Implicit Design Choices and Their Impact on Emotion Recognition Model Development and Evaluation

by

Mimansa Jaiswal

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in The University of Michigan
2023

Doctoral Committee:

 Professor Emily Mower Provost, Chair
 Professor Nikola Bancovic
 Professor Benjamin Fish
 Douwe Kiela, Contextual AI
 Professor V.G. Vinod Vydiswaran

Mimansa Jaiswal

mimansa@umich.edu

ORCID iD: 0009-0001-8290-7743

Dedicated to my parents and my late granddad, B.P. Bhagat, and my late grandmom, Smt.

Radha Devi for their boundless love and support

ACKNOWLEDGEMENTS

I would like to start by expressing my profound gratitude towards my PhD advisor, Emily *(Emily Mower Provost)*. She has been my go-to person for any research brainstorming and has shown me tremendous patience, support, and guidance throughout my PhD journey. Without her persistence and suggestions, completing this PhD would not have been possible. I am also incredibly grateful for my thesis committee members: Vinod *(VG Vinod Vydiswaran)*, Nikola *(Nikola Banovic)*, Benjamin *(Benjamin Fish)*, and Douwe *(Douwe Kiela)*. Their valuable insights during my thesis proposal helped shape the final version of the thesis.

For putting up with me during my PhD journey, my immense gratitude goes towards my family, especially my parents, grandparents, and Bert. My parents have been a rock for me over the past six years. Though I could not visit them often or talk to them much, they were always there when I needed someone. They seem to have aged fifteen years in the six years of my PhD, stressed about me, but their support never wavered. My mom *(Archana Kumari)* received her own PhD in 2021, and my dad *(Abhay Kumar)* became the Vice-Chancellor of a new IIIT—two accomplishments that were their lifelong dreams and inspired me immensely. I unfortunately lost two of my grandparents during the PhD program, and I will never forget their blessings and excitement for me embarking on my higher education journey. During early 2020, in the midst of COVID, I adopted a cat named Bert—yes, named after the language model. Without him, I would not have maintained my sanity during the dark, lonely nights and tiring, long work days. His purring loudly into my ear calmed me down on the worst of nights.

I was lucky enough to secure three internships and have amazing research mentors for

way or format. Conversations with him have always left me rejuvenated, happy, and feeling peppier—a testament to how amazing a best friend he is. Sagarika *(Sagarika Srishti)*, for all her support, both in India and when she came to the US. Her move to the US during my PhD was a major personal highlight. Ariba *(Ariba Javed)*, thanks for all the discussions, talks, and emotional conversations, and for always being up for anything interesting, including a pottery class. Shobhit *(Shobhit Narain)* has been an amazing companion, helping me with job applications and always being the sarcastic, serious, yet most helpful guy I have had the pleasure of calling a friend. And finally, Sai *(Sairam Tabibu)* helped me fill out the PhD application for UMich on the exact deadline, without which, I would not be here at all.

This is probably an unconventional paragraph in acknowledgments, but these were unconventional times during COVID. For the two years of lockdown, I turned to Among Us when I felt lonely or lost in my research. I am really thankful for the streamers whose broadcasts provided some semblance of social interaction. For almost three years, I watched them stream at least 8 hours a day while I worked, to simulate a social environment. And when my research progress stalled, I turned to anonymous Discord communities, playing Among Us and golf for hours, which helped alleviate feelings of depression and sadness, providing a much-needed uplift.

My PhD journey wasn't easy, and a lot happened over the six years, but I made it through. The credit for that goes to all the people mentioned here, to whom I am forever indebted.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Emotion recognition is a complex task due to the inherent subjectivity in both the perception and production of emotions. The subjectivity of emotions poses significant challenges in developing accurate and robust computational models. This thesis examines critical facets of emotion recognition, beginning with the collection of diverse datasets that account for psychological factors in emotion production. To address these complexities, the thesis makes several key contributions.

To handle the challenge of non-representative training data, this work collects the Multimodal Stressed Emotion dataset, which introduces controlled stressors during data collection to better represent real-world influences on emotion production. To address issues with label subjectivity, this research comprehensively analyzes how data augmentation techniques and annotation schemes impact emotion perception and annotator labels. It further handles natural confounding variables and variations by employing adversarial networks to isolate key factors like stress from learned emotion representations during model training. For tackling concerns about leakage of sensitive demographic variables, this work leverages adversarial learning to strip sensitive demographic information from multimodal encodings. Additionally, it proposes optimized sociological evaluation metrics aligned with cost-effective, real-world needs for model testing.

The findings from this research provide valuable insights into the nuances of emotion labeling, modeling techniques, and interpretation frameworks for robust emotion recognition. The novel datasets collected help encapsulate the environmental and personal variability prevalent in real-world emotion expression. The data augmentation and annotation studies improve label consistency by accounting for subjectivity in emotion perception. The stressor-

controlled models enhance adaptability and generalizability across diverse contexts and datasets. The bimodal adversarial networks aid in generating representations that avoid leakage of sensitive user information. Finally, the optimized sociological evaluation metrics reduce reliance on extensive expensive human annotations for model assessment.

This research advances robust, practical emotion recognition through multifaceted studies of challenges in datasets, labels, modeling, demographic and membership variable encoding in representations, and evaluation. The groundwork has been laid for cost-effective, generalizable emotion recognition models that are less likely to encode sensitive demographic information.

CHAPTER I

Introduction

In human communication, perceiving and responding to others' emotions in interpersonal conversations play a crucial role [76]. To create systems that can aid in human-centered interpersonal situations, it is necessary for these systems to possess the capability to recognize emotions effectively [220]. Robust Emotion Recognition (ER) models can be beneficial in various situations, such as crisis text lines or passive mental health monitoring [160]. However, these ML models often lack robustness when faced with unseen data situations, making deploying them in high-risk situations or healthcare a challenging task [241].

Recongizing emotion is a challenging task because it is subjective in both perception and production [178]. The labels used to train emotion recognition models are perceptually subjective [28]. The same emotion can be perceived differently by different people, depending on their cultural background, personal experiences, and other factors [147]. Additionally, there is production subjectivity. The same emotion can be expressed differently by different people, depending on their individual personality, cultural background, physiological and other factors [13]. The subjectivity of emotion recognition makes it difficult to develop accurate and robust models that account for these numerous variations [221].

In addition to the challenges posed by subjectivity, there are challenges that relate to the information that is learned in addition to and beyond the expression of emotion itself. The manner in which emotions are expressed are correlated with a person's demographic and

identifying features. Hence, systems trained to recognize emotion can often learn implicit associations between an individual's demographic factors and emotion [208]. When used as a component in larger systems, these implicit associations can lead to either the leakage of demographic information, or can bias the larger system's output based on demographic information, even when not explicitly trained to do so.

Training any robust machine learning model necessitates having access to large amounts of diverse and labelled data. Training models for emotion recognition faces the challenge of not having access to large quantities of diverse data. Scraping data over the internet, as is done for other areas, leads to a dataset that is often demographically biased, and, often exaggerated for entertainment purposes. On the other hand, data collected in laboratory environments is intentionally cleaner and often exaggerated in case of scripted sessions. Therefore, both of these data collection methods do not encapsulate possible environmental and personal factors, which leads to models often being trained on either highly skewed or non-representative data. The resulting models are either fragile or biased, and ultimately unable to handle real-world variability.

In this dissertation, critical facets of emotion recognition are thoroughly explored, beginning with the collection of datasets, which take into account psychological factors in producing emotions. This is followed closely by examining the influence that alterations in data augmentation processes have on emotion labels, while also challenging and interrogating the validity of previously established labels. Alterations in labeling techniques and the resulting effects on annotator-assigned labels are also scrutinized. Simultaneously, the research develops robust models specifically trained to disregard certain physiological emotion production factors. Integral to the research is the creation of bimodal models that generate representations aiming to tackle the reduction of leakage of sensitive demographic variables. The concluding portion of the study involves an in-depth evaluation of the robustness and impartiality of these models, carried out in a human-centric manner, ensuring an emphasis on minimal costs for data annotation. From this extensive research, valuable

insights are gained into the complexities of emotion recognition, which pave the way for more nuanced and robust labeling, modeling, and interpretation techniques. It also lays the groundwork for future efforts in the development of robust and cost-effective emotion recognition models.

1.1 Emotion Theories and the Impact on Emotion Recognition Model Development

To better understand the subjectivity inherent in emotion recognition and its correlation with the research gaps and challenges, we must first explore the contrasts between emotion production and emotion perception theories. These theories elucidate the distinct factors related to the subjectivity of emotions in both production and recognition processes and offer valuable insights for developing robust and unbiased emotion recognition models.

1.1.1 Emotion Production and Emotion Perception

Emotion production refers to experiencing and generating emotional responses, encompassing several factors, including cognitive appraisal, physiological response, behavior and expression, and subjective experience. These components work together to create the unique process of producing emotions within each person.

Emotion perception, conversely, focuses on recognizing and interpreting others' emotional signals, influenced by factors such as emotional cues, context and environment, past experiences and learning, and individual differences. This process involves making sense of others' emotions based on various internal and external factors.

1.1.2 Emotion Theories, Research Challenges, and Implications for Emotion Recognition

Various theories of emotion provide insights into the challenges faced in developing computational models for emotion recognition in speech or text. Below, we discuss the relevance and implications of some prominent theories in the context of speech or text-based (bimodal) emotion recognition.

- James-Lange Theory and Cannon-Bard Theory [204]: Both theories emphasize physiological responses' importance in emotion. In speech or text-based recognition, it is vital to consider correlations between observable features (e.g., vocal tonality, speech patterns) and underlying physiological responses. Accounting for these correlations can help capture emotions, even though the relationship might be subjective due to personal and cultural differences.

- Schachter-Singer Two-Factor Theory [204]: This theory stresses the importance of both physiological arousal and cognitive appraisal for experiencing emotions. In speech or text-based emotion recognition, cognitive appraisal aspects such as semantic content, contextual factors, and discourse patterns can be extracted. However, the subjectivity of cognitive appraisal processes presents challenges given personal experiences' impact on interpretation.

- Lazarus Cognitive-Mediational Theory [204]: Centered around the role of cognitive appraisal, this theory highlights the need for emotion recognition systems to account for individuals' interpretations of situations through cues that may suggest appraisal (e.g., word choice, phrase structure, conversational context). Advanced models might need to factor in users' personal and demographic features to better understand cognitive appraisal processes. This approach introduces more subjectivity and potential privacy concerns, as individual perspectives and experiences can vary significantly.

Integrating insights from these theories can aid unraveling the complexities and subjective nature of emotions expressed through language, as speech or text-based emotion recognition relies primarily on linguistic patterns, tone, and content analysis.

1.1.3   Addressing Challenges Through Thesis Contributions

The thesis contributions align with and address the subjectivity challenges in emotion production and perception, thus tackling the complexities involved in developing robust and unbiased emotion recognition models.

- Collecting datasets that account for psychological factors in emotion production: By considering psychological factors influencing unique emotional experiences, more diverse datasets are created, allowing models to account for subjectivity in emotion production and generalize across emotions.

