next utterance, the branches designated for these two speakers predicted the same emotional responses. However, in the majority of the cases, this phenomenon is not observed.

Chapter 7: Subjective Emotions using Personality Traits

TABLE 7.22: The same dialogue is passed through all the MT branches. Each branch is trained with respect to each speaker’s personality traits. The outputs represent the predicted emotional responses in 3-classes from all the speakers against the same dialogue. GT stands for Ground Truth, red color indicates the wrong prediction and the green indicates the correct prediction. *U1*, *U2*, and *U3* represent Utterances 1, Utterances 2, and Utterances 3 respectively.

Speaker	Dialogue	GT	Chandler	Joey	Monica	Phoebe	Rachel	Ross
Chandler	<i>U1</i> : Hey, I’ve been honing! <i>U2</i> : What was with the dishes? <i>U3</i> : Oh, uh..	Negative	Negative	Neutral	Neutral	Positive	Neutral	Negative
Joey	<i>U1</i> : I feel like I can’t do stuff! <i>U2</i> : What kinda stuff? <i>U3</i> : Will you grow up? I’m not talking about sexy stuff, but, like, when I’m cooking naked.	Negative	Positive	Negative	Positive	Neutral	Positive	Neutral
Monica	<i>U1</i> : How does she do that? <i>U2</i> : I can not sleep in public places. <i>U3</i> : Would you look at her? she is so peaceful.	Positive	Negative	Positive	Neutral	Negative	Neutral	Positive
Phoebe	<i>U1</i> : Left! Thank you. <i>U2</i> : You’re welcome. <i>U3</i> : Ross!	Positive	Neutral	Neutral	Neutral	Positive	Neutral	Neutral
Rachel	<i>U1</i> : Yeah, it’s a real shame you can’t make it to that one-woman show tonight. <i>U2</i> : Oh, I’d love to, but I gotta get up so early the next day and so, you know me, work comes first <i>U3</i> : Oh, yeah, yeah, yeah...	Neutral	Negative	Negative	Negative	Negative	Neutral	Positive
Ross	<i>U1</i> : This is my wedding! <i>U2</i> : All right, you know what? We are really late! Let’s go! Let’s go! Let’s go! <i>U3</i> : Fine, you will- you will watch it in a video when we get back.	Negative	Positive	Positive	Neutral	Neutral	Positive	Positive

Chapter 7: Subjective Emotions using Personality Traits

TABLE 7.23: The same dialogue is passed through all the MT branches. Each branch is trained with respect to each speaker’s personality traits. The outputs represent the predicted emotional responses in 7-classes from all the speakers against the same dialogue. GT stands for Ground Truth, red color indicates the wrong prediction and the green indicates the correct prediction. *U1*, *U2*, and *U3* represent Utterances 1, Utterances 2, and Utterances 3 respectively.

Speaker	Dialogue	GT	Chandler	Joey	Monica	Phoebe	Rachel	Ross
Chandler	<i>U1</i> : Hey, I’ve been honing! <i>U2</i> : What was with the dishes? <i>U3</i> : Oh, uh..	Surprise	Surprise	Neutral	Neutral	Joy	Neutral	Neutral
Joey	<i>U1</i> : I feel like I can’t do stuff! <i>U2</i> : What kinda stuff? <i>U3</i> : Will you grow up? I’m not talking about sexy stuff, but, like, when I’m cooking naked.	Surprise	Neutral	Surprise	Joy	Fear	Neutral	Joy
Monica	<i>U1</i> : Phoebe why don’t you just call her? You obviously want to. <i>U2</i> : You think you know me so well. <i>U3</i> : Well, don’t wanna?	Neutral	Joy	Neutral	Neutral	Surprise	Neutral	Neutral
Phoebe	<i>U1</i> : Left! Thank you. <i>U2</i> : You’re welcome. <i>U3</i> : Ross!	Neutral	Surprise	Neutral	Neutral	Neutral	Neutral	Neutral
Rachel	<i>U1</i> : Yeah, it’s a real shame you can’t make it to that one-woman show tonight. <i>U2</i> : Oh, I’d love to, but I gotta get up so early the next day and so, you know me, work comes first <i>U3</i> : Oh, yeah, yeah, yeah...	Anger	Fear	Anger	Neutral	Neutral	Anger	Neutral
Ross	<i>U1</i> : No you do know what, you’re not gonna suck me into this. <i>U2</i> : Oh sure I am, because you always have to be right. <i>U3</i> : I do not always have to be okay, okay.	Anger	Neutral	Neutral	Neutral	Neutral	Neutral	Anger

(iv) **Replace Personality Information of Speakers in Multi-Task**

(MT) Approach: A pre-trained Multi-Task (MT) model is used for this analysis. During testing, we replace the personality information of the speaker S_i with another speaker S_j but pass through the same branch that is designated for that speaker S_i in Multi-Task (MT) model to get the emotional responses. For example, the first dialogue from Table 7.24 belongs to speaker Chandler, and during testing, we replace Chandler’s personality with Joey’s personality but use the same branch that is designated for speaker Chandler to get the emotional responses. Tables 7.24 and 7.25 show that when replacing the personalities of speakers then the emotional responses are also affected by this change except for the neutral emotion.

TABLE 7.24: Emotional responses of speakers when replacing their personalities with other speakers. GT stands for Ground Truth, MT stands for Multi-Task, and Replace Personality represents the speaker whose personality is used to replace the original speaker’s personality. The red color indicates the wrong prediction and the green indicates the correct prediction. $U1$, $U2$, and $U3$ represent Utterances 1, Utterances 2, and Utterances 3 respectively. The predicted emotional responses are in 3 classes.

Speaker	Dialogue	GT	Replace Personality	MT
Chandler	$U1$: Hey, I’ve been honing! $U2$: What was with the dishes? $U3$: Oh, uh..	Negative	Joey	Neutral
Joey	$U1$: I feel like I can’t do stuff! $U2$: What kinda stuff? $U3$: Will you grow up? I’m not talking about sexy stuff, but, like, when I’m cooking naked.	Negative	Chandler	Positive
Monica	$U1$: Phoebe why don’t you just call her? You obviously want to. $U2$: You think you know me so well. $U3$: Well, don’t wanna?	Neutral	Phoebe	Neutral
Phoebe	$U1$: Left! Thank you. $U2$: You’re welcome. $U3$: Ross!	Positive	Monica	Neutral
Rachel	$U1$: Yeah, it’s a real shame you can’t make it to that one-woman show tonight. $U2$: Oh, I’d love to, but I gotta get up so early the next day and so, you know me, work comes first $U3$: Oh, yeah, yeah, yeah...	Neutral	Ross	Neutral
Ross	$U1$: No you do know what, you’re not gonna suck me into this. $U2$: Oh sure I am, because you always have to be right. $U3$: I do not always have to be okay, okay.	Negative	Rachel	Neutral

TABLE 7.25: Emotional responses of speakers when replacing their personalities with other speakers. GT stands for Ground Truth, MT stands for Multi-Task, and Replace Personality represents the speaker whose personality is used to replace the original speaker’s personality. The red color indicates the wrong prediction and the green indicates the correct prediction. *U1*, *U2*, and *U3* represent Utterances 1, Utterances 2, and Utterances 3 respectively. The predicted emotional responses are in 7 categories.

Speaker	Dialogue	GT	Replace Personality	MT
Chandler	<i>U1</i> : Do I ever. <i>U2</i> : Chris says they’re closing down the bar <i>U3</i> : No way!	Surprise	Joey	Neutral
Joey	<i>U1</i> : You liked it? <i>U2</i> : You really liked it ? <i>U3</i> : Oh-oh-oh	Surprise	Chandler	Joy
Monica	<i>U1</i> : Hey, did you pick a roommate? <i>U2</i> : You betcha! <i>U3</i> : Is it the Italian guy?	Neutral	Phoebe	Neutral
Phoebe	<i>U1</i> : Okay. Does it have to do with Ross and Rachel? <i>U2</i> : No <i>U3</i> : Does it have to do with Joey?	Neutral	Monica	Fear
Rachel	<i>U1</i> : But I told you, I didn’t have the time! <i>U2</i> : Yeah, well you never have the time. I mean, I didn’t feel like I even have a girlfriend anymore, Rachel. <i>U3</i> : Wh, Ross what do you want from me?	Anger	Ross	Fear
Ross	<i>U1</i> : I loved this place! To tell you the truth, I wish I didn’t have to move. <i>U2</i> : Uhh, are you saying that you’re not entirely happy about this? <i>U3</i> : Well, I mean if uh, if Emily gave me a choice	Sadness	Rachel	Fear

7.6 Discussion

Automatically predicting emotional responses in the conversation has been a great interest in Human-Computer Interaction. One of the exciting research areas is the development of an open-domain dialogue system. The common approach to this research is predicting emotional responses without considering the subjectivity of the speakers. Since emotions are highly subjective, a single utterance may generate different emotional responses in different speakers.

To get the subjective emotional response of each individual speaker, our study presents an approach that models each emotional response separately along with the speaker's personality information. In particular, two different types of approaches have been proposed, Single-Task (ST) and Multi-Task (MT). In Single-Task (ST), every time a separate model is trained for each speaker to get the subjective emotional responses for the next utterance. On the other hand, in Multi-Task (MT), a single model is trained for multiple speakers to get subjective emotional responses with respect to each speaker. The MT approach has multiple outputs in predicting emotional response; one for every single speaker. The Multi-Task (MT) learning was tested with 3-categories and 7-categories emotion classification. The results show that the Multi-Task (MT) approach helps in predicting subjective emotional responses.

In conclusion, the Multi-Task (MT) approach performed better in order to generalize the subjective emotional response of each speaker individually compared to Single-Task (ST) and the state-of-the-art. The reason behind this is that Multi-Task (MT) learning helps the network to learn a better representation of the data. As a result, the network has more generalization capabilities for each individual speaker. In this study, only the textual modality is used, whereas, in a daily conversation, other non-verbal cues such as speech signals, body poses, and facial expressions also play an important part in generating subjective emotional responses. Dialogues with non-verbal cues change the whole context of the conversation and, as a result, produce different subjective emotional responses.

Chapter 7: Subjective Emotions using Personality Traits

An interesting line of future work would be a multi-modal approach to predicting subjective emotional responses.

Unpublished manuscript

- The paper is submitted to the 11th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2023), and it is currently under review.

Chapter 8

Automatic Recognition of Personality Traits

*“All personality traits have their
good side and their bad side.”*

Susan Cain

This chapter presents the research work on the automatic recognition of personality traits using speech data. The study investigates the performance of pre-trained weights for underlying audio classification tasks as compared to hand-crafted features. Later, this chapter explains how to convert a non-interpretable CNN model into an interpretable CNN model for big-five personality traits. The proposed interpretable model discovered the distinct frequency patterns for each personality trait. We believe interpretability in affective machine learning systems helps to understand affective subjectivity in more detail.

In recent years, personality computing [239] has become a very active research subject that focuses on computational techniques related to human personality. It mainly addresses three fundamental problems: automatic personality trait recognition, perception, and synthesis [240]. The first one aims at correctly identifying or predicting the actual (self-assessed) personality traits of human beings. This allows the construction of an apparent personality of an unacquainted individual. Automatic personality trait perception concentrates on analyzing the different subjective factors such as cultural, social, contextual, gender, and appearance that affect the personality perception of a given individual. Automatic personality trait synthesis tries to generate artificial personalities through artificial agents and robots. This research work focuses on the first problem of personality computing, i.e. automatic personality trait recognition. The reason behind this is that personality traits provide a promising additional source of information for personalization, which goes beyond context- and device-specific behavior and preferences.

Automatic personality trait recognition from social media content has recently attracted much attention in the fields of artificial intelligence and computer vision, e.g. [240–242]. In 2016, the well-known European Conference on Computer Vision (ECCV) released a challenge named ChaLearn Looking At People (LAP) [243]. The main objective of the challenge was to develop a machine-learning architecture that quantitatively evaluates the recognition of the Big Five personality

traits of speakers. For ChaLearn 2016 challenge, the First Impression dataset [243] was released, which is still the largest public database for apparent personality trait estimation. Our study focuses on the First Impression dataset.

Most previous works focus on personality trait modeling and prediction from different cues, both behavioral and verbal. Junior et al. [241], Mehta et al. [242], and Zhao et al. [240] presented very detailed surveys on recently developed techniques that used single or multi-modalities for automatic recognition of personality traits. We observe that works approaching the problem from a multi-modal perspective were the ones showing the best results [244] on the First Impression dataset. Most of these multi-modal approaches usually combine image and audio, extracting visual features with Convolutional Neural Networks (CNN), and using a late fusion with hand-crafted audio features.

In contrast, we only chose audio modality for the recognition of personality traits. The reason behind this is that in psychology, it is widely accepted that speech conveys a great deal of information about the speaker’s personality traits [245–247]. The objective of our work is to explore audio modality in more depth, and the possibility of using end-to-end methods for the audio modality. The use of hand-crafted features for audio processing is usually motivated by the lack of a large corpus of data for training end-to-end systems. However, recent audio databases such as AudioSet [140] provide enough data to train a Deep Learning model. This work takes advantage of the pre-trained model on AudioSet to fine-tune the personality trait recognition task on the First Impressions dataset. We found that the pre-trained weights are a good initialization point for the intended task, improving the results obtained by hand-crafted features.

8.1 Related Work

8.1.1 Automatic Personality Recognition

In most cases, Automatic Personality Recognition (APR) approaches aim at inferring emotional and social phenomena from machine-detectable behavioral evidence such as facial expression, speech, and text, which are used to predict personality traits. Mohammadi et al. [248] used the Praat tool [249] to extract the audio features pitch, energy, and voiced segments as well as statistical features (maximum, minimum, mean, relative entropy) from speech clips. These features were used with the combination of Logistic Regression and Support Vector Machine to predict the personality traits. The speech clips were taken from SSPNet Speaker Personality Corpus [248]. Bietal et al. [250, 251] used speaking activity, prosodic features, looking activity, verbal content, and facial expression in their work for personality prediction.

At European Conference on Computer Vision (ECCV) 2016, a challenge named ChaLearn Looking At People (LAP) [243] was introduced. The main objective of the challenge was to develop a machine-learning architecture that takes Human-Centered YouTube videos and quantitatively evaluates the recognition of the Big Five personality traits of speakers. In this challenge, Subramaniam et al. [252] proposed a two-stream model: one stream for visual and another stream for audio representation. The audio stream uses 68 dimension handcrafted features vector for audio processing. These features are extracted by splitting the raw signal into short-term windows (frames) and computing a number of features for each frame. Then the mean and standard deviation of each feature sequence are computed. These features concatenated with the visual features are taken as input to the multi-modal neural network for personality prediction.

Similarly, Zhang et al. [244] proposed the Deep Bimodal Regressor (DBR) framework, which uses CNN for visual representation and linear regression for the audio stream. For audio processing, the Mel-Frequency Cepstral Coefficients (MFCCs) features are first extracted from raw wave data of each video and then

the Mel Filter bank energies are computed. These energies are the input of the linear regressor. Finally, the visual modality and audio modality are lately fused by averaging the scores of the visual and the audio modules. DBR framework ranked first place in the ECCV ChaLearn LAP challenge 2016.

8.1.2 Audio Classification

In the past decades, most of the works on automatic audio classification are based on hand-crafted features. Some of these features are time domain features, and others are frequency domain features. Some examples are Mel Frequency Cepstral Coefficients (MFCC) [253], Linear Prediction Cepstral Coefficient (LPCC) [254], and Bark Frequency Cepstral Coefficient (BFCC) [254]. Recently, Convolution Neural Networks (CNN) have shown encouraging effectiveness and attracted more attention to audio processing [138].

Panagiotis Tzirakis et al. [255] trained an end-to-end convolutional neural network for extracting audio features from raw data on an emotion recognition task. These features are considered as input for a 2-layer LSTM (Long Short-Term Memory) for contextual information extraction. The proposed model produces state-of-the-art results for the RECOLA [75] dataset of the AVEC 2016 research challenge on emotion recognition. Satt et al. [256] remove non-speech signal such as crowd noise/music from 2D-Mel-Spectrogram and feeds this into the CNNs for emotion recognition.

Samarth Tripathi et al. [257] use a multi-modal deep learning approach for emotion recognition using the IEMOCAP dataset [80] with different modalities: text, speech, and image. The concatenation of all modalities outperforms the state-of-art results. To our best knowledge, most CNNs have been used for emotion recognition, but CNN based on audio cues have not been considered for the personality recognition task. Note that CNN methods usually require large amounts of labeled data to be trained. For tasks without enough available data

for training, a common approach is to use a similar dataset for pre-training the system and then fine-tuning it for the target task.

In the previous context, AudioSet [140] can potentially be an interesting resource to use. The upper ontology of the dataset covers a wide range of everyday sounds, including the domain ontology of human sounds, which consists of the human voice (speech, shouting, screaming), respiratory sounds (breathing, coughing, sneezing), heart sounds/heartbeat (heart murmur), among others. As the First Impressions dataset comprises clips extracted from different YouTube high-definition (HD) videos of people facing and speaking in English to a camera. People in videos show different gender, ages, nationalities, and ethnicity.

8.2 End-to-End Approach for Personality Trait Recognition from Audio

8.2.1 Audio Pre-Processing

An audio signal is the electronic representation of a sound wave like speech, music, or any other type of sound. An audio signal may be represented in a digital or analog format. We deal only with the analog representation of an audio signal. A spectrogram is a general way to represent an analog representation of an audio signal. The spectrogram is a graph with two geometric dimensions: the horizontal axis represents the time and the vertical axis represents the frequency. A third dimension indicating the amplitude of a particular frequency at a particular time is represented by the intensity or color of each point in the graph. We use two variations of the spectrogram.

Clip-Spectrogram (CS)- To generate the clip-spectrogram we proceed as follows. First, we extract an audio sequence from the video. Each audio sequence is divided into clips of 960 milliseconds (ms). After getting clips, the next step is to extract spectral information of each clip i.e. how much energy the windowed signal

contains at different frequency bands. The Short-Time Fourier Transform (STFT) is used for extracting spectral information with $25ms$ window length and 10 ms window shift. STFT transformed the 960 ms clip into 64 Mel-spaced frequencies, and the magnitude of each bin is log-transformed. This gives the $2D$ log-Mel-spectrogram patch of $96*64$ bins for the input of our model.

Summary-Spectrogram (SS)- The window length ($25ms$) that we used to get the spectral information of each $960ms$ frame holds only a small fraction of audio information which may not be sufficient for the personality prediction task. To consider the whole audio information of a video we concatenate the clip-based $2D$ log-Mel spectrograms along with the temporal domain. This way, we obtain a $1248*64$ dimensional spectrogram. Then, an average pool operation is performed to reduce the size of the spectrogram. We take an average of 60 ms frame across all 64 Mel-spaced frequency bins, obtaining, as a result, a $208*64$ spectrogram, which we refer to as Summary-Spectrogram.

8.2.2 Architectures and Implementation Details

To get the personality traits prediction from CS and SS, we decided to fine-tune the VGG model [138] on the First Impression dataset. The reason behind this is that the VGG model has got state-of-the-art results on many audio event classifications. The VGG model has multiple convolution layers for feature extraction. The first layer has 64 filters, the 2nd layer has 128 filters, and the 3rd and 4th layers are in a pair of two convolution layers having 256 and 512 filters respectively. Max-pooled operation is applied to reduce the dimensionality of the convolved filters. After that, two fully-connected layers having 4096 neurons are applied for feature mapping. Lastly, the output layer consists of 5 neurons, one for each personality trait. The architecture is shown in Fig. 8.1. In the training phase, the model is fine-tuned on First Impression Dataset using the pre-trained model for audio event classification in AudioSet [140] with a learning rate of 10^{-6} and Adam optimizer for optimization of the model.

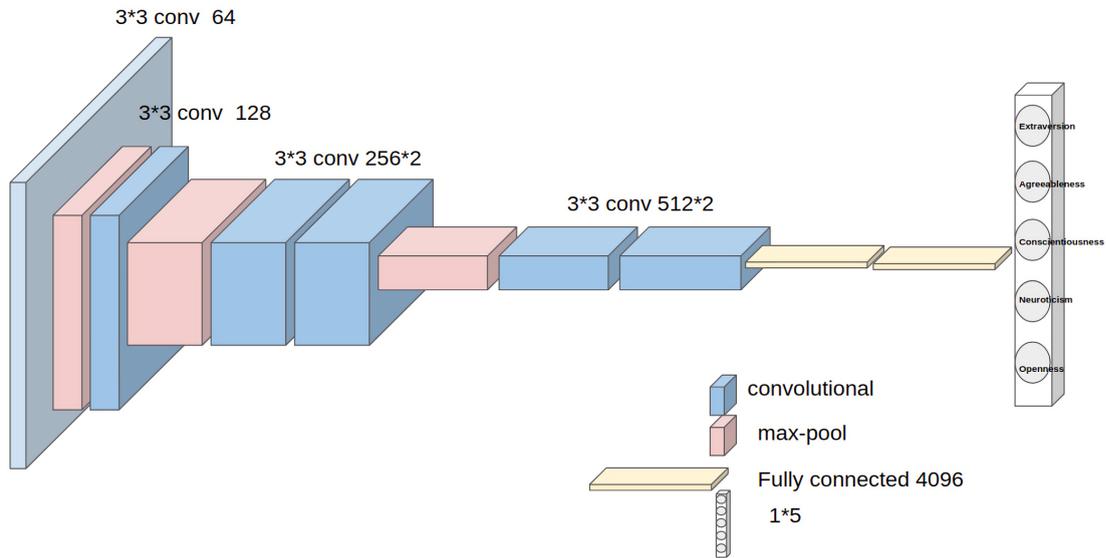


FIGURE 8.1: Audio-based convolutional neural network for big five personality prediction using Clip-Spectrogram (CS) and Summary-Spectrogram (SS).

8.3 Experiments

8.3.1 First-Impression Dataset

The First Impressions dataset [243] consists of 10,000 video clips extracted from more than 3,000 different YouTube high-definition (HD) videos. These videos are human-centered in which a person is facing and speaking English to a camera (see Fig. 8.2). People in the videos belong to different gender, ages, nationalities, and ethnicity. Each video clip was annotated by AMT workers. Each worker was shown two videos and asked to answer which of the two subjects presented individual traits more strongly (see Fig. 8.3). The pairwise data is converted in [243, 258] to continuous values using [259]. This method individually converts the ordinal ratings of each dimension into continuous values (such as the level of “Extraversion”) by fitting a Bradley-Terry-Luce (BTL) [243] model with maximum likelihood, which is further scaled to be in the range of [0, 1]. This way, each video sample in the dataset will have a continuous value associated with each trait dimension, which can be used by any supervised learning method, in a classification or regression task.