- Examining the influence of data augmentation processes on emotion perception labels: This contribution seeks to understand data augmentation's impact on ground truth labels, creating better representations of emotions in the datasets, accounting for subjectivity in emotion perception.

- Analyzing labeling setups' impact on annotators' emotion perception labels: This investigates how labeling setups influence emotion perception, aiming to improve label consistency and reduce inter-annotator disagreement, thus better representing subjectivity in emotion perception.

- Training robust models by explicitly disregarding emotion production factors: This minimizes the impact of subjective elements associated with emotion production, enabling models to focus on core emotional cues.

- Developing bimodal models for generation of emotion representations that are debiased and reduce encoding of demographic and membership information: This creates

models that consider multiple emotional cues while disregarding sensitive features, addressing subjectivity challenges in both emotion production and perception.

- Evaluating models in a human-centric manner: Designing evaluation methods aligned with real-world expectations and without incurring significant annotation costs ensures the models effectively tackle subjectivity challenges in a practical way.

By focusing on these contributions, the thesis emphasizes the connection between emotion production and perception's subjectivity and its influence on model development, advancing the creation of more robust and unbiased emotion recognition models.

1.2 Emotion Recognition

Emotion recognition models are customarily trained using laboratory-collected data encompassing video, audio, and corresponding text. These algorithms strive to capture the speaker's underlying emotional state either autonomously or as part of a larger pipeline, such as response generation. Supervised learning techniques predominantly train these models. Obtaining ground truth labels for the dataset samples is crucial for successfully training a supervised learning model. The emotion theories presented earlier are intrinsically linked with the complexity of emotion recognition. Understanding the interplay between these theories and model development is essential.

1.2.1 Emotion Labels

Emotion labels typically fall into two categories: categorical and dimensional. Categorical variables aim to discretely categorize emotion attributes, such as excitement, happiness, anger, or sadness. These labels' limitations align with the James-Lange and Cannon-Bard theories—emotions are subjective, making it difficult to define universal emotions across cultures. This subjectivity is intensified by both personal physiological responses to stimuli

6

and cultural context.

Dimensional emotional labels describe emotions across two dimensions, valence (sad to happy) and arousal (calm to excited). The dimensional approach is more consistent with the James-Lange and Cannon-Bard theories, addressing the physiological components of emotions, as well as the cognitive components emphasized by the Schachter-Singer Two-Factor Theory and Lazarus Cognitive-Mediational Theory. However, these dimensional labels also face the challenge of cultural and personal influences on the perception and expression of emotions.

1.2.2 Emotion Features

Three primary modalities are used in combination to train emotion recognition models: text, audio, and video. This thesis focuses predominantly on audio and its corresponding text as the feature set for these models.

Mel-filterbanks (MFBs) are often used as inputs to neural network models in speech. MFBs can capture correlations between vocal tonality, speech patterns, and underlying physiological responses. Nevertheless, factors like pitch, volume, or other nuances of speech may be affected by cultural and linguistic contexts. Furthermore, personal characteristics can influence these features, further complicating emotion recognition in cross-cultural or highly diverse settings.

Language features, which provide contextualized representations for words, capture the cognitive appraisal aspects (semantic content, contextual factors, and discourse patterns). The Lazarus Cognitive-Mediational Theory further highlights the need for models that account for user demographics. More advanced models may need to balance the understanding of individual emotions with ethical considerations.

1.2.3 Emotion Recognition Models

Audio-based emotion recognition models initially relied on Hidden Markov Models (HMMs) or Gaussian Mixture Models (GMMs) and later shifted focus to LSTMs and RNNs. These models aim to capture the dynamic and time-varying nature of speech, reflecting the James-Lange Theory and Cannon-Bard Theory's emphasis on physiological responses. However, these models must also account for the inherent cultural and linguistic differences in the way emotions are expressed through speech.

Language-based models, like recent advances in transformer architectures, address long and indirect contextual information challenges, in line with the Schachter-Singer Two-Factor Theory's cognitive appraisal aspects. These models strive to understand the nuances of language, cultural expressions, and individual semantic and contextual differences in recognizing emotions.

Multi-modal models exploit relevant information from text, audio, or video to form powerful emotion recognition models. Informed by the emotion theories, these models take into account the subjectivity of emotions by leveraging different modalities to discern the nuances of emotion expression. By combining these modes, models can better account for the emotional complexity that arises from intercultural and personal differences in perception, expression, and context.

1.3 Challenges in Emotion Recognition

The variable and subjective nature of emotions make it challenging to train models that can accurately identify emotion in any given scenario. Addressing three major challenges is necessary for any emotion recognition model deployed in a real-world setting: (a) Non-representative training data, (b) Subjective labels, (c) Unintentional encoding and leakage of sensitive information. Previous work has looked at varying ways to counter these challenges, talked about in detail in Chapter III, Section 2.1.

1.3.1 Non-representative data

Emotion production in real-world settings is influenced by various factors, including data collection settings, demographics, and personal factors. Addressing these confounding factors aligns with the implications of the earlier-discussed emotion theories. Researchers can tackle this challenge by developing more robust models, incorporating real-world variability through dataset augmentation or mitigating confounding factors.

1.3.2 Label Subjectivity

As highlighted in the emotion theories, emotions are inherently subjective and deeply influenced by personal experiences, culture, and context. This subjectivity leads to difficulty in pinpointing an objective and universal ground truth for training emotion recognition models. Researchers should account for label subjectivity by using diverse and representative datasets, annotations from multiple sources, and considering multiple emotion theories during the model design process.

1.3.3 Unintentional encoding and leakage of sensitive information

Variability can lead unintentional encoding and leakage of sensitive information concerns, specifically in human centered tasks, such as emotion recognition models, as the associative nature of the task and sensitive demographic variables may inadvertently lead to encoding personal information.

1.4 Proposed Methods

A robust and effective emotion recognition system must successfully navigate a range of challenges, including addressing subjectivity in emotion production and perception, handling natural variations and confounding variables, reducing encoded sensitive information, and providing relevant evaluation metrics. Here, we present a series of proposed methods aligned

with the outlined contributions to address these challenges.

1.4.1 Dataset Collection for Emotion Recognition

Tackling the challenge of subjectivity in emotion production, it's essential that we consider the issues in widely used emotion recognition datasets that arise due to design choices, methodology of data collection, and inherent subjectivity. Emotion datasets traditionally aim for minimal variation to ensure generalizability. However, this can result in non-robust models that struggle with unexpected variability. We propose the construction and validation of a new dataset called Multimodal Stressed Emotion (MuSE), which introduces a controlled situational confounder (stress) to better account for subjectivity. In addition, we discuss the use of domain adversarial networks to achieve more stable and reliable cross-corpus generalization while avoiding undesired characteristics in encodings.

1.4.2 Data Augmentation with Noise in Emotion Datasets

Addressing the challenge of subjectivity in emotion perception, we examine data augmentation with noise in emotion datasets, focusing on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, which features dyadic interactions with text, video, and audio modalities. Introducing realistic noisy samples through environmental and synthetic noise, we evaluate how ground truth and predicted labels change due to noise sources. We discuss the effects of commonly used noisy augmentation techniques on human emotion perception, potential inaccuracies in model robustness testing, and provide recommendations for noise-based augmentation and model deployment.

1.4.3 Annotations of Emotion Datasets

To further address subjectivity in emotion perception, we investigate how design choices in the annotation collection process impact the performance of trained models. Focusing on contextual biasing, we examine how annotators perceive emotions differently in the presence

or absence of context. Commonly-used emotion datasets often involve annotators who have knowledge of previous sentences, but models are frequently evaluated on individual utterances. We explore the implications of this discrepancy on model evaluation, and its potential for generating errors.

1.4.4 Methods for Handling Natural Variations and Confounding Variables

As mentioned earlier, we collect a dataset of differences in similar emotion production under varying levels of stress. Emotion recognition models may spuriously correlate these stress-based factors to perceived emotion labels, which could limit generalization to other datasets. Consequently, we hypothesize that controlling for stress variations can improve the models' generalizability. To achieve this, we employ adversarial networks to decorrelate stress modulations from emotion representations, examining the impact of stress on both acoustic and lexical emotion predictions. By isolating stress-related factors from emotion representations, we aim to enhance the model's ability to generalize across different stress conditions. Furthermore, we analyze the transferability of these refined emotion recognition models across various domains, assessing their adaptability to evolving contexts and scenarios. Ultimately, our approach aims to improve emotion recognition model robustness by addressing the inherent variability of emotional expression due to stress and ensuring greater applicability across multiple domains.

1.4.5 Approaches for Tackling Sensitive Information Leakage in Trained Emotion Recognition Models

Emotions are inherently related to demographic factors such as gender, age, and race. Consequently, emotion recognition models often learn these latent variables even if they are not explicitly trained to do so. This learning behavior poses a risk to user privacy, as the models inadvertently capture sensitive demographic information. Storing representations instead of raw data does not fully mitigate this issue, as latent variables can still compromise

11

user privacy. To address this challenge, we present approaches for mitigating the learning of certain demographic factors in emotion recognition embeddings. Furthermore, we tackle the issue of user-level membership identification by employing an adversarial network that strips this information from the final encoding, reduced leakage of sensitive information from generated representations.

1.4.6 Methods for Model Evaluation and Perception

Large language models face limitations in subjective tasks like emotion recognition due to inadequate annotation diversity and data coverage. Acquiring comprehensive annotations and evaluations is often costly and time-consuming. To address these challenges, we propose cost-effective sociological metrics for emotion generalization and reduced demographic vairable leakage. These metrics reduce reliance on expensive human-based feedback while still capturing the nuances of human emotions. By evaluating model performance and demographic variables encoded in generated representations, the proposed metrics improve cross-corpus results and allow for the development of accurate, relevant emotion recognition models in a more economic manner.

1.5 Contributions

This dissertation proposes several investigations and novel solutions to address various concerns related to real-world emotion recognition model deployment.

The contributions of the works in this dissertation can be summarized as follows:

- Chapter IV:

    - Introduction of Multimodal Stressed Emotion (MuSE) dataset.

    - Detailed data collection protocol.

    - Potential uses and emotion content annotations.

- Performance measuring baselines for emotion and stress classification.

- Chapter V:

  - Speech emotion recognition's impact under influence of various factors such as noise.

  - Investigation of noise-altered annotation labels and their aftermath.

  - Consequences on evaluation of ML models considering noise.

  - Specific recommendations for noise augmentations in emotion recognition datasets.

- Chapter VI:

  - Crowdsourced experiments to study the subjectivity in emotion expression and perception.

  - Contextual and randomized annotation schemes of the MuSE dataset.

  - Comparative analysis revealing contextual scheme's closeness to speaker's self-reported labels.

- Chapter VII:

  - Examination of emotion expressions under stress variations.

  - Utilization of adversarial networks to separate stress modulations from emotion representations.

  - Exploration of stress's impact on acoustic and lexical emotional predictions.

  - Evidence of improved generalizability with stress control during model training.

- Chapter VIII:

  - Highlighting the unintentional leak of sensitive demographic information in multimodal representations.

– Use of adversarial learning paradigm to improve sensitive information reduction metric.

– Maintenance of primary task performance, despite improvements to privacy.

• Chapter IX:

– New template formulation to derive human-centered, optimizable and cost-effective metrics.

– Correlation establishment between emotion recognition performance, biased representations and derived metrics.

– Employment of metrics for training an emotion recognition model with increased generalizability and decreased bias.

– Finding of positive correlation between proposed metrics and user preference.

1.6 Outline of the dissertation

Initiating with Chapter III, it delves into a comprehensive review of pertinent literature spanning from emotion recognition and privacy preservation to adversarial networks, model interpretability, and crowdsourcing designs. Moving forward, Chapter II provides an introduction to the common datasets, and features employed throughout this research. Subsequent chapters, from Chapter IV to IX, engage in a thorough exploration and discussion of the research work undertaken, characterized in the Contributions section. Lastly, Chapter X serves as a conclusive summary encapsulating the primary contributions made, elaborating on the proposed future works.

CHAPTER II

Related Work: Modeling Emotions

Emotion recognition is a complex, multifaceted field drawing on various research areas. This chapter explores the various methods and considerations in this field, from the use of crowdsourcing to the importance of context, and from handling confounding factors to the impact of noise on machine learning models. We explore the ethical considerations of unintentional encoding of sensisitive variables in data collection and neural networks, the role of interpretability in model trustworthiness, and the importance of automating human in the loop feedback. We also delve into the challenge of generalizability in emotion recognition.

2.1 Concerns with Emotion Recognition Datasets

Some aspects of the above mentioned datasets limit their applicability, including: a lack of naturalness, unbalanced emotion content, unmeasured confounding variables, small size, small number of speakers, and presence of background noise. These datasets are also limited in the number of modalities they use, usually relying on visual and acoustic/lexical information.

2.1.1 Recorded Modalities

As shown in Table 2.1, the most common modalities are video, acoustics, and text. In addition to these modalities, we chose to record two more modalities: thermal and physiological. Previous research has shown that thermal recordings perform well as non-invasive measurement of physiological markers like, cardiac pulse and skin temperature [173, 172, 80]. They have been shown to be correlated to stress symptoms, among other physiological measures. We used the physiological modality to measure stress responses [234, 210] to psychological stressors. This modality has been previously noted in literature for measuring stress [96], usually measured in polygraph tests. We perform baseline experiments to show that the modalities collected in the dataset are indeed informative for identifying stress and emotion.

2.1.2 Lack of Naturalness

A common data collection paradigm for emotion is to ask actors to portray particular emotions. These are usually either short snippets of information [36], a single sentence in a situation [38], or obtained from sitcoms and rehearsed broadcasts [47]. A common problem with this approach is that the resulting emotion display is not natural [113]. These are more exaggerated versions of singular emotion expression rather than the general, and messier, emotion expressions that are common in the real world [12, 21, 72]. Further, expressions in the real world are influenced by both conversation setting and psychological setting. While some datasets have also collected spontaneous data [36, 38], these utterances, though emotionally situated, are often neutral in content when annotated. The usual way to get natural emotional data is to either collect data using specific triggers that have been known to elicit a certain kind of response or to completely rely on in-the wild data, which however often leads to unbalanced emotional content in the dataset [183].

2.1.3 Unbalanced Emotion Content

In-the-wild datasets are becoming more popular [47, 118, 138]. The usual limitation to this methodology is that, firstly, for most people, many conversations are neutral in emotion expression. This leads to a considerable class imbalance [183]. To counter this issue, MSP-Podcast [143] deals with unbalanced content by pre-selecting segments that are more likely to have emotional content. Secondly, data collected in particular settings, e.g., therapy [162], or patients with clinical issues [130] comprise mostly of negative emotions because of the recruitment method used in the collection protocol.

2.1.4 Presence of Interactional Variables

The common way of inducing emotions involves either improvisation prompts or scripted scenarios. Emotion has been shown to vary with a lot of factors that are different from the intended induction [198, 240, 156]. These factors in general can be classified into: (a) recording environment confounders and (b) collection confounders. Recording environment-based variables hamper the models' ability to to learn the emotion accurately. These can be environment noise [16], placement of sensors or just ambient temperature [31].

Table 2.1: Summary of some of the existing emotion corpora. Lexical modality is mentioned for manually transcribed datasets. A - Audio, L - Lexical, T- Thermal, V- Visual, P - Physiological.

	Corpus	Size	Speakers	Rec. Type	Language	Modality	Annotation Type
1.	IEMOCAP	12h26m	10	improv/acted	English	A, V, L	Ordinal, Categorical
2.	MSP-Improv	9h35m	12	improv/acted	English	A, V	Ordinal
3.	VAM	12h	47	spontaneous	German	A, V	Ordinal
4.	SEMAINE	6h21m	20	spontaneous	English	A, V	Ordinal, Categorical
5.	RECOLA	2h50m	46	spontaneous	French	A, V, P	Ordinal
6.	FAU-AIBO	9h12m	51	spontaneous	German	A, L	Categorical
7.	TUM AVIC	0h23m	21	spontaneous	English	A, V, L	Categorical
8.	Emotion Lines	30k samples	-	spont/scripted	English	A, L	Categorical
9.	OMG-Emotion	2.4k samples	-	spontaenous	English	A, V, L	Ordinal
10.	MSP-Podcast	27h42m	151	spontaneous	English	A	Ordinal, Categorical
11.	MuSE	10h	28	spontaneous	English	A, V, L, T, P	Ordinal (Random, Context)

2.1.5 Demographics in Dataset Collection Recruitment

The data collection variations influence both the data generation and data annotation stages. The most common confounders are gender, i.e., ensuring an adequate mix of male vs female, and culture, i.e., having a representative sample to train a more general classifier. Another confounding factor includes personality traits [242], which influence how a person both produces [242] and perceives [158] emotion. Another confounder that can occur at the collection stage is the familiarity between the participants, like RECOLA [183], which led to most of the samples being mainly positive due to the colloquial interaction between the participants. They also do not account for the psychological state of the participant. Psychological factors such as stress [132], anxiety [229] and fatigue [26] have been shown previously to have significant impact on the display of emotion. But the relation between these psychological factors and the performance of models trained to classify emotions in these situations has not been studied.

2.2 Crowdsourcing and Context in Emotion Recognition

Crowdsourcing has emerged as a highly efficient approach for gathering dependable emotion labels, as extensively investigated by Burmania et al. [33]. In addition to this, previous studies have concentrated on enhancing the dependability of annotations by employing quality-control methods. For instance, Soleymani et al. [201] have proposed the utilization of qualification tests to weed out spammers from the crowd, thereby ensuring the quality of collected data. Furthermore, Burmania et al. [35] have explored the use of gold-standard samples to continuously monitor the reliability and fatigue levels of annotators.

The interpretation of emotions is heavily influenced by the context in which they are expressed. Various factors such as tone, choice of words, and facial expressions can significantly impact how individuals perceive and understand emotions [129]. It is noteworthy that this contextual information is implicitly incorporated in the labeling schemes

of commonly used emotion datasets like IEMOCAP [36] and MSP-Improv [38]. However, a notable disparity often exists between the information available to human annotators and that accessible to emotion classification systems. This discrepancy arises because emotion recognition systems are typically trained on individual utterances [10, 3, 157, 190].

2.3 Handling Confounding Factors

2.3.1 Singularly Labeled or Unlabeled Factors

To address confounding factors that are either labeled singularly or cannot be labeled, researchers have devised specific methods. For instance, Ben-David et al. [23] conducted a study wherein they showed that a sentiment classifier, trained to predict the sentiment expressed in reviews, could also implicitly learn to predict the category of the products being reviewed. This finding highlights the potential of classifiers to capture additional information beyond their primary task. In a similar vein, Shinohara [196] employed an adversarial approach to train noise-robust networks for automatic speech recognition. By leveraging this technique, Shinohara aimed to enhance the network's ability to handle noisy and distorted speech signals.

2.3.2 Explicitly Labeled Factors

In addition to addressing confounding factors that are singularly or unlabeled, researchers have also developed methods to handle confounding factors that are explicitly labeled during the data collection process. One such approach involves the use of adversarial multi-task learning, which aims to mitigate variances caused by speaker identity [153]. By incorporating this technique, researchers can reduce the influence of speaker-specific characteristics on the emotion recognition system, thereby enhancing its generalizability. Furthermore, a similar approach has been employed to prevent networks from learning publication source characteristics, which could introduce biases in the classification process [149]

20

2.4 Noise and Approaches to Dealing with it in Machine Learning Models

The impact of noise on machine learning models has been the subject of extensive research, which can be broadly classified into three main directions: robustness in automatic speech recognition, noise-based adversarial example generation, and performance improvement through model augmentation with noise.

One area of focus is the robustness of models in automatic speech recognition (ASR) when exposed to noisy environments. Researchers have explored various techniques to enhance the performance of ASR systems in the presence of noise. This includes the development of noise-robust feature extraction methods, such as mel-frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) features [135]. These techniques aim to minimize the impact of noise on the accuracy of speech recognition systems, enabling them to effectively operate in real-world, noisy conditions.

Another line of research involves the generation of noise-based adversarial examples, which are intentionally crafted to deceive machine learning models. Adversarial attacks exploit vulnerabilities in models by adding imperceptible noise to input samples, causing the models to misclassify or produce incorrect outputs. Carlini and Wagner [42] and Gong et al. [89] have proposed methodologies for generating adversarial audio examples that can fool ASR systems. These techniques highlight the importance of understanding and addressing the susceptibility of machine learning models to adversarial noise.

Furthermore, researchers have explored the potential benefits of incorporating noise during the training and augmentation process of machine learning models. By augmenting the training data with various types of noise, models can become more robust and adaptable to real-world conditions. For instance, Sohn et al. [200] and Wallace et al. [224] have investigated the effectiveness of noise augmentation techniques in improving the performance of models across different tasks. These methods aim to enhance model generalization and

reduce overfitting, ultimately leading to better model performance in noise-affected scenarios.

While evaluating model robustness to noise or adversarial attacks, researchers commonly introduce noise into the dataset and assess the model's performance [5]. However, when it comes to emotion recognition, introducing noise while ensuring that the perception of emotions remains intact can be highly challenging. It is crucial to strike a balance between adding noise for robustness evaluation purposes and preserving the original emotional content. This ensures that the introduced noise does not distort or alter the true emotional expression, enabling accurate and reliable emotion recognition systems.

2.5 Unintentional Sensisitve Variable Encoding, and Ethical Considerations in Data Collection and Neural Networks

The preservation of privacy in data collection has been a key area of focus in early research. Various methods such as rule-based systems and the introduction of background noise have been explored in order to achieve this goal [88, 69]. However, more recent studies have shifted their attention towards privacy preservation in the context of neural networks. In particular, researchers have primarily concentrated on ensuring that the input data used in these networks are not memorized and cannot be retrieved even when the model is deployed [41, 2].