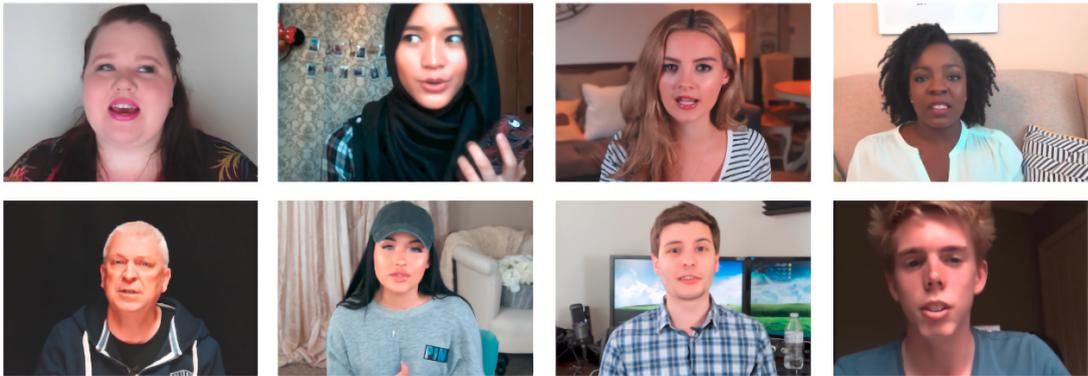


FIGURE 8.2: Few examples of Human-Centered videos in which people are talking to the camera about any predefined topic [260].

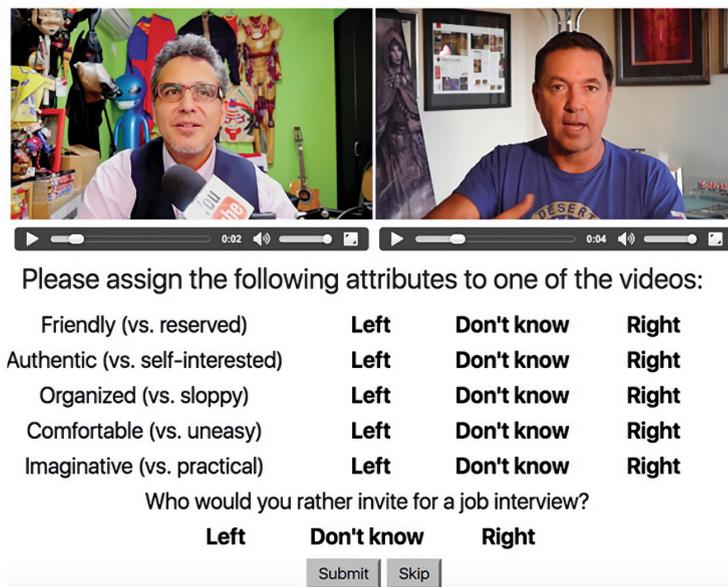


FIGURE 8.3: Snapshots of the interface for labeling videos [243]. The “big five” traits are characterized by adjectives: Extroversion = Friendly (versus Reserved); Agreeableness = Authentic (versus Self-interested); Conscientiousness = Organized (versus Sloppy); (non-)Neuroticism = Comfortable (versus Uneasy); Openness = Imaginative (versus Practical).

8.3.2 Evaluation metric

The output of the models consists of a set of 5 continuous prediction values in the range $[0, 1]$ and the performance is evaluated by computing the mean accuracy over all five traits and videos. Accuracy for each trait is defined in Eq. 8.1.

$$A = 1 - \frac{1}{N_t} \sum_{i=1}^{N_t} |t_i - p_i| / \sum_{i=1}^{N_t} |t_i - \bar{t}| \quad (8.1)$$

where p_i are the predicted scores, t_i are the ground truth scores, with the sum running over the N_t test videos, and \bar{t} is the average ground truth score over all videos.

8.3.3 Results

This section presents the performances of the models that were trained using Clip-Spectrogram (CS) and Summary-Spectrogram (SS). The results are presented in Table 8.1. The model that was trained using Clip-Spectrogram (CS) has got the highest performance as compared to Summary-Spectrogram (SS) and NJU-LUMDA [244]. The results show that the weights that are trained on a large-scale audio dataset help in learning for other audio classification tasks and perform better as compared to the hand-crafted features.

Furthermore, the results also show that the low-frequency patterns are also very important along with high-frequency patterns presented in the spectrogram for mapping personality traits. This is why the model which was trained using Clip-Spectrogram (CS) has achieved the highest performance against Summary-Spectrogram (SS).

TABLE 8.1: Accuracies Evaluation Results E refers to Extraversion, A to Agreeableness, C to Conscientiousness, N to Neuroticism, and O to Openness.

Model Name	Mean Accuracy	E	A	C	N	O
NJU-LUMDA [244]	0.8900	0.890	0.892	0.886	0.896	0.888
Ours-SS	0.8944	0.897	0.895	0.894	0.896	0.890
Ours-CS	0.9009	0.900	0.908	0.899	0.901	0.905

8.4 Interpretability for Personality Trait Recognition from Audio

Deep learning models [261] have achieved remarkable performance in a variety of tasks, from visual recognition, natural language processing, and reinforcement learning to recommendation systems, where deep models have produced results comparable to and in some cases superior to human experts. Due to their nature of overparameterization (involving more than millions of parameters and stacked with more than hundreds of layers), it is often difficult to understand the prediction results of deep models [262]. Explaining their behaviors remains challenging because of their hierarchical non-linearity in a black-box fashion. The lack of interpretability raises a severe issue about the trust of deep models in high-stakes prediction applications, such as autonomous driving, healthcare, criminal justice, and financial services [263].

Interpretability is the extraction of relevant knowledge from a machine learning model concerning relationships either contained in data or learned by the model [264]. For example, to predict the value of a house, the machine learning model would learn patterns from past house sales. The higher the interpretability of a machine learning system, the easier it is for humans to comprehend why certain decisions or predictions have been made.

Our main goal to introduce interpretability is to understand what parts of the audio are the most informative for the model to recognize the personality traits. For this goal, we used a technique called Class Activation Map (CAM) [265]. CAM was originally introduced for explaining the outputs of image classification CNNs. More concretely, given an image classification task, CAM is a method for visualizing the regions of the image that are highly informative for the classifier. Specifically, given a particular category of the classification task, CAM generates a heatmap indicating the discriminative image regions used by CNN to identify that specific category.

Previously, Ventura et al. [266] proposed an interpretable Convolutional Neural Network (CNN) model for automatically inferring the personality traits of people talking to a camera. This work only considers the visual modality in order to interpret the results of CNN. The experiments were done using the FirstImpression dataset [243]. The authors used the CAM [265] technique to introduce interpretability in CNN. Later, they used face detection and Action Unit (AUs) recognition systems to show the parts of human faces that the model used to predict different personality traits of people.

Notice that CAM can be easily applied to a CNN that takes spectrograms as input, since the spectrograms themselves can be represented as images. Thus, the application of CAM, in that case, would allow locating the regions that have discriminative frequencies and their amplitude with respect to time. That region represents the frequencies with different ranges across time. For frequency analysis, we average all the corresponding frequencies with respect to the time and get the 1-dimensional representation of frequencies. CAM has also been adapted for the case that the CNN takes the raw audio data as input. As it can be seen in Fig. 8.4, the output CAM is a 1D signal that represents how discriminative each component of the input signal is. In this case, since the input is the raw signal in the frequency domain, CAM is applied to find the most discriminative frequency components for the personality trait regression problem.

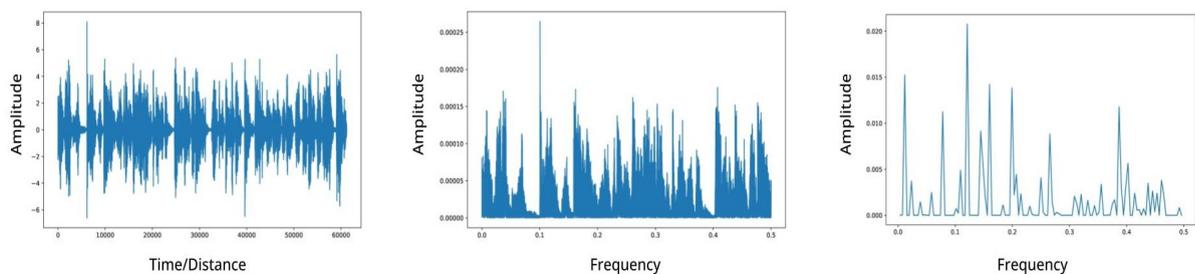


FIGURE 8.4: Left: Representation of the audio signal in the Time domain. Middle: Representation of the audio signal in the Frequency domain. Right: CAM generated discriminative frequency patterns for a particular personality trait in an audio signal.

8.4.1 Interpretable CNN for Clip-Spectrogram (Our-ICS) and Summary-Spectrogram (Our-ISS)

We use fine-tuned VGG network [138] with some additional modifications for interpretability: we remove all the fully-connected layers between the last convolutional layer and the output layer. Then we add a Global Average pooling (GAP) layer before the output. This layer performs the average operation on the convolutional features of the Conv-4 layer and uses those features for a fully-connected layer that produces the prediction results. Fig. 8.5 represents the interpretable CNN architecture. The same architecture is used for both: CS and SS. We followed the same training procedure to fine-tune our interpretable models for clip-spectrogram (CS) and summary-spectrogram (SS) as we adopted to fine-tune our non-interpretable models.

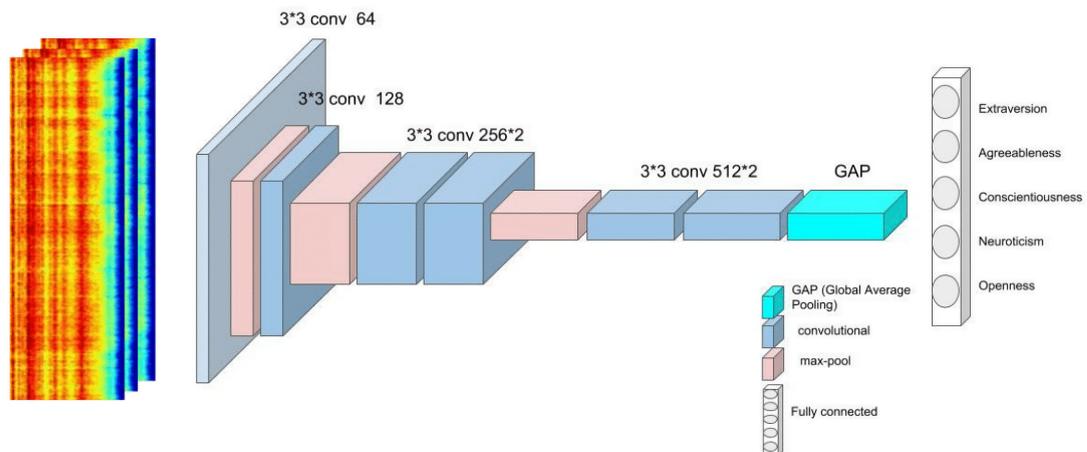


FIGURE 8.5: Interpretable CNN: GAP (Global Average Pooling) is introduced in the network for interpretability

8.4.1.1 Results

Interpretable Clip-Spectrogram (Ours-ICS): This model corresponds to the architecture displayed in Fig. 8.5, which includes the CAM module for interpretability. To observe the discriminative regions of the spectrograms, we have generated the CAMs for 20 clips that have the highest prediction values

with respect to each personality trait. We have not found any clear evidence of a pattern for the discriminative regions on the spectrogram with respect to each big five personality trait. Since there is no clear evidence, we only take one trait for visualization, i.e. extraversion, which is displayed in Fig. 8.6. To extend this analysis, we decided to aggregate the CAM values without any binarization along the time domain of the spectrogram to see if there is any pattern in the frequency domain. As it can be seen in Fig. 8.7, we found these time-aggregated CAMs have a very similar pattern for all 20 clips with the highest predicted value for the extraversion personality trait. These frequency pattern has been also found for the other personality traits. However, we did not find any discriminative pattern related to each specific personality trait, which indicates that the model decision about all personality traits is based on the same features. Regarding the accuracy achieved by this model (0.8990, see Table 8.2), we would like to highlight that although the accuracy is slightly worse than the non-interpretable model (0.9009, see Table 8.1), this model is considerably lighter (it does not include any fully connected layer) besides the fact of being interpretable with the CAM technique. Furthermore, it outperforms the NJU-LAMDA baseline (0.8900).

TABLE 8.2: Accuracies Evaluation Results E refers to Extraversion, A to Agreeableness, C to Conscientiousness, N to Neuroticism, and O to Openness.

Model Name	Model Type	Model Accuracy	E	A	C	N	O
NJU-LUMDA [244]	Non-Interpretable	0.8900	0.890	0.892	0.886	0.896	0.888
Ours-ICS	Interpretable	0.8990	0.897	0.906	0.891	0.897	0.902
Ours-ISS	Interpretable	0.8963	0.892	0.896	0.894	0.898	0.893
Ours-IRA	Interpretable	0.8946	0.890	0.892	0.894	0.896	0.890

Interpretable Summary-Spectrogram (Our-ISS): This model also corresponds to the architecture displayed in Fig. 8.5, which includes the CAM module for interpretability but uses the Summary-Spectrograms instead of the Clip Spectrograms. Regarding the visualization of the discriminative regions using CAM for the 20 videos with the highest predictive value for each personality trait (see Fig. 8.8), we have also found that there is not a clear pattern in the spectrograms. As done with the clip spectrogram model, a frequency-based analysis

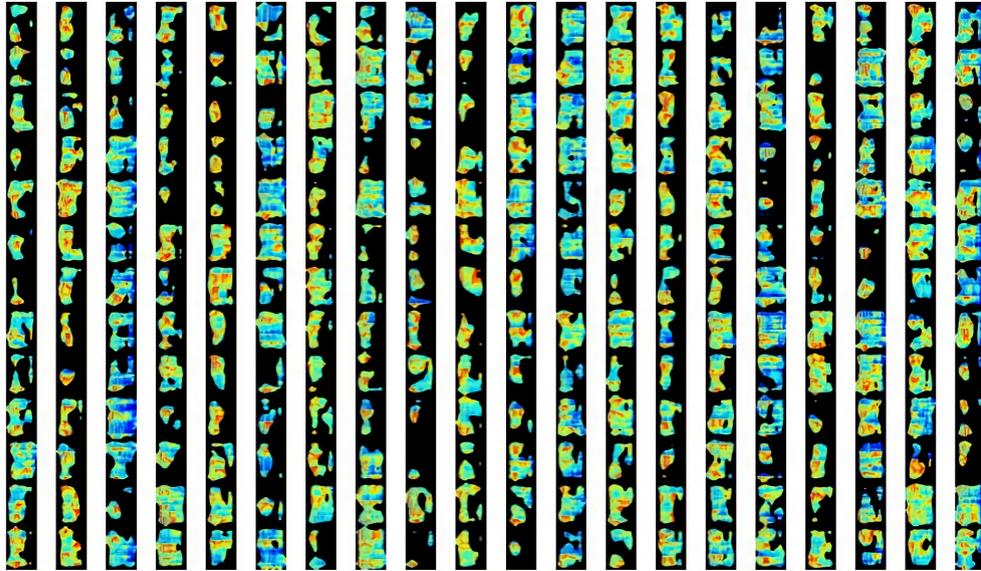


FIGURE 8.6: Class activation maps of 20 highest predictions of extraversion trait (from left to right: Each Column represents one video (13 clips))

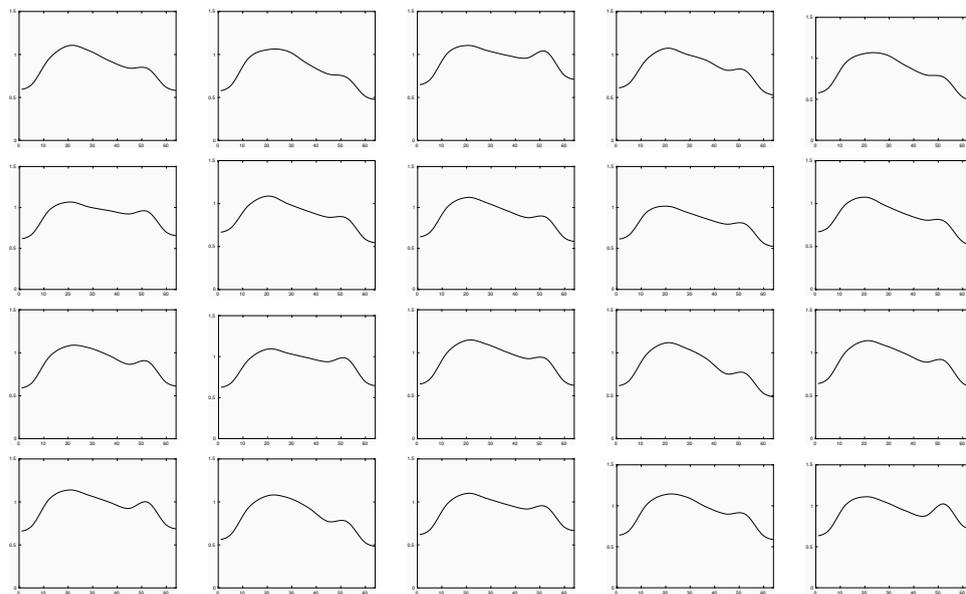


FIGURE 8.7: Class activation maps of 20 highest prediction of Extraversion trait (from left to right: Each CAM represents the frequency component of a clip)

has also been performed (see Fig. 8.9). As a result, we have also found a frequency pattern, which is very similar to the frequency pattern found for the clip spectrogram model. However, as it can be seen in Table 8.2, the Summary-Spectrogram model achieves an accuracy of (0.8963) which is lower than the accuracy achieved by the Clip Spectrogram model (0.8990) but higher than the NJU-LAMDA baseline (0.8900).

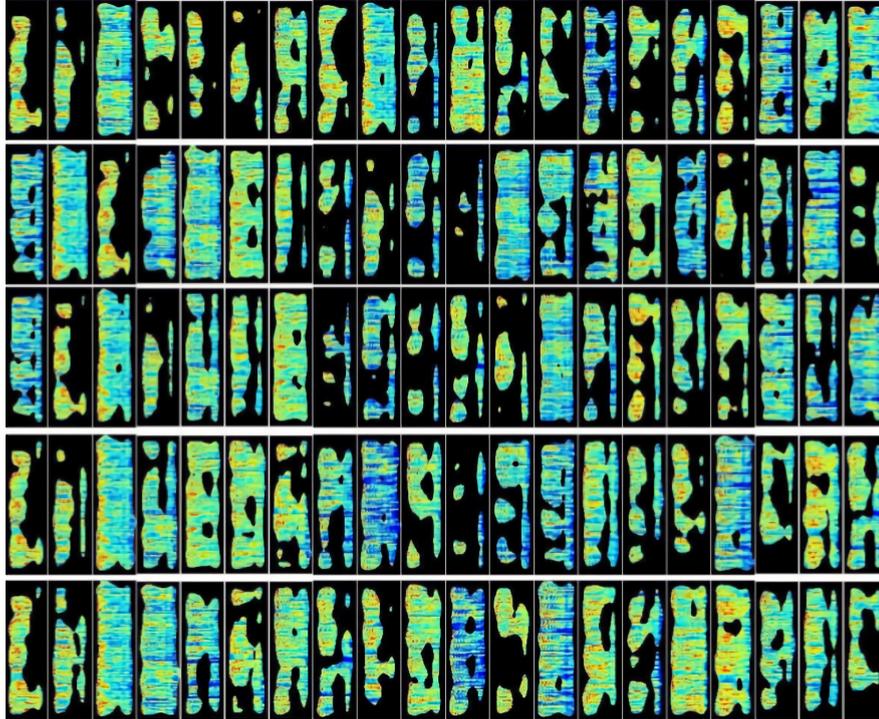


FIGURE 8.8: Class activation maps of 20 highest predictions videos of all traits. From top to bottom: Row 1: Extraversion, Row 2 : Agreeableness, Row 3 : Conscientiousness, Row 4: Neuroticism, Row 5 : Openness

8.4.2 Interpretable CNN for Raw Audio Data (Our-IRA)

Besides using the spectrogram, we have also considered feeding the raw audio data directly as input to the neural network. We convert the original signal from the time domain into the frequency domain by (FFT) [267]. FFT decomposes a sequence of values into a component of different frequencies and we use these frequencies as input to our network. The reason behind this transformation is to find a pattern of discriminative regions in the frequency domain, which may be repeated for different sequences for the same personality trait. The architecture is

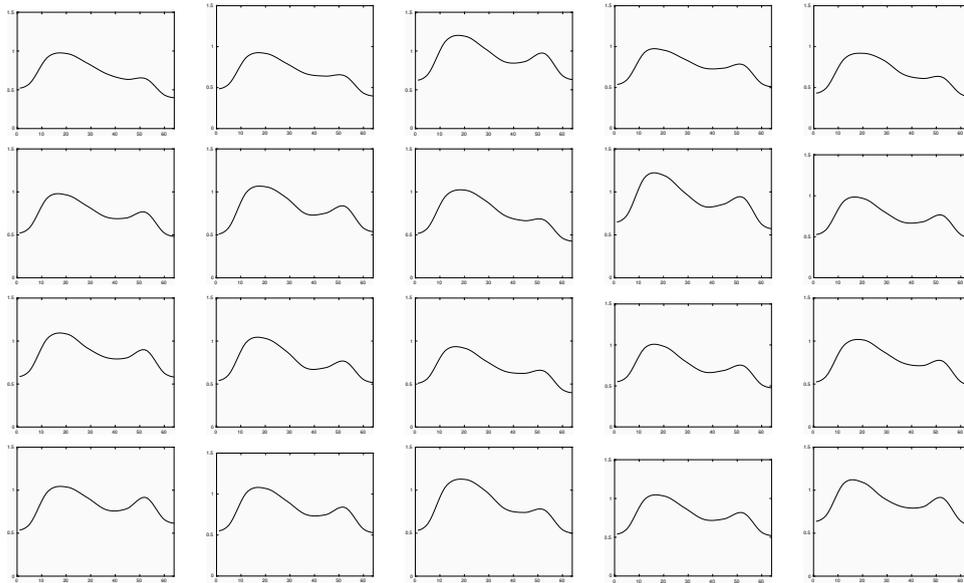


FIGURE 8.9: Class activation maps of 20 highest predictions of extraversion trait (from left to right: Each CAM represents the frequency component of a whole video)

displayed in Figure 8.10, which takes as input the raw audio signal in the frequency domain and can be analyzed by applying the CAM technique as illustrated in Fig. 8.4.