Another crucial consideration in the field of privacy preservation is fair algorithmic representation. The objective here is to develop networks that are invariant to specific attributes, often related to demographic information, in order to ensure fairness [29, 59, 61]. Although certain methods have demonstrated promise in achieving fairness, they may still inadvertently lead to privacy violations [108].

2.6   The Role of Interpretability in Model Trustworthiness

The aspect of interpretability plays a crucial role in establishing trustworthiness of models. Studies have indicated that individuals are more inclined to trust the decisions made by a model if its explanations align with their own decision-making processes [203, 73, 195]. In addition, interpretability methods can be employed by model designers to evaluate and debug a trained model [68]. These methods provide insights into the inner workings of the model and facilitate a better understanding of its decision-making process.

2.7   Automating Human in the Loop Feedback

In order to automate human in the loop feedback, several approaches have been proposed. One such approach involves the utilization of a teacher-student feedback model, where feedback from human teachers is used to improve the performance of the model [179]. Another avenue of research focuses on enhancing active learning techniques, which aim to select the most informative data points for annotation by human experts, thereby reducing the overall labeling effort required [115].

These methods often incorporate a combination of fine-tuning and prompt-based learning techniques, which further enhance the model's ability to learn from human feedback and adapt its performance accordingly [217]. By fine-tuning the model based on the feedback received and utilizing prompts as guiding cues, these approaches enable the model to continually improve its performance, making it more effective in addressing the specific task or problem at hand.

2.8   Generalizability in Emotion Recognition

Achieving generalizability in emotion recognition poses a significant challenge for researchers. To address this challenge, various methods have been explored in order to obtain models that can generalize well across different datasets and scenarios. One approach is the

use of combined and cross-dataset training, where multiple datasets are combined during the training process to create a more comprehensive and diverse training set. This helps the model learn a wider range of emotion patterns and improves its ability to generalize to unseen data [137].

Another technique that has been investigated is transfer learning, which involves leveraging knowledge acquired from pre-trained models on a related task and applying it to the emotion recognition task. By transferring the learned representations and weights from a pre-trained model, the model can benefit from the general knowledge and feature extraction capabilities it has acquired, leading to improved generalizability in emotion recognition [137].

Furthermore, researchers have also explored the concept of generalizability from the perspective of noisy signals. Emotion recognition often deals with noisy data, such as speech with background noise or facial expressions with occlusions. By developing models that are robust to such noise and can effectively extract emotion-related information from imperfect signals, the generalizability of the models can be enhanced [93].

2.9   Conclusion

The field of emotion recognition is complex, with many factors and considerations influencing the development and deployment of effective models. This chapter has explored some of the key areas in this field, highlighting the importance of crowdsourcing, context, handling confounding factors, dealing with noise, and ensuring that the representations don't inadverdently encode sensitive demographic or membership information. The role of interpretability in model trustworthiness and the challenge of automating human in the loop feedback were also discussed. Although progress has been made in many of these areas, the challenge of generalizability in emotion recognition remains, and future research will need to continue to address this issue.

CHAPTER III

Datasets and Pre-processing

This thesis focusses on emotion recognition as a task. For this purpose, we use a standard set of datasets and features as described in this chapter. This allows us to perform experiments with a set of known and commonly used datasets, keeping them uniform across experimental variables.

3.1 Datasets Used In Thesis

In the past years, there have been multiple emotional databases collected and curated to develop better emotion recognition systems. Table 2.1 shows the major corpora that are used for emotion recognition.

3.1.1 IEMOCAP

The IEMOCAP dataset was created to explore the relationship between emotion, gestures, and speech. Pairs of actors, one male and one female (five males and five females in total), were recorded over five sessions. Each session consisted of a pair performing either a series of given scripts or improvisational scenarios. The data were segmented by speaker turn, resulting in a total of 10,039 utterances (5,255 scripted turns and 4,784 improvised turns).

3.1.2 MSP-Improv

The MSP-Improv dataset was collected to capture naturalistic emotions from improvised scenarios. It partially controlled for lexical content by including target sentences with fixed lexical content that are embedded in different emotional scenarios. The data were divided into 652 target sentences, 4,381 improvised turns (the remainder of the improvised scenario, excluding the target sentence), 2,785 natural interactions (interactions between the actors in between recordings of the scenarios), and 620 read sentences for a total of 8,438 utterances.

3.1.3 MSP-Podcast

The MSP-Podcast dataset was collected to build a naturlisitic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. This was done using machine learning algorithms, which along with a cost-effective annotation process using crowdsourcing, led to a vast and balanced dataset. We use a pre-split part of the dataset which has been identified for gender of the speakers which comprises of 13,555 utterances. The dataset as a whole contains audio recordings.

3.1.4 MuSE

The MuSE dataset consists of recordings of 28 University of Michigan college students, 9 female and 19 male, in two sessions: one in which they were exposed to an external stressor (final exams period at University of Michigan) and one during which the stressor was removed (after finals have concluded). Each recording is roughly 45-minutes. We expose each subject to a series of emotional stimuli, short-videos and emotionally evocative monologue questions. These stimuli are different across each session to avoid the effect of repetition, but capture the same emotion dimensions. At the start of each session, we record a short segment of the user in their natural stance without any stimuli, to establish a baseline. We record their behavior using four main recording modalities: 1) video camera, both close-up on the face and wide-angle to capture the upper body, 2) thermal camera,

close-up on the face, 3) lapel microphone, 4) physiological measurements, in which we choose to measure heart rate, breathing rate, skin conductance and skin temperature (Figure 4.1). The data include self-report annotations for emotion and stress (Perceived Stress Scale, PSS) [56, 57], as well as emotion annotations obtained from Amazon Mechanical Turk (AMT). To understand the influence of personality on the interaction of stress and emotion, we obtain Big-5 personality scores [87], which was filled by 18 of the participants, due to the participation being voluntary.

3.2 Data Pre-Processing

We use these features consistently across the thesis to have a standardized set of inputs, aiming to avoid variability that comes from different labelling or pre-processing schemas. Our preprocessing corresponds to converting Likert scale emotion annotations to classes based on quartiles. The feature processing has 2 components, acoustic and lexical, for training, testing or fine-tuning speech-only, text-only or bimodal models.

3.2.1 Emotion Labels

Each utterance in the MuSE dataset was labeled for *activation* and *valence* on a nine-point Likert scale by eight crowd-sourced annotators [105], who observed the data in random order across subjects. We average the annotations to obtain a mean score for each utterance, and then bin the mean score into one of three classes, defined as, {"*low*": [min, 4.5], "*mid*": (4.5, 5.5], "*high*": (5.5, max]}. The resulting distribution for activation is: {"*high*": 24.58%, "*mid*": 40.97% and "*low*": 34.45%} and for valence is {"*high*": 29.16%, "*mid*": 40.44% and "*low*": 30.40%}. Utterances in IEMOCAP and MSP-Improv were annotated for valence and activation on a five-point Likert scale. The annotated activation and valence values were averaged for an utterance and binned as: {"*low*": [1, 2.75], "*mid*": (2.75, 3.25], "*high*": (3.25, max]}

3.2.2 Stress Labels

Utterances in the the MuSE dataset include stress annotations, in addition to the activation and valence annotations. The stress annotations for each session were self-reported by the participants using the Perceived Stress Scale (PSS) [58]. We perform a paired t-test for subject wise PSS scores, and find that the scores are significantly different for both sets (16.11 vs 18.53) at $p < 0.05$. This especially true for question three (3.15 vs 3.72), and hence, we double the weightage of the score for this question while obtaining the final sum. We bin the original nine-point adjusted stress scores into three classes, {"*low*": (min, mean−2], "*mid*": (mean−2, mean+2], "*high*": (mean+2, max]}. We assign the same stress label to all utterances from the same session. The distribution of our data for stress is "*high*": 40.33%, "*mid*": 25.78% and "*low*": 38.89%

Improvisation Labels. Utterances in the IEMOCAP dataset were recorded in either a scripted scenario or an improvised one. We label each utterance with a binary value {"*scripted*", "*improvised*"} to reflect this information.

3.3 Lexical and Acoustic Feature Extraction

3.3.1 Acoustic

We use Mel Filterbank (MFB) features, which are frequently used in speech processing applications, including speech recognition, and emotion recognition [116, 126]. We extract the 40-dimensional MFB features using a 25-millisecond Hamming window with a step-size of 10-milliseconds. As a result, each utterance is represented as a sequence of 40-dimensional feature vectors. We *z*-normalize the acoustic features by session for each speaker.

3.3.2 Lexical

We have human transcribed data available for MuSE and IEMOCAP. We use the word2vec representation based on these transcriptions, which has shown success in sentiment and

emotion analysis tasks [121]. We represent each word in the text input as a 300-dimensional vector using a pre-trained word2vec model [155], replacing out-of-vocab words with the $\langle unk \rangle$ token. Besides, we also incorporate BERT embeddings for enhanced contextual understanding. These embeddings, generated from the pre-trained BERT model, provide deep, bidirectional representations by understanding the text context from both directions. Each utterance is eventually represented as a sequence of 768-dimensional feature vectors. We use just acoustic inputs for MSP-Improv because human transcriptions are not available.

Emotion Recognition Dataset: MuSE

4.1 Motivation and Contributions

Endowing automated agents with the ability to provide support, entertainment and interaction with human beings requires sensing of the users' affective state. These affective states are impacted by a combination of emotion inducers, current psychological state, and various contextual factors. Although emotion classification in both singular and dyadic settings is an established area, the effects of these additional factors on the production and perception of emotion is understudied. This chapter presents a dataset, Multimodal Stressed Emotion (MuSE), to study the multimodal interplay between the presence of stress and expressions of affect. We describe the data collection protocol, the possible areas of use, and the annotations for the emotional content of the recordings. The chapter also presents several baselines to measure the performance of multimodal features for emotion and stress classification.

4.2 Introduction

Virtual agents have become more integrated into our daily lives than ever before [144]. For example, Woebot is a chatbot developed to provide cognitive behavioral therapy to a user [74]. For this chatbot agent to be effective, it needs to respond differently when the user

is stressed and upset versus when the user is calm and upset, which is a common strategy in counselor training [213]. While virtual agents have made successful strides in understanding the task-based intent of the user, social human-computer interaction can still benefit from further research [54]. Successful integration of virtual agents into real-life social interaction requires machines to be emotionally intelligent [27, 238].

But humans are complex in nature, and emotion is not expressed in isolation [90]. Instead, it is affected by various external factors. These external factors lead to interleaved user states, which are a culmination of situational behavior, experienced emotions, psychological or physiological state, and personality traits. One of the external factors that affects psychological state is stress. Stress can affect everyday behavior and emotion, and in severe states, is associated with delusions, depression and anxiety due to impact on emotion regulation mechanisms [122, 193, 216, 225]. Virtual agents can respond in accordance to users' emotions only if the machine learning systems can recognize these complex user states and correctly perceive users' emotional intent. We introduce a dataset designed to elicit spontaneous emotional responses in the presence or absence of stress to observe and sample complex user states.