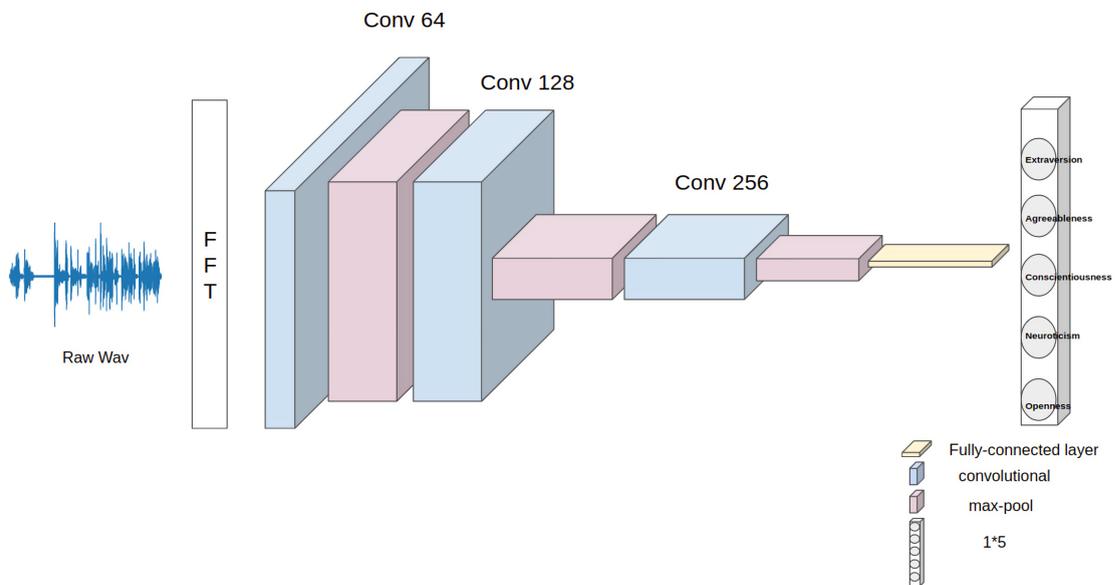


FIGURE 8.10: The proposed convolutional neural network for personality prediction when using raw audio data. An FFT module is applied to the raw signal before feeding it to a CNN for prediction. No fully connected layers are used in order to design an interpretable model.

8.4.2.1 Results

For interpretability, we have averaged the CAMs obtained for the 20 videos with the highest predicted values for each personality for visualization in Fig. 8.11. We observe that CAMs for all personality traits are not identical, which means that the prediction made by the model is based on different frequency components depending on the personality trait to be predicted. Regarding the accuracy, as it can be seen in Table 8.2, this model works slightly worse than the Interpretable Clip-Spectrogram model (0.8946 vs 0.8990) and Summary-Spectrogram (0.8946 vs 0.8963), but this model is even lighter in terms of the number of parameters and still outperforms the NJU-LAMDA baseline (0.8900), which is based on hand-crafted audio features.

8.5 Discussion

In this research work, we proposed an audio feature extraction scheme for big five personality prediction. We proposed a scheme that takes advantage of the learned features from a large-scale audio dataset and fine-tuned it for personality traits recognition in speech data. The proposed scheme outperforms the state-of-the-art results and shows that the learned features perform better than the hand-crafted features. Furthermore, we proposed a fully convolutional architecture that allows finding the discriminative regions from different audio input representations (spectrogram and frequency) by using the CAM technique, which is usually applied to images. Thanks to the use of this technique, we proposed a model which is better than state-of-the-art techniques based on hand-crafted audio features both in terms of accuracy and interpretability. The accuracies achieved by interpretable models are very close to the non-interpretable models. Furthermore, the significant advantage over non-interpretable models is that the interpretable models are lighter in the number of parameters and take less time to converge.

The results show that when CAM was applied to the spectrogram the model could not learn consistent discriminative regions for each personality trait in all

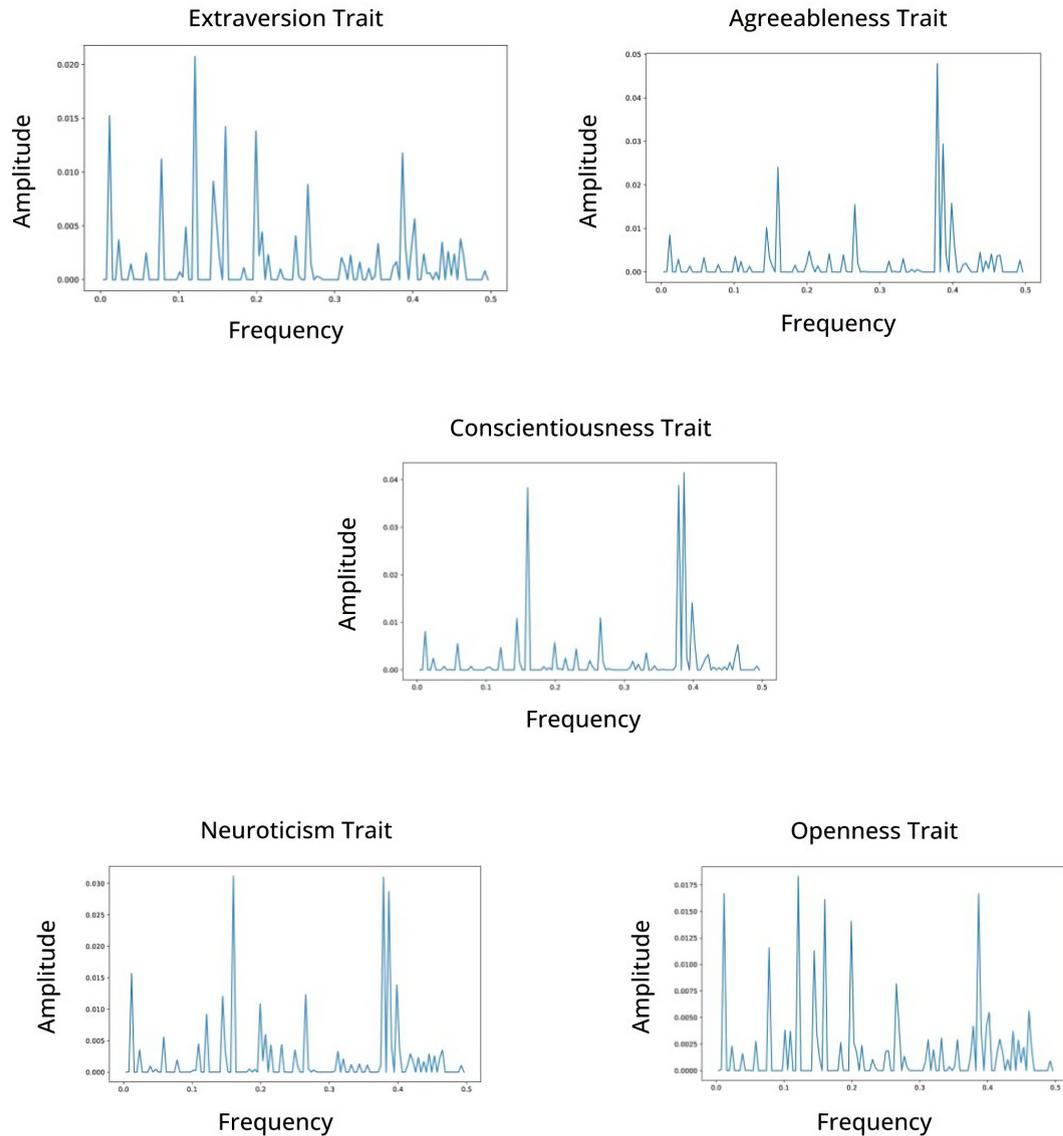


FIGURE 8.11: Frequency-based class activation maps for top 20 predictions of each personality traits. Each frequency pattern is the average of the 20 Frequency-based class activation maps of that personality trait.

speakers. On the other hand, the CAM performed well on the frequency dimension and found consistent discriminative regions for each personality trait in all speakers. The videos in the FirstImpression dataset are human-centered videos in which people are talking to the camera (a non-responded machine). To find subjective discriminative patterns with respect to personality traits in a speech signal, the community needs datasets that are large enough to capture subjective emotional responses based on the personality traits of a speaker.

We want to notice that this work was done in 2018 and published in 2019. In the past recent years, much work has been proposed that adopted a multi-modalities approach including audio for personality traits recognition using the FirstImpression dataset. Furthermore, with the recent gaining on fairness in machine learning, the FirstImpression dataset was also examined by the researchers [260] and found that the labels are biased due to human perception. This biased behavior is related to the annotator's subjectivity in Part I of this thesis. However, Part I focuses on subjectivity in the context of annotations on affect instead of the subjectivity of annotations on personality. We think that studying subjectivity in the context of personality trait recognition would be an interesting direction for future research.

Publication

- Hassan Hayat, Carles Ventura, Agata Lapedriza, ‘On the use of Interpretable CNN for Personality Trait Recognition from Audio’, **Frontiers in Artificial Intelligence and Applications**, 2019. DOI: 10.3233/FAIA190116
 - Source code available at: [GitHub Repository Link](#)

Conclusion

Important challenges in Affective Computing are to develop systems that are able to generalize and, at the same time, can understand the subjective emotional expressions and responses of each individual. The research work presented in this thesis was carried out toward the development of systems that are aware of emotional subjectivity. This research work addressed emotional subjectivity from two different dimensions:

Subjectivity in Affect Labeling- In supervised machine learning, most of the affective datasets only release hard labels associated with the data samples. Hard labels are aggregated labels that come from multiple soft labels. Generally, soft labels represent the annotator's individual emotional perception. Each annotator has some agreement and disagreement with other annotators. In the process of getting hard labels, developed functions only considered the agreement part and discard the disagreement which actually represents the subjective emotional perception of each annotator. The research community is divided into two considerations with the annotator's disagreement: subjectivity as noise and subjectivity as information. This research work considered annotators' disagreement as information in modeling subjective emotional perception. More specifically, we developed a multi-modal Multi-Task (MT) learning technique that learns the annotator's agreement and disagreement simultaneously. The proposed machine learning model has multiple outputs. Each output comes from a separate block that is only designated to learn a specific annotator's emotional perception. These outputs are trained using soft labels with respect to every single annotator. Along these soft labels, there is an additional output that is trained on the hard labels,

Conclusion

i.e. aggregated labels. The purpose of this output is to learn also the aggregated emotional perception. All these output blocks are connected to a shared backbone that basically learns the shared emotional patterns between each individual and the aggregated annotator.

The multi-modal Multi-Task (MT) approach was tested on three different datasets named COGNIMUSE, IEMOCAP, and SemEval.2007, ranging from binary to multi-class emotion classification. The results show that multi-modal Multi-Task (MT) approach has surpassed the state-of-the-art approaches. Besides these significant improvements shown by MT, the approach also has a limitation, i.e. the proposed architecture is limited to model the number of annotators. In the experiments, the MT was tested with a maximum of 8 annotators, including the aggregated annotator, since we only found very few datasets that released the annotator-level annotations, which is the key to modeling emotional subjectivity in supervised machine learning. The results of this research work encourage the affective computing community to release the annotator-level annotations, along with aggregated annotations, when publicly releasing any affective datasets.

Subjective emotional responses in dialogue systems- Automatically generating emotional responses is a challenging area in dialogue systems. Dialogue systems consider dialogue as a conversation between the user and a system. The aim of this research work is to automatically generate subjective emotional responses in dialogue systems with respect to each speaker. Researchers in psychology established a strong correlation between personality traits and subjective emotional responses of a speaker. Based on these findings, we developed an approach that modeled the personality traits of a speaker with the emotional variations presented in the preceding utterances to generate subjective emotional responses of a speaker.

For this study, we used PELD (Personality EmotionLines Dataset). To our best knowledge, this is the only available dataset in which the utterances of the dialogues are annotated with the speaker id, textual representation, emotion categories, sentiment categories, and the personality traits of a single speaker. To

Conclusion

model subjective emotional responses for each speaker, the training data should consist of multiple datasets. Every single dataset has samples that belong to only a particular speaker. In the PELD dataset, the first utterance of each dialogue is annotated with the emotion category, which acts as an initial emotion of the conversation and with the personality traits of the speaker. This personality information is used to categorize all the dataset dialogues into speaker-specific dialogues to generate multiple datasets for training a Multi-Task (MT) model. The big five personality information and the initial emotions are transformed into VAD (Valence, Arousal, and Dominance) vectors before feeding the utterances into the Multi-Task (MT) model.

In the Multi-Task (MT) model training, preceding utterances of a single speaker are first transformed into semantically meaningful vectors. The Bert-Base language model is used to obtain these vectors. Then these contextual embeddings are encoded into the contextual VAD vectors. These contextual VAD vectors represent the emotional variation in the preceding utterances. The preceding emotional variations are influenced by the personality traits of the speaker, i.e. the learned personality VAD vector. The resultant vector represents the emotional variations caused by the personality traits of the speaker. Later, the sum of the personality-based emotion variations and the contextual VAD vector is used to generate the subjective emotional responses of the speaker. In a single epoch, each speaker’s training sample is used to train a Multi-Task (MT) model.

The proposed Personality based Multi-Task (MT) learning was tested with two emotion category representations: 7-class emotion categories and 3-class emotion categories. The results show that the Multi-Task (MT) approach has surpassed the state-of-the-art results for both emotion category representation set. To our best knowledge, this is the first work that separately modeled the emotional response of every single speaker using the PELD dataset. The lack of a publicly available dialogue dataset, in which the utterances are annotated with emotion classes, speaker id, and the personality traits of the speaker create a barrier to applying Personality based Multi-Task (MT) learning on other datasets.

Conclusion

It is also a serious concern in enhancing the progress toward subjective aware dialogue systems.

Overall, the presented research work addressed the problem of emotional subjectivity in affective computing systems. To attain subjective aware affective models, a Multi-Task (MT) learning approach is proposed, i.e. separate modeling of each emotional perception in the data. The proposed Multi-Task (MT) learning was tested in developing subjective aware affect models from two different perspectives (*i*) subjectivity in affect labeling and (*ii*) subjective emotional responses of speakers in dialogues. The results show that the separate modeling of each emotional perception in the data converges the machine learning system better and has more generalization capabilities compared to the combined modeling of each emotional perception in the data.

Lastly, this thesis presents the research work on interpretable models for personality trait recognition. The research work explored a Convolutional Neural Network (CNN) based architecture that learns the audio cues to predict the Big Five personality traits score of a speaker. The model takes advantage of a pre-trained model on a large database for audio event recognition (AudioSet) and has been finetuned on the First Impression Dataset to obtain an audio representation for personality trait recognition. Then we interpret the model and generate the visual correlation between the model parameters and learned representations with Class Activation Maps (CAM). Lastly, we explored another CNN model that was trained from scratch, which takes the raw audio data in the frequency domain as an input, finding some frequency patterns discriminative for each personality trait. The interpretability part reveals the inter-mechanism of the model, showing that some frequency bands are more discriminative for personality trait recognition than others.

Bibliography

- [1] RW Picard. *Affective computing* cambridge. MA: MIT Press [Google Scholar], 1997.
- [2] William James. *The emotions*. 1922.
- [3] James R Averill. *A semantic atlas of emotional concepts*. American Psycholog. Ass., Journal Suppl. Abstract Service, 1975.
- [4] James R Averill. A constructivist view of emotion. In *Theories of emotion*, pages 305–339. Elsevier, 1980.
- [5] Carroll E Izard. Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Annual review of psychology*, 60:1–25, 2009.
- [6] Jennifer S Lerner, Ye Li, Piercarlo Valdesolo, and Karim S Kassam. Emotion and decision making. *Annual review of psychology*, 66(1), 2015.
- [7] Edmund T Rolls. *Emotion and decision-making explained*. OUP Oxford, 2013.
- [8] Chai M Tyng, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik. The influences of emotion on learning and memory. *Frontiers in psychology*, 8:1454, 2017.
- [9] Rosalind W Picard. *Affective computing*. MIT press, 2000.
- [10] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

Conclusion

- [11] Rosalind W Picard. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1-2):55–64, 2003.
- [12] Paul R Kleinginna and Anne M Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379, 1981.
- [13] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 2022.
- [14] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.
- [15] Paul Ekman, Wallace V Friesen, Maureen O’sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.
- [16] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200, 1992.
- [17] David Watson and Auke Tellegen. Toward a consensual structure of mood. *Psychological bulletin*, 98(2):219, 1985.
- [18] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015.
- [19] Behzad Hasani and Mohammad H Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Proceedings of*

Conclusion

- the IEEE conference on computer vision and pattern recognition workshops*, pages 30–40, 2017.
- [20] Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Björn Schuller. Speech emotion classification using attention-based lstm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1675–1685, 2019.
- [21] Pengcheng Wei and Yu Zhao. A novel speech emotion recognition algorithm based on wavelet kernel sparse classifier in stacked deep auto-encoder model. *Personal and Ubiquitous Computing*, 23(3):521–529, 2019.
- [22] Fatemeh Noroozi, Dorota Kaminska, Ciprian Corneanu, Tomasz Sapinski, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, 2018.
- [23] Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, and Gholamreza Anbarjafari. Emotion recognition from skeletal movements. *Entropy*, 21(7):646, 2019.
- [24] Peixiang Zhong and Chunyan Miao. ntuer at semeval-2019 task 3: Emotion classification with word and sentence representations in rcnn. *arXiv preprint arXiv:1902.07867*, 2019.
- [25] Armin Seyeditabari, Narges Tabari, Shafie Gholizadeh, and Wlodek Zadrozny. Emotion detection in text: focusing on latent representation. *arXiv preprint arXiv:1907.09369*, 2019.
- [26] Kiri Wagstaff. Machine learning that matters. *arXiv preprint arXiv:1206.4656*, 2012.
- [27] Cynthia Rudin and Kiri L Wagstaff. *Machine learning for science and society*, 2014.
- [28] Samuele Lo Piano. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, 7(1):1–7, 2020.

Conclusion

- [29] Maryam Ashoori and Justin D Weisz. In ai we trust? factors that influence trustworthiness of ai-infused decision-making processes. *arXiv preprint arXiv:1912.02675*, 2019.
- [30] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31(2):447–464, 2021.
- [31] David Spiegelhalter. Should we trust algorithms? 2020.
- [32] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.
- [33] Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.
- [34] Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part I*, volume 8188. Springer, 2013.
- [35] V John Mathews and Zhenhua Xie. A stochastic gradient adaptive filter with gradient adaptive step size. *IEEE transactions on Signal Processing*, 41(6):2075–2087, 1993.
- [36] Guannan Qu and Na Li. Accelerated distributed nesterov gradient descent for smooth and strongly convex functions. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 209–216. IEEE, 2016.
- [37] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.
- [38] Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, 2019.

Conclusion

- [39] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336, 2020.
- [40] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.
- [41] Georgios Rizos and Björn W Schuller. Average jane, where art thou?—recent avenues in efficient machine learning under subjectivity uncertainty. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 42–55. Springer, 2020.
- [42] Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 932–939. JMLR Workshop and Conference Proceedings, 2010.
- [43] Maarit Koponen, Wilker Aziz, Luciana Ramos, and Lucia Specia. Post-editing time as a measure of cognitive effort. In *Workshop on Post-Editing Technology and Practice*, 2012.
- [44] Dogs vs. Cats, 2013. dataset downloaded from Kaggle competition, [http://https://www.kaggle.com/c/dogs-vs-cats](https://www.kaggle.com/c/dogs-vs-cats).
- [45] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173, 2008.
- [46] Niall Twomey, Tom Diethe, Meelis Kull, Hao Song, Massimo Camplani, Sion Hannuna, Xenofon Fafoutis, Ni Zhu, Pete Woznowski, Peter Flach, et al. The sphere challenge: Activity recognition with multimodal sensor data. *arXiv preprint arXiv:1603.00797*, 2016.

Conclusion

- [47] Flávio Luis De Mello and Sebastião Alves de Souza. Psychotherapy and artificial intelligence: A proposal for alignment. *Frontiers in psychology*, 10: 263, 2019.
- [48] Morton Wagman. *Computer psychotherapy systems: Theory and research foundations*. Routledge, 2018.
- [49] David D Luxton. An introduction to artificial intelligence in behavioral and mental health care. In *Artificial intelligence in behavioral and mental health care*, pages 1–26. Elsevier, 2016.
- [50] Elaheh Yadegaridehkordi, Nurul Fazmidar Binti Mohd Noor, Mohamad Nizam Bin Ayub, Hannyzzura Binti Affal, and Nornazlita Binti Hussin. Affective computing in education: A systematic review and future research. *Computers & Education*, 142:103649, 2019.
- [51] Chih-Hung Wu, Yueh-Min Huang, and Jan-Pan Hwang. Review of affective computing in education/learning: Trends and challenges. *British Journal of Educational Technology*, 47(6):1304–1323, 2016.
- [52] Olga C Santos. Emotions and personality in adaptive e-learning systems: an affective computing perspective. In *Emotions and personality in personalized services*, pages 263–285. Springer, 2016.
- [53] Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas By. Sentiment analysis on social media. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 919–926. IEEE, 2012.
- [54] Carlos A. Iglesias and Antonio Moreno. *Sentiment Analysis for Social Media*. MDPI, apr 2020. doi: 10.3390/books978-3-03928-573-0. URL <https://doi.org/10.3390/books978-3-03928-573-0>.
- [55] Jan Mizgajski and Mikołaj Morzy. Affective recommender systems in online news industry: how emotions influence reading choices. *User Modeling and User-Adapted Interaction*, 29(2):345–379, 2019.

Conclusion

- [56] Rafael A Calvo and Sidney D’Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010.
- [57] Rainer Reisenzein. Broadening the scope of affect detection research. *IEEE Transactions on Affective Computing*, 1(1):42–45, 2010.
- [58] Junzheng Li, Wei Zhu, Yanchun Yang, and Xiyuan Zheng. A cross-media retrieval method based on semisupervised learning and alternate optimization. *Mobile Information Systems*, 2021, 2021.
- [59] Jie Cao, Shengsheng Qian, Huaiwen Zhang, Quan Fang, and Changsheng Xu. Global relation-aware attention network for image-text retrieval. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 19–28, 2021.
- [60] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60): 16, 1999.
- [61] Jesse J Prinz. *Gut reactions: A perceptual theory of emotion*. oxford university Press, 2004.
- [62] Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer, 2012.
- [63] Paul Ekman, E Richard Sorenson, and Wallace V Friesen. Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88, 1969.
- [64] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.
- [65] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Lisbon, Portugal, 2004.
- [66] Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michał R Wróbel. Modeling emotions for affect-aware applications. *Information Systems Development and Applications*, pages 55–69, 2015.