There has been a rich history of visual [235, 111], speech [143], linguistic [207], and multimodal emotion datasets [38, 36, 183]. Vision datasets have focused both on facial movements [111] and body movement [131]. Speech datasets have been recorded to capture both stress and emotion separately but do not account for their inter-dependence [185, 97, 127, 243]. Stress datasets often include physiological data [234, 210].

Existing datasets are limited because they are designed to elicit emotional behavior, while neither monitoring external psychological state factors nor minimizing their impact by relying on randomization. However, emotions produced by humans in the real world are complex. Further, our natural expressions are often influenced by multiple factors (e.g., happiness *and* stress) and do not occur in isolation, as typically assumed under laboratory conditions. The primary goal of this work is to collect a multimodal stress+emotion dataset
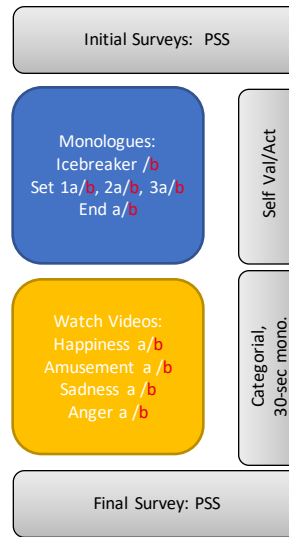
Figure 4.1: Experimental Protocol For Recording

– Multimodal Stressed Emotion (MuSE) – to promote the design of algorithms that can recognize complex user states.

For MuSE, The extracted features for each modality, and the anonymized dataset (other than video) will be released publicly along with all the corresponding data and labels. We present baseline results for recognizing both emotion and stress in the chapter, in order to validate that the presence of these variables can be computationally extracted from the dataset, hence enabling further research.

4.3   MuSE Dataset

4.3.1   Experimental Protocol

We collect a dataset that we refer to as Multimodal Stressed Emotion (MuSE) to facilitate the learning of the interplay between stress and emotion. The protocol for data collection is shown in Figure 4.1. There were two sections in each recording: monologues and watching emotionally evocative videos. We measure the stress level at the beginning and end of each recording. The monologue questions and videos were specifically chosen to cover all

categories of emotions. At the start of each recording, we also recorded a short one-minute clip without any additional stimuli to register the baseline state of the subject.

Previous research has elicited situational stress such as public speaking [123, 85, 8], mental arithmetic tasks [139] or use Stroop Word Test [215]. However, these types of stress are often momentary and fade rapidly in two minutes [139]. We alleviate this concern by recording both during and after final exams (we anticipate that these periods of time are associated with high stress and low stress, respectively) in April 2018. We measure stress using Perceived Stress Scale [57] for each participant. We measure their self-perception of the emotion using Self-Assessment Manikins (SAM) [30]. The recordings and the survey measures were coordinated using Qualtrics[1] enabling us to ensure minimal intervention and limit the effect of the presence of another person on the emotion production.

Each monologue section comprised of five questions broken into sections meant to elicit a particular emotion (Table 4.1). These questions were shown to elicit thoughtful and emotional responses in their data pool to generate interpersonal closeness [11]. We include an icebreaker and ending question to ensure cool off periods between change in recording section, i.e., from neutral to monologues, and from monologues to videos, hence decreasing the amount of carry-over emotion from the previous monologue to the next. Each subject was presented with a different set of questions over the two recordings to avoid repetition effect. We also shuffle the order of the other three questions to account for order effects [133]. Each subject was asked to speak for a minimum of two minutes. After their response to each question, the subjects marked themselves on two emotion dimensions: activation and valence on a Likert Scale of one to nine using self-assessment manikins [30].

For the second part of the recording, the subjects were asked to watch videos in each of the four quadrants i.e., the combination of {*low, high*} × {*activation, valence*} of emotion. These clips were selected from the corpus [140, 20], which tested for the emotion elicited from the people when watching these clips (Table 4.2). The subjects were monitored for

---

[1]umich.qualtrics.com

33

their reaction to the clips. After viewing a clip, subjects are asked to speak for thirty seconds about how the video made them feel. After their response, they marked a emotion category, e.g., angry, sad, etc. for the same clip. When switching videos, the subjects were asked to view a one-minute neutral clip to set their physiological and thermal measures back to the baseline [189].

The 28 participants were also asked to fill out an online survey used for personality measures on the big-five scale [87], participation being voluntary. This scale has been validated to measure five different dimensions named OCEAN (openness, conscientiousness, extraversion, agreeableness, and neuroticism) using fifty questions and has been found to correlate with passion [60], ambition [19], and emotion mechanisms [181]. We received responses for this survey from 18 participants. These labels can be used in further work to evaluate how these personality measures interact with the affects of stress in emotion production, as previously studied in [242].

4.3.2 Equipment Setup

The modalities considered in our setup are: thermal recordings of the subject's face, audio recordings of the subject, color video recording of the subject's face, a wide-angle color video recording the subject from the waist up and physiological sensors measuring skin conductance, breathing rate, heart rate and skin temperature. For these modalities we have set up the following equipment:

1. **FLIR Thermovision A40 thermal camera** for recording the close-up thermal recording of the subject's face. This camera provides a 640x512 image in the thermal infrared spectrum.

2. **Raspberry Pi with camera module V2 with wide-angle lens** used for the waist up shot of the subject. We have chosen Raspberry Pi's due to its low price and support for Linux OS, which integrates easily into a generic setup.

3. **Raspberry Pi with camera module V2** used to record the subject from the waist up.

Table 4.1: Emotion elicitation questions.

Icebreaker

1. Given the choice of anyone in the world, whom would you want as a dinner guest?
2. Would you like to be famous? In what way?

Positive

1. For what in your life do you feel most grateful?
2. What is the greatest accomplishment of your life?

Negative

1. If you could change anything about the way you were raised, what would it be?
2. Share an embarrassing moment in your life.

Intensity

1. If you were to die this evening with no opportunity to communicate with anyone, what would you most regret not having told someone?
2. Your house, containing everything you own, catches fire. After saving your loved ones and pets, you have time to safely make a final dash to save any one item. What would it be? Why?

Ending

1. If you were able to live to the age of 90 and retain either the mind or body of a 30-year old for the last 60 years of your life, which would you choose?
2. If you could wake up tomorrow having gained one quality or ability, what would it be?

4. **TASCAM DR-100 mk II** used to record audio. We chose this product for its high fidelity. It can record 24-bit audio at 48kHz.

5. **ProComp$^\infty$-8 channel biofeedback and neurofeedback system v6.0** used to measure blood volume pulse (BVP sensor), skin conductance (SC sensor), skin temperature (T sensor), and abdominal respiration (BR sensor)

The equipment operator started and marked the synchronization point between video and audio recordings using a clapper. Subsequent time stamps are recorded by the qualtrics

Table 4.2: Emotion elicitation clips.

Movie	Description
Low Valence, Low Activation (Sad)	
City of Angels	Maggie dies in Seth's arms
Dangerous Minds	Students find that one of their classmates has died
Low Valence, High Activation (Anger)	
Sleepers	Sexual abuse of children
Schindler's List:	Killing of Jews during WWII
High Valence, Low Activation (Contentment)	
Wall-E	Two robots dance and fall in love
Love Actually	Surprise orchestra at the wedding
High Valence, High Activation (Amusement)	
Benny and Joone	Actor plays the fool in a coffee shop
Something About Mary	Ben Stiller fights with a dog
Neutral	
A display of zig-zag lines across the screen	
Screen-saver pattern of changing colors	

survey using subject click timings.

4.3.3 Post-processing

Splitting of the Recordings. Each modality is split into neutral recordings of one-minute, five questions and four video recordings with associated monologues, resulting in fourteen recordings for emotional content, thus 28 recordings per subject. In total we have 784 distinct recordings over five modalities, 28 subjects and two stress states, for a total of 3920 recording events. Temperatures are clamped to between 0^oC and 50^oC. This helps reduce the size of the thermal recording files after being zipped.

Utterance Construction. The five monologues extracted above were divided into utterances. However, since the monologues are a form of spontaneous speech, there are no clear sentence boundaries marking end of utterance. We manually created utterances by identifying prosodic or linguistic boundaries in spontaneous speech as defined by [125].
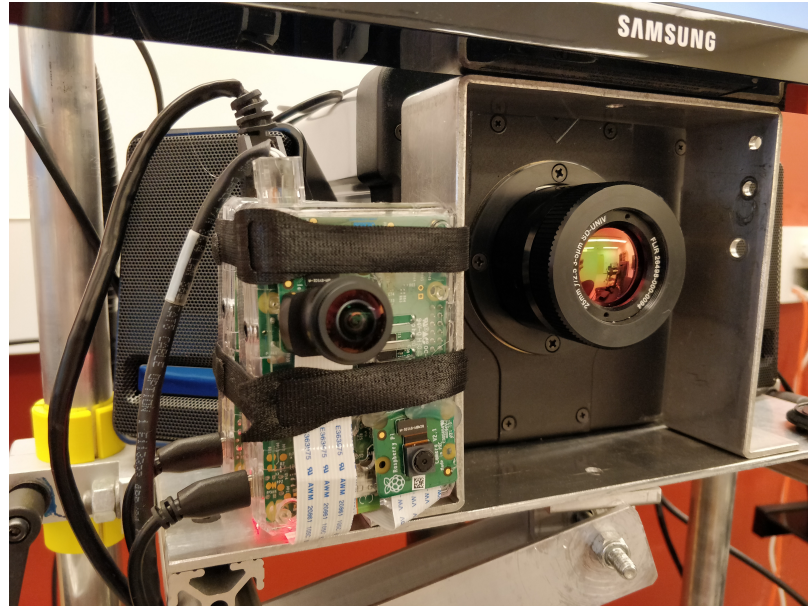
Figure 4.2: Close-up view of the thermal and video recording equipment.

The boundaries used for this work are: (a) clear ending like a full stop or exclamation, (b) a change in context after filler words or completely revising the sentence to change meaning, or (c) a very long pause in thought. This method has been previously shown to be effective in creating utterances that mostly maintain a single level of emotion [118].

The dataset contains 2,648 utterances with a mean duration of 12.44 ± 6.72 seconds (Table 4.3). The mean length of stressed utterances (11.73 ± 5.77 seconds) is significantly different (using two-sample t-test) from that of the non-stressed utterances (13.30 ± 6.73 seconds). We remove utterances that are shorter than 3-seconds and longer than 35-seconds and end up retaining 97.2% of our dataset. This allows us to to avoid short segments that may not have enough information to capture emotion, and longer segments that can have variable emotion, as mentioned in [118]. Because our dataset is comprised of spontaneous utterances, the mean length of utterance is larger than those in a scripted dataset [38] due to more corrections and speech overflow.

Stress State Verification. We perform a paired t-test for subject wise PSS scores, and find that the mean scores are significantly different for both sets (16.11 vs 18.53) at $p < 0.05$. This implied that our hypothesis of exams eliciting persistently more stress than

normal is often true. In our dataset, we also provide levels of stress which are binned into three categories based on weighted average (using questions for which the t-test score was significant).