Conclusion

- [67] Klaus R Scherer and Paul Ekman. *Approaches to emotion*. Psychology Press, 2014.
- [68] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. *The measurement of meaning*. Number 47. University of Illinois press, 1957.
- [69] Lisa A Feldman. Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of personality and social psychology*, 69(1):153, 1995.
- [70] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.
- [71] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.
- [72] Michelle SM Yik, James A Russell, and Lisa Feldman Barrett. Structure of self-reported current affect: Integration and beyond. *Journal of personality and social psychology*, 77(3):600, 1999.
- [73] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.
- [74] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011.
- [75] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

Conclusion

- [76] Ellen Douglas-Cowie, Cate Cox, Jean-Claude Martin, Laurence Devillers, Roddy Cowie, Ian Sneddon, Margaret McRorie, Catherine Pelachaud, Christopher Peters, Orla Lowry, et al. The humaine database. In *Emotion-Oriented Systems*, pages 243–284. Springer, 2011.
- [77] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and emotion*, 24(7):1153–1172, 2010.
- [78] Yoann Baveye, Jean-Noël Bettinelli, Emmanuel Dellandréa, Liming Chen, and Christel Chamaret. A large video database for computational models of induced emotion. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 13–18. IEEE, 2013.
- [79] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quang, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-Hélène Demarty, and Liming Chen. The mediaeval 2015 affective impact of movies task. In *MediaEval*, 2015.
- [80] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.
- [81] Athanasia Zlatintsi, Petros Koutras, Georgios Evangelopoulos, Nikolaos Malandrakis, Niki Efthymiou, Katerina Pastra, Alexandros Potamianos, and Petros Maragos. Cognimuse: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, 2017(1):1–24, 2017.
- [82] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, 2007.

Conclusion

- [83] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.
- [84] Paul Vicol, Makarand Tapaswi, Lluís Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8581–8590, 2018.
- [85] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.
- [86] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.
- [87] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010.
- [88] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. Emotional speech: Towards a new generation of databases. *Speech communication*, 40(1-2):33–60, 2003.
- [89] Ellen Douglas-Cowie, Laurence Devillers, Jean-Claude Martin, Roddy Cowie, Suzie Savvidou, Sarkis Abrilian, and Cate Cox. Multimodal databases of everyday emotion: Facing up to complexity. In *Ninth European conference on speech communication and technology*, 2005.
- [90] Sarkis Abrilian, Laurence Devillers, S Buisine, and Jean-Claude Martin. Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *HCI International*, volume 401, pages 407–408, 2005.

Conclusion

- [91] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech communication*, 49(10-11):787–800, 2007.
- [92] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. The vera am mittag german audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo*, pages 865–868. IEEE, 2008.
- [93] Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 505–510, 2013.
- [94] Geoffrey Leech. Introducing corpus annotation. *Corpus annotation: Linguistic information from computer text corpora*, pages 1–18, 1997.
- [95] Petra Saskia Bayerl and Karsten Ingmar Paul. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725, 2011.
- [96] Karën Fort, Claire François, Olivier Galibert, and Maha Ghribi. Analyzing the impact of prevalence on the evaluation of a manual annotation campaign. In *International Conference on Language Resources and Evaluation (LREC)*, 2012.
- [97] Stephanie Glen. Correlation coefficient: Simple definition, formula, easy steps. *StatisticsHowTo.com*. Available online: <https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/> (accessed on 3 August 2020), 2021.
- [98] Charles Spearman. The proof and measurement of association between two things. 1961.

Conclusion

- [99] Yue Zhang, Andrea Michi, Johannes Wagner, Elisabeth André, Björn Schuller, and Felix Weninger. A generic human-machine annotation framework based on dynamic cooperative learning. *IEEE Transactions on Cybernetics*, 50(3):1230–1239, 2019.
- [100] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.
- [101] Jean Véronis. A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*, pages 2–4. Citeseer, 1998.
- [102] Klaus Krippendorff. Computing krippendorff’s alpha-reliability.(2011). *Annenberg School for Communication Departmental Papers: Philadelphia*, 2011.
- [103] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.
- [104] William A Scott. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325, 1955.
- [105] Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th linguistic annotation workshop*, pages 92–100, 2011.
- [106] Kilem Gwet. Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters. *Gaithersburg, MD: STATAXIS Publishing Company*, 2001.
- [107] Jet Hoek, Merel CJ Scholman, and Ted JM Sanders. Is there less annotator agreement when the discourse relation is underspecified?
- [108] Andrew Rosenberg and Ed Binkowski. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. 2004.

Conclusion

- [109] Kateryna Ignatova, Cigdem Toprak, Delphine Bernhard, and Iryna Gurevych. Annotating question types in social q&a sites. In *Tagungsband des GSCL Symposiums 'Sprachtechnologie und eHumanities*, pages 44–49. Citeseer, 2009.
- [110] Nourah Alswaidan and Mohamed El Bachir Menai. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987, 2020.
- [111] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. An overview on image sentiment analysis: Methods, datasets and current challenges. *ICETE (1)*, pages 296–306, 2019.
- [112] Anvita Saxena, Ashish Khanna, and Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79, 2020.
- [113] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010.
- [114] Pablo Morales-Álvarez, Pablo Ruiz, Raúl Santos-Rodríguez, Rafael Molina, and Aggelos K Katsaggelos. Scalable and efficient learning from crowds with gaussian processes. *Information Fusion*, 52:110–127, 2019.
- [115] Trevor Cohn and Lucia Specia. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, 2013.
- [116] Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [117] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. Beyond black & white: Leveraging annotator

Conclusion

- disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, 2021.
- [118] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *2016 international joint conference on neural networks (IJCNN)*, pages 566–570. IEEE, 2016.
- [119] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- [120] Huang-Cheng Chou and Chi-Chun Lee. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5886–5890. IEEE, 2019.
- [121] Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima, and Yushi Aono. Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4964–4968. IEEE, 2018.
- [122] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2227–2231. IEEE, 2017.
- [123] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- [124] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*, pages 845–850, 2015.

Conclusion

- [125] Yongxin Yang and Timothy M Hospedales. Trace norm regularised deep multi-task learning. *arXiv preprint arXiv:1606.04038*, 2016.
- [126] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997.
- [127] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Personalized multitask learning for predicting tomorrow’s mood, stress, and health. *IEEE Transactions on Affective Computing*, 11(2):200–213, 2017.
- [128] Daniel Lopez-Martinez and Rosalind Picard. Multi-task neural networks for personalized pain recognition from physiological signals. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 181–184. IEEE, 2017.
- [129] Carl Edward. Rasmussen and christopher ki williams. gaussian processes for machine learning. *MIT Press*, 211:212, 2006.
- [130] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [131] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.
- [132] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- [133] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.
- [134] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018.

Conclusion

- [135] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018.
- [136] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [137] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
- [138] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.
- [139] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [140] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.
- [141] Hassan Hayat, Carles Ventura, and Agata Lapedriza. Recognizing emotions evoked by movies using multitask learning. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

Conclusion

- [142] Tuan-Linh Nguyen, Swathi Kavuri, and Minhoo Lee. A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips. *Neural Networks*, 118:208–219, 2019.
- [143] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [144] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [145] American Psychological Association (APA). Personality traits definition, 2017. URL <https://dictionary.apa.org/personality-trait>.
- [146] David C. Funder. Personality. *Annual Review of Psychology*, 52(1):197–221, 2001. doi: 10.1146/annurev.psych.52.1.197. URL <https://doi.org/10.1146/annurev.psych.52.1.197>. PMID: 11148304.
- [147] Daniel J Ozer and Veronica Benet-Martinez. Personality and the prediction of consequential outcomes. *Annual review of psychology*, 57:401, 2006.
- [148] James S Uleman, S Adil Saribay, and Celia M Gonzalez. Spontaneous inferences, implicit impressions, and implicit theories. *Annu. Rev. Psychol.*, 59: 329–360, 2008.
- [149] David C Funder. Global traits: A neo-allportian approach to personality. *Psychological Science*, 2(1):31–39, 1991.
- [150] Helen Palmer. The enneagram: Understanding yourself and the others in your life (harpersanfrancisco), 1991.
- [151] Heather EP Cattell and Alan D Mead. The sixteen personality factor questionnaire (16pf). *The SAGE handbook of personality theory and assessment*, 2:135–159, 2008.
- [152] Hans Jurgen Eysenck. *A model for personality*. Springer Science & Business Media, 2012.

Conclusion

- [153] Isabel Briggs Myers. The myers-briggs type indicator: Manual (1962). 1962.
- [154] Dan P McAdams. The five-factor model in personality: A critical appraisal. *Journal of personality*, 60(2):329–361, 1992.
- [155] Warren T Norman. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The journal of abnormal and social psychology*, 66(6):574, 1963.
- [156] Dean Peabody and Lewis R Goldberg. Some determinants of factor structures from personality-trait descriptors. *Journal of personality and social psychology*, 57(3):552, 1989.
- [157] Lewis R Goldberg. An alternative” description of personality”: the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990.
- [158] Gordon W Allport and Henry S Odbert. Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1):i, 1936.
- [159] Robert Hogan, Gordon J Curphy, and Joyce Hogan. What we know about leadership: Effectiveness and personality. *American psychologist*, 49(6):493, 1994.
- [160] Adrian Furnham, Chris J Jackson, and Tony Miller. Personality, learning style and work performance. *Personality and individual differences*, 27(6):1113–1122, 1999.
- [161] Jon F Sigurdsson. Computer experience, attitudes toward computers and personality characteristics in psychology undergraduates. *Personality and Individual Differences*, 12(6):617–624, 1991.
- [162] J Philippe Rushton, Harry G Murray, and Stephen Erdle. Combining trait consistency and learning specificity approaches to personality, with illustrative data on faculty teaching performance. *Personality and Individual Differences*, 8(1):59–66, 1987.

Conclusion

- [163] Adrian Furnham and Jean Mitchell. Personality, needs, social skills and academic achievement: A longitudinal study. *Personality and Individual Differences*, 12(10):1067–1073, 1991.
- [164] Meera Komarraju and Steven J Karau. The relationship between the big five personality traits and academic motivation. *Personality and individual differences*, 39(3):557–567, 2005.
- [165] Adam A Augustine, Scott H Hemenover, Randy J Larsen, and Tirza E Shulman. Composition and consistency of the desired affective state: The role of personality and motivation. *Motivation and Emotion*, 34(2):133–143, 2010.
- [166] Lameese Eldesouky and Tammy English. Individual differences in emotion regulation goals: Does personality predict the reasons why people regulate their emotions? *Journal of personality*, 87(4):750–766, 2019.
- [167] Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003.
- [168] Henry Kellerman. Emotion and the organization of primary process. In *Emotion, psychopathology, and psychotherapy*, pages 89–113. Elsevier, 1990.
- [169] Philip J Corr. Reinforcement sensitivity theory (rst): Introduction. 2008.
- [170] Charles D Spielberger, Sumner J Sydeman, Ashley E Owen, and Brian J Marsh. *Measuring anxiety and anger with the State-Trait Anxiety Inventory (STAI) and the State-Trait Anger Expression Inventory (STAXI)*. Lawrence Erlbaum Associates Publishers, 1999.
- [171] Auke Tellegen, David Watson, and Lee Anna Clark. On the dimensional and hierarchical structure of affect. *Psychological science*, 10(4):297–303, 1999.
- [172] Paula M Niedenthal, Nathalie Dalle, and Anette Rohmann. Emotional response categorization as emotionally intelligent behavior. 2002.