4.4 Emotional Annotation

4.4.1 Crowdsourcing

Crowdsourcing has previously been shown to be an effective and inexpensive method for obtaining multiple annotations per segment [99, 34]. We posted our experiments as Human Intelligence Tasks (HITs) on *Amazon Mechanical Turk* and used selection and training mechanisms to ensure quality [106]. HITs were defined as sets of utterances in a monologue. The workers were presented with a single utterance and were asked to annotate the activation and valence values of that utterance using Self-Assessment Manikins [30]. Unlike the strategy adopted in [47], the workers could not go back and revise the previous estimate of the emotion. We did this to ensure similarity to how a human listening into the conversation might shift their perception of emotion in real time. These HITs were presented in either the contextual or the random presentation condition defined below.

In the contextual experiment, we posted each HIT as a collection of ordered utterances from each section of a subject's recording. Because each section's question was designed to elicit an emotion, to randomize the carry-over effect in perception, we posted the HITs in a random order over the sections from all the subjects in our recording. For example, a worker might see the first HIT as *Utterance 1...N from Section 3 of Subject 4's stressed recording* and see the second HIT as *Utterance 1...M from Section 5 of Subject 10's non-stressed recording* where *N, M* are the number of utterances in those sections respectively. This ensures that the annotator adapts to the topic and fluctuations in speaking patterns over the monologue being annotated.

In the randomized presentation, each HIT is an utterance from any section, by any speaker,

Table 4.3: Data summary (R:random, C:context, F:female, M:male).

Monologue Subset	
Mean no. of utterances/monologue	9.69 ± 2.55
Mean duration of utterances	12.44 ± 6.72 seconds
Total no. of utterances	2,648
Selected no. of utterances	2,574
Gender distribution	19 (M) and 9 (F)
Total annotated speech duration	~ 10 hours
Crowdsourced Data	
Num of workers	160 (R) and 72 (C)
Blocked workers	8
Mean activation	3.62±0.91 (R) 3.69±0.81 (C)
Mean valence	5.26±0.95 (R) 5.37±1.00 (C)

in random order. So, a worker might see the first HIT as *Utterance 11 from Section 2 of Subject 1's stressed recording monologue* and see the second HIT as *Utterance 1 from Section 5 of Subject 10's non-stressed monologue recording*. We use this method of randomization to ensure lack of adaptation to both speaker specific style and the contextual information. The per-utterance and the contextual labels can be used to train different machine learning models that are apt for either singular one-off instances or for holding multiple turn natural conversation, respectively.

4.4.2 Emotion Content Analysis

We show the distribution of the annotations received in both the random and contextual setting in Table 4.3 and Figure 4.3. The labels obtained for our dataset form a distribution that mostly covers negative and neutral levels of activation, and all but extremities for valence. This can also be seen in the data summary in Table 4.3. We performed a paired t-test between the labels obtained from random vs contextual presentation and found that these labels are significantly different (using paired t-test at $p < 0.05$ for both activation and valence for utterances in the non-stressed situation). Although the obtained labels are significantly
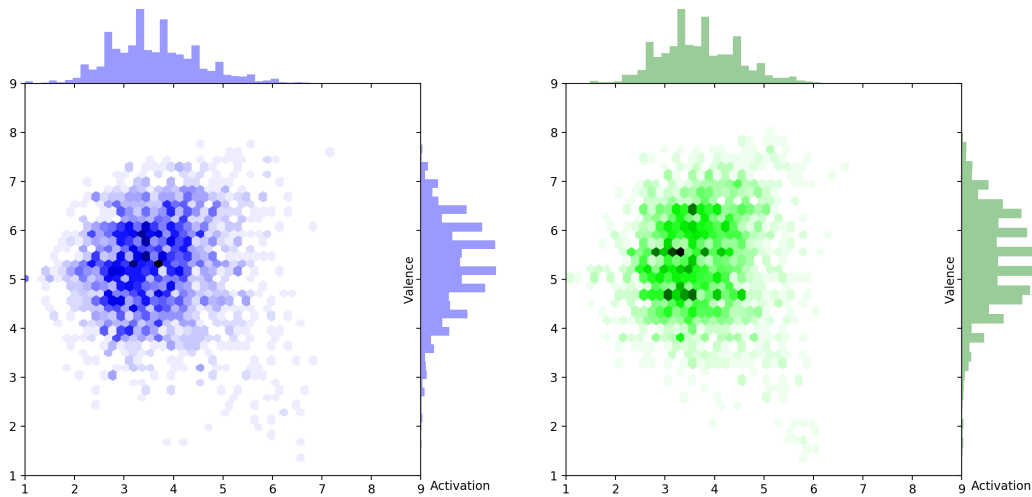
Figure 4.3: Distribution of the activation and valence ratings in random labeling scheme (on left) and contextual labeling scheme (on right).

different for valence in the stressed category using the same method as above, the same does not hold true for the activation annotations in this category.

4.5 Experiments

In this section, we describe our baseline experiments for predicting emotion and stress in the recorded modalities. We have a more granular marked annotation of emotion, i.e., over each utterance, as compared to stress over the complete monologue. Hence, we extract features for each modality over continuous one second frame intervals for predicting stress, and over the complete utterance for emotion. Audio and lexical features are still extracted over a complete utterance for stress due to higher interval of variation over time.

4.5.1 Evaluation of Emotion Recognition

We use the following set of features for our baseline models:

1. **Acoustic Features.** We extract acoustic features using OpenSmile [71] with the eGeMAPS configuration [70]. The eGeMAPS feature set consists of 88 utterance-

Labelling emotion of a short (~25s) audio clip

Instructions (Click to expand)

In this task, you will listen to an audio clip which is a few seconds long and then answer two multiple choice questions.

Please listen to the entire audio after accepting the HIT, before answering the questions.

For the following two questions, you will each select an option (appear as an image) that is the **closest to your perception of the speaker's emotion**. Each question will ask about a different aspect.

- Q1: How **negative vs. positive** is the main speaker?
  - Example negative emotions: angry, sad, tired, annoyed, frustrated
  - Example positive emotions: happy, content, pleasant, relaxed
- Q2: How **calm vs. excited** is the main speaker?
  - Note: excitement does not imply positiveness.
  - Example calm emotions: sleepy, bored, calm
  - Example excited emotions: rage, glee, excited, anxious

Please listen to the **entire audio after accepting the HIT**, before answering the questions.

▶ 0:00 / 0:12 ● ——— ◀) ⋮

Figure 4.4: An overview of the instructions provided to the annotators for annotating an utterance.

level statistics over the low-level descriptors of frequency, energy, spectral, and cepstral parameters. We perform speaker-level z-normalization on all features.

2. **Lexical Features.** We extract lexical features using Linguistic Inquiry and Word Count (LIWC) [174]. These features have been shown to be indicative of stress, emotion, veracity and satisfaction [86, 161, 164]. We normalize all the frequency counts by the total number of words in the sentence accounting for the variations due to utterance length.

3. **Thermal Features.** For each subject a set of four regions were selected in the thermal image: the forehead area, the eyes, the nose and the upper lip as previously used in [172, 80, 6]. These regions were tracked for the whole recording and a 150-bin histogram of temperatures was extracted from the four regions per frame, i.e., 30 frames a second for thermal recordings. We further reduced the histograms to the first four measures of central tendency, e.g. Mean, Standard Deviation, Skewness and Kurtosis. We combined these features over the utterance using first delta measures (min, max, mean, SD) of all the sixteen extracted measures per frame, resulting in 48 measures in total.
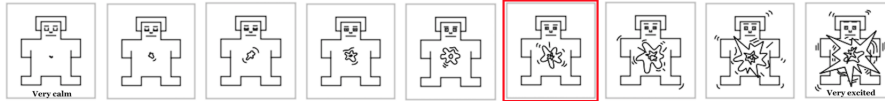
Figure 4.5: Annotation scale used by MTurk workers to annotate the emotional content of the corpus. They annotate valence and activation for each utterance.

4. **Close-up Video Features.** We use OpenFace [15] to extract the subject's facial action units. The AUs used in OpenFace for this purpose are AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28 and AU25 comprising of eyebrows, eyes and mouth. These features have been previously shown to be indicative of emotion [227, 64] and have been shown to be useful for predicting deception [110]. We summarize all frames into a feature using summary statistics (maximum, minimum, mean, variance, quantiles) across the frames and across delta between the frames resulting in a total of 144 dimensions.

Network Setup. We train and evaluate multiple unimodal Deep Neural Networks (DNN) models for predicting valence and activation using Keras [91]. [106] have shown that a match between the context provided to the classifier and the annotator leads to better classification performance. Because we are performing single utterance classification, for all further experiments, we use the annotations obtained in a random manner as mentioned above. In all cases, we predict the continuous annotation using regression.

We also use an ensemble of these four networks (audio, lexical, visual and thermal) to measure multimodal performance. For each network setup, we follow a five-fold subject

Table 4.4: RMSE for emotion classification models using multiple modalities. Significance established at $p < 0.05$.

	Activation	Valence
Unimodal Models		
Acoustic (A)	**1.004***	1.122
Lexical (L)	1.343	0.980
Close Video (V)	1.111	**0.879****
Thermal (T)	2.012	1.565
Ensemble		
A+L	0.987	0.981
A+V	0.970	0.899
L+V	0.981	0.901
A+L+V	0.972	**0.856***
A+L+V+T (All)	**0.961***	0.868

independent evaluation scheme and report the average RMSE across the folds. For each test-fold, we use the previous fold for hyper-parameter selection and early stopping. The hyper-parameters include: number of layers $\{2, 3, 4\}$ and layer width $\{64, 128, 256\}$. We use ReLU activation and train the networks with MSE loss using the Adam optimizer.

We train our networks for a maximum of 50 epochs and monitor the validation loss after each epoch. We perform early stopping if the loss doesn't decrease for 15 consecutive epochs. We save the weights that achieved the lowest validation performance during training. We train each network five times with different seeds and average the predictions to account for variations due to random initialization.

Results. We show our results in Table 4.4. We find that between acoustic and lexical modalities, the acoustic modality carries more information about activation and the lexical for valence. This is in line with previous research [232, 39]. We also note that the visual modality significantly outperforms both the speech and lexical modalities for valence prediction.

When we merge these networks using late voting on each modality (decision fusion), we find that the combination of all modalities performs the best for predicting activation. But

for predicting valence, the best performance is shown by the combination of acoustic, lexical, visual and thermal modalities. We believe this is true because previous work has shown that thermal features are mostly indicative of intensity and discomfort [94] and hence improves performance on activation prediction, while the visual expressions are most informative about valence [186].

4.5.2 Evaluation of Presence of Stress

We use the following set of features for our baseline models. Given that stress vs non-stressed state is classified for the complete section (monologue or neutral recording), we extract visual features differently to use the the sequential information over the whole segment, i.e., a monologue. We also use physiological features for our network, since we found that even though they are highly variable over shorter segments (utterances), they are informative for recognizing physiological state on a whole section.