Conclusion

- [173] Ainize Peña-Sarrionandia, Moira Mikolajczak, and James J Gross. Integrating emotion regulation and emotional intelligence traditions: a meta-analysis. *Frontiers in psychology*, 6:160, 2015.
- [174] Marie T Dasborough. Emotional intelligence as a moderator of emotional responses to leadership. In *Emotions and leadership*, volume 15, pages 69–88. Emerald Publishing Limited, 2019.
- [175] Peter Salovey and John D Mayer. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211, 1990.
- [176] John D Mayer. Controversies in emotional intelligence. 2004.
- [177] Kostantinos V Petrides and Adrian Furnham. Trait emotional intelligence: Psychometric investigation with reference to established trait taxonomies. *European journal of personality*, 15(6):425–448, 2001.
- [178] Kostantinos V Petrides. Ability and trait emotional intelligence. 2011.
- [179] Alexandra Martins, Nelson Ramalho, and Estelle Morin. A comprehensive meta-analysis of the relationship between emotional intelligence and health. *Personality and individual differences*, 49(6):554–564, 2010.
- [180] Federica Andrei, Giacomo Mancini, Bruno Baldaro, Elena Trombini, and Sergio Agnoli. A systematic review on the predictive utility of the trait emotional intelligence questionnaire (teique). *BPA-Applied Psychology Bulletin (Bollettino di Psicologia Applicata)*, 62(271), 2014.
- [181] Donald H Saklofske, Elizabeth J Austin, and Paul S Minski. Factor structure and validity of a trait emotional intelligence measure. *Personality and Individual differences*, 34(4):707–721, 2003.
- [182] Kenneth S Kendler, John M Myers, and Corey LM Keyes. The relationship between the genetic and environmental influences on common externalizing psychopathology and mental wellbeing. *Twin Research and Human Genetics*, 14(6):516–523, 2011.

Conclusion

- [183] Alexander B Siegling, Adrian Furnham, and Konstantinos V Petrides. Trait emotional intelligence and personality: Gender-invariant linkages across different measures of the big five. *Journal of psychoeducational assessment*, 33(1):57–67, 2015.
- [184] Juan Carlos Pérez-González and Maria-Jose Sanchez-Ruiz. Trait emotional intelligence anchored within the big five, big two and big one frameworks. *Personality and Individual Differences*, 65:53–58, 2014.
- [185] Ramón López-Cózar, Zoraida Callejas, David Griol, and José F Quesada. Review of spoken dialogue systems. *Loquens*, 1(2):012, 2014.
- [186] Salla Syvänen and Chiara Valentini. Conversational agents in online organization–stakeholder interactions: a state-of-the-art analysis and implications for further research. *Journal of Communication Management*, 24(4):339–362, 2020.
- [187] Robert Dale. The return of the chatbots. *Natural Language Engineering*, 22(5):811–817, 2016.
- [188] Gina Neff. Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 2016.
- [189] Alexander Maedche, Christine Legner, Alexander Benlian, Benedikt Berger, Henner Gimpel, Thomas Hess, Oliver Hinz, Stefan Morana, and Matthias Söllner. Ai-based digital assistants. *Business & Information Systems Engineering*, 61(4):535–544, 2019.
- [190] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.
- [191] Kenneth Mark Colby. PARRY. <https://www.chatbots.org/chatbot/parry/>, 1971. [Online; accessed 15-August-2022].
- [192] Rodrigo Bavaresco, Diórgenes Silveira, Eduardo Reis, Jorge Barbosa, Rodrigo Righi, Cristiano Costa, Rodolfo Antunes, Marcio Gomes, Clauter

Conclusion

- Gatti, Mariangela Vanzin, et al. Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review*, 36:100239, 2020.
- [193] Xiaojie Wang and Caixia Yuan. Recent advances on human-computer dialogue. *CAAI Transactions on Intelligence Technology*, 1(4):303–312, 2016.
- [194] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [195] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.
- [196] Xianda Zhou and William Yang Wang. Mojitalk: Generating emotional responses at scale. *arXiv preprint arXiv:1711.04090*, 2017.
- [197] Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.
- [198] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32, 2020.
- [199] Janneke M Van der Zwaan, Virginia Dignum, and Catholijn M Jonker. A bdi dialogue agent for social support: Specification and evaluation method. 2012.
- [200] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.

Conclusion

- [201] Pawel Dybala, Michal Ptaszynski, Rafal Rzepka, and Kenji Araki. Activating humans with humor—a dialogue system that users want to interact with. *IEICE TRANSACTIONS on Information and Systems*, 92(12):2394–2401, 2009.
- [202] Russell Beale and Chris Creed. Affective interaction: How emotional agents affect users. *International journal of human-computer studies*, 67(9):755–776, 2009.
- [203] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.
- [204] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191, 2020.
- [205] Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model. *arXiv preprint arXiv:2005.05298*, 2020.
- [206] Yunyi Yang, Yunhao Li, and Xiaojun Quan. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238, 2021.
- [207] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.
- [208] Xinyan Zhao, Feng Xiao, Haoming Zhong, Jun Yao, and Huanhuan Chen. Condition aware and revise transformer for question answering. In *Proceedings of The Web Conference 2020*, pages 2377–2387, 2020.

Conclusion

- [209] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, 2020.
- [210] Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, and Jiaxing Shen. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5010–5020, 2021.
- [211] Renee D Goodwin and Howard S Friedman. Health status and the five-factor personality traits in a nationally representative sample. *Journal of health psychology*, 11(5):643–654, 2006.
- [212] Sarah E Hampson, Lewis R Goldberg, Thomas M Vogt, and Joan P Dubanoski. Forty years on: teachers’ assessments of children’s personality traits predict self-reported health behaviors and outcomes at midlife. *Health psychology*, 25(1):57, 2006.
- [213] Dave Korotkov and T Edward Hannah. The five-factor model of personality: Strengths and limitations in predicting health status, sick-role and illness behaviour. *Personality and Individual Differences*, 36(1):187–199, 2004.
- [214] Timothy W Smith. Personality as risk and resilience in physical health. *Current directions in psychological science*, 15(5):227–231, 2006.
- [215] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. Affect-driven dialog generation. *arXiv preprint arXiv:1904.02793*, 2019.
- [216] Rohola Zandie and Mohammad H Mahoor. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. In *The Thirty-Third International Flairs Conference*, 2020.
- [217] Peixiang Zhong, Yan Zhu, Yong Liu, Chen Zhang, Hao Wang, Zaiqing Nie, and Chunyan Miao. Endowing empathetic conversational models with personas. *arXiv preprint arXiv:2004.12316*, 2020.

Conclusion

- [218] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*, 2019.
- [219] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer, 2018.
- [220] Jia Li, Xiao Sun, Xing Wei, Changliang Li, and Jianhua Tao. Reinforcement learning based emotional editing constraint conversation generation. *arXiv preprint arXiv:1904.08061*, 2019.
- [221] Xiao Sun, Xinmiao Chen, Zhengmeng Pei, and Fuji Ren. Emotional human machine conversation generation based on seqgan. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE, 2018.
- [222] Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1401–1410, 2019.
- [223] Mark A Thornton and Diana I Tamir. Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences*, 114(23):5982–5987, 2017.
- [224] Ameneh Gholipour Shahraki and Osmar R Zaïane. Lexical and learning-based emotion mining from text. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, 2017.
- [225] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.

Conclusion

- [226] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.
- [227] Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Dario Bertero, Yan Wan, Ricky Ho Yin Chan, and Chien-Sheng Wu. Zara: a virtual interactive dialogue system incorporating emotion, sentiment and personality recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 278–281, 2016.
- [228] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.
- [229] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.
- [230] Sayyed M Zahiri and Jinho D Choi. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aai conference on artificial intelligence*, 2018.
- [231] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [232] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.
- [233] Andrew S Fox, Regina C Lapate, Alexander J Shackman, and Richard J Davidson. *The nature of emotion: Fundamental questions*. Oxford University Press, 2018.

Conclusion

- [234] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.
- [235] Albert Mehrabian. Analysis of the big-five personality factors in terms of the pad temperament model. *Australian journal of Psychology*, 48(2):86–92, 1996.
- [236] Lisa Feldman Barrett and James A Russell. The structure of current affect: Controversies and emerging consensus. *Current directions in psychological science*, 8(1):10–14, 1999.
- [237] Ranier Reisenzein. Pleasure-arousal theory and the intensity of emotions. *Journal of personality and social psychology*, 67(3):525, 1994.
- [238] Justin Storbeck and Gerald L Clore. Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Social and personality psychology compass*, 2(5):1824–1843, 2008.
- [239] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.
- [240] Xiaoming Zhao, Zhiwei Tang, and Shiqing Zhang. Deep personality trait recognition: A survey. *Frontiers in Psychology*, page 2390, 2022.
- [241] Julio Cezar Silveira Jacques Junior, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel AJ Van Gerven, Rob Van Lier, et al. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing*, 2019.
- [242] Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4):2313–2339, 2020.
- [243] Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset

Conclusion

- and results. In *European conference on computer vision*, pages 400–418. Springer, 2016.
- [244] Xiu-Shen Wei, Chen-Lin Zhang, Hao Zhang, and Jianxin Wu. Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, 9(3):303–315, 2017.
- [245] Bruce L Brown, William J Strong, and Alvin C Rencher. Acoustic determinants of perceptions of personality from speech. 1975.
- [246] Ernest Kramer. Judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychological Bulletin*, 60(4):408, 1963.
- [247] David W Addington. The relationship of selected vocal characteristics to personality perception. 1968.
- [248] Gelareh Mohammadi and Alessandro Vinciarelli. Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, 3(3):273–284, 2012.
- [249] Jean-Philippe Goldman. Easyalign: an automatic phonetic alignment tool under praat. In *Interspeech’11, 12th Annual Conference of the International Speech Communication Association*, 2011.
- [250] Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. Hi youtube! personality impressions and verbal content in social video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 119–126, 2013.
- [251] Joan-Isaac Biel and Daniel Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2012.
- [252] Arulkumar Subramaniam, Vismay Patel, Ashish Mishra, Prashanth Balasubramanian, and Anurag Mittal. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *European conference on computer vision*, pages 337–348. Springer, 2016.

Conclusion

- [253] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.
- [254] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.
- [255] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5089–5093. IEEE, 2018.
- [256] Aharon Satt, Shai Rozenberg, and Ron Hoory. Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech*, pages 1089–1093, 2017.
- [257] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*, 2018.
- [258] Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Julio Jacques, Meysam Madadi, Xavier Baró, Stephane Ayache, Evelyne Viegas, Yağmur Güçlütürk, Umut Güçlü, et al. Design of an explainable machine learning challenge for video interviews. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3688–3695. IEEE, 2017.
- [259] Baiyu Chen, Sergio Escalera, Isabelle Guyon, Víctor Ponce-López, Nihar Shah, and Marc Oliu Simón. Overcoming calibration problems in pattern labeling with pairwise ratings: application to personality traits. In *European Conference on Computer Vision*, pages 419–432. Springer, 2016.
- [260] Julio Junior, CS Jacques, Agata Lapedriza, Cristina Palmero, Xavier Baro, and Sergio Escalera. Person perception biases exposed: Revisiting the first impressions dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 13–21, 2021.

Conclusion

- [261] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. *Deep learning. Nature*, 521(7553):436–444, 2015.
- [262] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [263] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.
- [264] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.
- [265] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.
- [266] Carles Ventura, David Masip, and Agata Lapedriza. Interpreting CNN models for apparent personality trait regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 55–63, 2017.
- [267] Michael Heideman, Don Johnson, and Charles Burrus. Gauss and the history of the fast Fourier transform. *IEEE ASSP Magazine*, 1(4):14–21, 1984.