1. **Acoustic, Lexical, and Thermal Features.** We use the same features as extracted for predicting emotion.

2. **Wide-angle Video Features.** We extract the subject's pose using OpenPose [40, 199, 228] at 25 frames per second. For each frame, we extract 14 three-dimensional points representing anchor points for the upper body. For classification of each 3D point is interpolated over one second using a 5th order spline [167, 102]. The parameters of the splines are then used as features for classification.

3. **Close-up Video Features.** We use OpenFace to extract the subject's action units [15]. The features are extracted for every frame. In each frame, features include the gaze direction vectors, gaze angles, 2D eye region landmarks, head locations, rotation angles of the head, landmark locations, and facial action units. Landmarks locations offset by the nose location. We window the data into segments of one-second windows with 0.5 second overlap and calculate summary statistics (maximum, minimum, mean,

Table 4.5: Baseline results for classifying stressed and non-stressed situations per time unit, unless specified otherwise. A - Accuracy, P - Precision, R - Recall.

Recording Parts	A	P	R	F_1
Thermal				
Neutral	**0.61**	**0.67**	**0.62**	**0.64**
Questions	0.50	0.64	0.52	0.57
Wide-angle Video				
Neutral	0.66	0.41	0.96	0.58
Questions	0.69	0.45	0.82	0.58
Close-up Video				
Neutral	0.61	0.78	0.33	0.46
Questions	0.65	0.65	0.69	0.67
Physiological				
Neutral	0.66	0.47	0.89	0.64
Questions	0.70	0.55	0.88	0.67
Audio - Per utterance				
Questions	0.67	0.70	0.69	0.69
Text - Per utterance				
Questions	0.60	0.74	0.61	0.67
Late Fusion - Voting				
Questions	0.60	0.74	0.61	0.67

variance). We retain the top 300 features based on the F values between the training features and corresponding labels (stressed vs non-stressed).

4. **Physiological Features.** While the physiological features varied greatly per second to be informative for emotion, they are informative for recognizing presence or absence of stress. We consider the raw measurements for heart rate, breathing rate, skin conductance and skin temperature and compute the first four measures of central tendency, e.g. mean, standard deviation, skewness, and kurtosis.

Network. We train a DNN to perform binary classification, i.e., to recognize stressed vs. non-stressed situation using ReLU as activation, with softmax as the classification method.The final layer uses a soft-max activation. We train six different networks for thermal, wide-angle video, close-up video, physiological, audio, and lexical modalities. Each network is trained in a subject-independent manner. We train network to recognize

stress vs non-stress situation in both neutral recording,i.e., when the subject isn't speaking at the beginning of the recording, and during emotional monologue questions. To do so, we decide the final prediction by a majority vote over one-second predictions for the complete section of the recording. For the lexical and acoustic modality, we train the network for the question monologues, and decide the final prediction based on a majority vote over prediction for each utterance.

Results. We report our results for prediction of stress vs non-stress situation using various modalities in Table 4.5. We see that the captured modalities are indeed informative for recognizing stress vs non-stressed situations. We find that for recognizing this distinction when the subjects are speaking, audio and physiological features perform the best. This is in agreement with previous related work [131, 234, 96]. Interestingly, we also find that the thermal and physiological modality is apt at recognizing differences in stress, even in the neutral recording, i.e., when the subject is not speaking. This advantage of thermal modality has been previously documented by researchers [7, 173, 172, 80]. We find that answering emotional monologue questions interferes with the recorded thermal modality, leading to a poorer performance at stress recognition.

4.6   Conclusions and Future Work

In this chapter, we introduced a dataset that aims to capture the interplay between psychological factors such as stress and emotion. While various other datasets have explored the relationship between gender or personality measures and emotion production and perception, the relationship between psychological factors and emotion is understudied from a data collection point of view, and hence an automated modeling perspective.

We verified that the presence of emotion and stress can be detected in our dataset. Our baseline results for emotion classification using DNNs with acoustic, linguistic and visual features on our dataset are similar to reported results on other datasets such as IEMOCAP [36] and MSP-Improv [38].

CHAPTER V

Best Practices for Noise Based Augmentation in Emotion Datasets

5.1 Motivation and Contributions

Speech emotion recognition is an important component of any human centered system. But speech characteristics produced and perceived by a person can be influenced by a multitude of reasons, both desirable such as emotion, and undesirable such as noise. To train robust emotion recognition models, we need a large, yet realistic data distribution, but emotion datasets are often small and hence are augmented with noise. Often noise augmentation makes one important assumption, that the prediction label should remain the same in presence or absence of noise, which is true for automatic speech recognition but not necessarily true for perception based tasks. In this chapter we make three novel contributions. We validate through crowdsourcing that the presence of noise does change the annotation label and hence may alter the original ground truth label. We then show how disregarding this knowledge and assuming consistency in ground truth labels propagates to downstream evaluation of ML models, both for performance evaluation and robustness testing. We end the chapter with a set of recommendations for noise augmentations in speech emotion recognition datasets.

5.2 Introduction

Speech emotion recognition is increasingly included as a component in many real-world human-centered machine learning models. Modulations in speech can be produced for a multitude of reasons, both desirable and undesirable. In our case desirable modulations encode information that we want our model to learn and be informed by, such as speaker characteristics or emotion. Undesirable modulations encode information that are extrinsic factors change with the environment, such as noise. In order to handle these modulations, we need large datasets that capture the range of possible speech variations and their relationship to emotion expression. But, such datasets are generally not available for emotion tasks. To bridge this gap, researchers have proposed various methods to generate larger datasets. One of the most common is noise augmentation. The baseline assumption of noise augmentation is that the labels of the emotion examples do not change once noise has been added [169]. While this assumption can be confidently made for tasks such as automatic speech recognition (ASR), the same cannot be said for perception-based tasks, such as emotion recognition.

In this chapter, we question the assumption that the annotation label remains the same in the presence of noise. We first create a noise augmented dataset and conduct a perception study to label the emotion of these augmented samples, focused on the type of noise in samples whose perception has changed or remained the same given the agumentation. We use the results from this study to classify the complete set of augmentation noises into two categories, *perception-altering* (i.e., noises that may change the perception of emotion) and *perception-retaining* (i.e., noises that do not change the perception of emotion). We propose that the perception-altering noises should not be used in supervised learning or evaluation frameworks because we cannot confidently maintain that the original annotation holds for a given sample. We evaluate the effects of disregarding emotion perception changes by examining how the performance of emotion recognition models and analyses of their robustness change in unpredictable manners when we include samples that alter human perception in the training of these models. Lastly, we provide a set of recommendations for

noise based augmentation of speech emotion recognition datasets based on our results.

Researchers have considered the impact of noise on emotion perception and thereby the annotation of emotions. [X] looked at how pink and white noises in varying intensities change the perception of emotion. Another set of research has concentrated on training and validating noise robust models with the assumption that intent label prediction remains consistent in the presence of noise. For example, [X] have looked at training student teacher models that aim to ignore the effect of noise introduced to the model. On the other hand [X] have proposed copy pasting various emotion segments together along with neutral noise to balance the classes in an emotion dataset, thus improving performance.

In this chapter, we claim that the standard assumption about perception and hence, label retention of emotion in the presence of noise may not hold true in a multiple noise categories. To understand which noises impact emotion perception, we use a common emotion dataset, IEMOCAP and introduce various kinds of noises to it, at varying signal to noise ratio (SNR) levels as well as at different positions in the sample. We then perform a crowdsourcing experiment that asks workers to annotate their perception of emotion for both the clean and the corresponding noise-augmented sample. This enables us to divide noise augmentation options into groups characterized by their potential to either influence or not influence human perception.

The results of the crowdsourcing experiments inform a series of empircal analyses focused on model performance and model robustness. We first present an empirical evaluation of the effects of including perception-altering noises in training. It will allow us to observe how the inclusion of perception-altering noises creates an impression of performance improvement. We will discuss how this improvement is a myth, this new model will have learned to predict labels that are not truly associated with a given sample due to the perceptual effects of these noises. We consider both a general recurrent neural network (RNN) model and an end-to-end model for this purpose. We evaluate conditions in which novel augmentation noises are either introduced during training (matched) or seen for the first time during testing

(mismatched). The second empirical evaluation analyzes whether the gap in performance between the matched and mismatched conditions can be bridged using noise robust modeling techniques. The third and final evaluation is focused on the robustness of the model. It will allow us to observe how the inclusion of these perception altering noises ultimately leads to a model that is more susceptible to attack compared to a model that does not include these noises. We train an attack model for robustness testing. It considers a pool of noises and picks the best noise with a minimal SNR degradation that is able to change a model's prediction. We consider a condition in which the attack model has black-box access to the trained model. The attack has a fixed number of allowed queries to the trained model, but not the internal gradients or structure (i.e., the attack model can only provide input and can only access the trained model's prediction). We test and monitor the difference in the observed robustness of these aforementioned models.

We find that the crowdsourced labels do change in the presence of some kinds of noise. We then verify that the models perform worse on noisy samples when trained only on clean datasets. But, we show that this decrease in performance is different when using the complete set of noises for augmenting the test set vs. when only using the perception-retaining noises for augmentation. We show similar patterns for noise-robust models, specifically showing how there is an increased drop in performance for the end-to-end noise-robust model when excluding performance-altering noises during augmentation. We then discuss how our conventional metrics, those that look only at model performance, may be incorrectly asserting improvements as the model is learning to predict an emotion measure that is not in line with human perception. Troublingly, we find that the attack model is generally more effective when it has access to the set of all noises as compared to when excluding perception-altering noises for allowed augmentations. We also specifically find that given just a pool of carefully crafted reverberation modulations, the attack model can be successful in almost 65% of the cases with minimal degradation in SNR and in less than ten queries to the trained model. We end the chapter with a general set of recommendations for noise

augmentations in speech emotion recognition datasets.

5.3 Research Questions

In this chapter, we investigate five research questions:

Purpose 1: Premise Validation through Crowdsourcing

RQ1: Does the presence of noise affect emotion perception as evaluated by *human raters*? Is this effect dependent on the utterance length, loudness, the type of the added noise, and the original emotion?

Reason: Noise has been known to have masking effect on humans in specific situations. Hence, humans can often understand verbalized content even in presence of noise. Our goal is to understand whether the same masking effect extends to paralinguistic cues such as emotion, and to what extent. Our continuing claim from hereon remains that only noises that do not change human perception should be used for the training and evaluation of machine learning models. Not doing so, can lead to gains or drops in performance measurement that may not actually extend to real world settings. We call these changes "unverified" because we cannot, with certainity, be sure that the model should have predicted the original label (i.e., the label of the sample before noise was added) because the human did not neccessarily label the noisy instance with that same label.

Purpose 2: Noise Impact Quantification

RQ2: Can we verify previous findings that the presence of noise affects the performance of *emotion recognition models*? Does this effect vary based on the type of the added noise?

Reason: We have known that presence of noise in data shifts the data distribution [50]. This shift often leads to poor performance by machine learning models. We aim to quantify the amount of performance drop based on the type of noise in these systems, both, for any kind of noise, and then, specifically for noises that do not change human perception (perception-retaining).

Purpose 3: Denoising and Augmentation Benefits Evaluation

RQ3: Does dataset augmentation (Q3a) and/or sample denoising (Q3b) help improve the robustness of emotion recognition models to unseen noise?

Reason: We test whether the commonly-used methods for improving the performance of these models under distribution shifts is helpful. We focus on two main methods, augmentation and denoising. We specifically look at how performance changes when we augment with noises that include those that are perception-altering vs. when we exclude such noises.

Purpose 4: Model Robustness Testing Conditions

RQ4: How does the robustness of a model to attacks compare when we are using test samples that with are augmented with perception-retaining noise vs. samples that are augmented with all types of noise, regardless of their effect on perception?

Reason: Another major metric for any deployable machine learning algorithm is its performance on "unseen situations" or handling incoming data shifts (i.e., robustness testing). We test robustness using a noise augmentation algorithm that aims to forcefully and efficiently change a model's output by augmenting test samples with noise. We look at how often this algorithm is unsuccessful in being able to "fool" a model with its augmented samples. We look at the changes in frequency with which a model is successfully able to defend itself when the attack algorithm uses a set that includes all types of noises vs. when it only uses perception-retaining noises.

Purpose 5: Recommendations

RQ5: What are the recommended practices for speech emotion dataset augmentation and model deployment?

Reason: We then provide a set of recommendations based on our empirical studies for deploying emotion recognition models in real world situations.

5.4 Noise

We investigate the effects of two types of noise, environmental and signal distortion. Environmental noises are additive, while signal distortion noise involves other types of signal manipulation.

5.4.1 Environmental Noise

We define environmental noises (*ENV*) as additive background noise, obtained from the ESC-50 dataset[176][1]. ESC-50 is generally used for noise contamination and environmental sound classification [231]. These environmental sounds are representative of many types of noise seen in real world deployments, especially in the context of virtual and smart home conversational agents. We use the following categories:

- Natural soundscapes (*Nat*), e.g., rain, wind.

- Human, non-speech sounds (*Hum*), e.g., sneezing, coughing, laughing or crying in the background etc.

- Interior/domestic sounds (*Int*), e.g., door creaks, clock ticks etc.

We manipulate three factors when adding the noise sources:

- *Position*: The position of the introduction of sound that: (i) starts and then fades out in loudness or (ii) occurs during the entirety of the duration of the utterance. In the second case, this complete additive background would represent a consistent noise source in real world (e.g., fan rotation).

- *Quality Degradation*: The decrease in the signal to noise ratio (SNR) caused by the addition of the additive background noise at levels of 20dB, 10dB and 0dB. This is used only when noise is added to the entirety of the utterance.

---

[1]https://github.com/karoldvl/ESC-50

5.4.2 Signal Distortion

We define signal distortion noise as modulations that aren't additive in the background. These kinds of noise in the audio signal can occur from linguistic/paralinguistic factors, room environment, internet lags, or the physical locomotion of the speaker.

We use the nine following categories:

- *SpeedUtt*: The utterance is sped up by either 1.25× or 0.75×.

- *SpeedSeg*: A random segment within an utterance is sped up by 1.25×. The package pyAudio [2] that we used to speed up a segment did not permit slowing a segment down. Thus, the 0.75× was not used here.

- *Fade*: The loudness of the utterance is faded by 2% every second, which emulates the scenario of a user moving away from the speaker. The loudness is increased for fade in, and decreased for fade out.

- *Filler*: Non-verbal short fillers such as 'uh', 'umm' (from the same speaker) are inserted in the middle of a sentence. The insertion is either just the filler or succeeded and preceded by a long pause [3].

- *DropWord*: A randomly selected set of non-essential words belonging to the set: {a, the, an, so, like, and} are dropped from an utterance using word-aligned boundaries and stiching the audio segments together.

- *DropLetters*: Following the same approach as drop word, letters are dropped in accordance with various linguistic styles chosen from the set: {**/h/**+vowel, vowel+/n**d**/+consonant(next word), consonant+**/t/**+consonant(next word), vowel+**/r/**+consonant, /ihn**g**/}. This is supported by research that has studied phonological deletion or dropping of letters in the native US-English dialect [1, 237].

---

[2]https://people.csail.mit.edu/hubert/pyaudio/

[3]Fillers are obtained by parsing audio files for a given speaker and finding occurrences of any of the options from the above mentioned set. We will release the extracted fillers per speaker for IEMOCAP

- *Laugh/Cry*: "Sob" and "short-laughter" sounds are added to the utterance. They are obtained from AudioSet [82].

- *Pitch*: The pitch is changed by ± 3 half octaves using the pyAudio library.

- *Rev*: Room reverberation is added to the utterance using py-audio-effects (pysndfx[4]). We vary metrics such as reverberation ratio or room size to vary the type and intensity of reverberation added.

5.4.3 Sampling and Noise-Perturbations

We randomly select 900 samples from the IEMOCAP dataset, which is far larger than the ones used for previous perception studies [170, 191]. We select 100 samples from each activation and valence pair bin, i.e., 100 samples from the bin with activation: *low*, valence: *low*; 100 samples from the bin with activation: *low*, and valence: *mid*, and so on. This ensures that the chosen 900 samples cover the range of emotions expressed. We impose another constraint on these 100 samples from each bin, 30 of them are shorter than the first quartile or greater than fourth quartile of utterance length in seconds to cover both extremities of the spectrum, and the remaining 70 belong in the middle. We also ensure that the selected samples had a 50-50 even split amongst gender. We introduce noise to the 900 samples (Section 3.1). Each sample is modulated in ten ways: four randomly chosen types of environmental noise and six randomly chosen signal distortion noise modulations, giving us a total of 9,000 noisy samples[5].

5.5 User study

We first analyze the effects of noise on human perception by relabeling the noise-enhanced data using the Amazon Mechanical Turk (AMT) platform. We use insights from this experiment to guide the machine learning analyses that follow.

---

[4]https://github.com/carlthome/python-audio-effects
[5]We will release the script to create these files.

5.5.1 Crowdsourcing Setup

We recruited 147 workers using Amazon Mechanical Turk who self-identify as being from the United States and as native English speakers, to reduce the impact of cultural variability. We ensured that each worker had > 98% approval rating and more than 500 approved Human Intelligence Tasks (HITs). We ensured that all workers understood the meaning of activation and valence using a qualification task that asked workers to rank emotion content similar to [105]. The qualification task has two parts: (i) we explain the difference between valence and activation and how to identify those, and, (ii) we ask them to identify which of the two samples has a higher/lower valence and a higher/lower activation, to ensure that they have understood the concept of activation and valence annotations. All HIT workers were paid a minimum wage ($9.45/hr), pro-rated to the minute. Each HIT was annotated by three workers.

For our main task, we created pairs that contained one original and one modulated sample. We then asked each worker to annotate whether or not they perceived the pair to have the same emotion. If they said *yes* for both activation and valence, the noisy sample was labeled *same* and they could directly move to the next HIT. If they said *no*, the noisy sample was labeled *different*. In this case, they were asked to assess the activation and valence of the noisy sample using Self Assessment Manikins [30] on a scale of [1, 5] (similar to the original IEMOCAP annotation).

We also include three kinds of attention checks:

1. We show two samples that have not been modified and ask them to decide if the emotion represented was different. If the person says yes, then the experiment ends.

2. We observe the time spent on the task. If the time spent on the task is less than the combined length of both samples, then the user's qualification to annotate the HITs is rescinded and their responses are discarded.

3. We show two samples, one which has a gold standard label, and another, which has

been contaminated with significant noise (performance degradation >30dB), such that the resulting sample is incomprehensible. If people do not mark this set of samples as being different, the experiment ends.

The failure rate based on the above criteria was 8%. We ensured the quality of the annotations by paying bonuses based on time spent, not just number of HITs, and by disqualifying annotators if they annotated any sample (including those outside of the attention checks) more quickly than the combined length of the audio samples.

We then created two sets of labels for each noise-augmented clip. The *first type* of label compared a noise-augmented clip to its original. The noise-augmented clip was labeled the *same* if the modified and original clip were perceived to have the same valence or activation, otherwise it was labeled *different*. We created this label by taking the majority vote over all evaluations. The *second type* of label included valence and activation. A noise-augmented clip was given the average valence and activation over all evaluations.

The inter-annotator agreement was measured using Cohen's kappa. Conventionally, when estimating Cohen's kappa, annotators are not considered as individuals, instead reducing annotators to the generic 1, 2, and 3. The challenge is that this often leads to artificially inflated inter-annotator agreement because individual characteristics and behavior of a particular worker are not taken under consideration [95]. We take a different approach, creating a table for the calculation of the statistic that considers annotators as individuals with separate entries for each clip, following the approach of [95]. If an annotator didn't evaluate a given clip, the cell has a null (missing data) value. We found that the Cohen's kappa was 79% for activation and 76% for valence [6].

Randomly sample 1 noise variation from each category mentioned in Section 5.4.;
$numAttempts = 0$;
for *each noise in selected random noises:* **do**
    Add noise to the sample such that the decrease in SNR is 1.;
    Get the classifier output with this new sample variation.;
    $numAttempts+ = 1$;
    if $numAttempts > k$ **then**
        return *Exit Code = Failure*
    end
    if *classifier output changes* **then**
        return *Exit Code = Success*
    end
end
for *each noise in selected random noises:* **do**
    Add noise to the sample such that the decrease in SNR is 5.;
    Get the classifier output with this new sample variation.;
    $numAttempts+ = 1$;
    if $numAttempts > k$ **then**
        return *Exit Code = Failure*
    end
    if *classifier output changes* **then**
        return *Exit Code = Success*
    end
    if *classifier output changes* **then**
        while *classifier output does not change* **do**
            Iterate over all SNR decreases from 2-5;
            Get classifier output for the modified sample;
            $numAttempts+ = 1$;
            if $numAttempts > k$ **then**
                return *Exit Code = Failure*
            end
            if *classifier output changes* **then**
                return *Exit Code = Success*
            end
        end
    end
end
for *each noise in selected random noises:* **do**
    Add noise to the sample such that the decrease in SNR is 10.;
    Get the classifier output with this new sample variation.;
    if *classifier output changes* **then**
        return *Exit Code = Success*
    end
    if *classifier output changes* **then**
        while *classifier output does not change* **do**
            Iterate over all SNR decreases from 6-10 Get classifier output for the modified sample;
            $numAttempts+ = 1$;
            if $numAttempts > k$ **then**
                return *Exit Code = Failure*
            end
            if *classifier output changes* **then**
                return *Exit Code = Success*
            end
        end
    end
end
return *Exit Code = Failure*

58

Table 5.1: The table shows the ratio of samples marked by human evaluators as imperceptible to difference in emotion perception. *V: Valence, A: Activation, δV:Average change in Valence on addition of noise, δA:Average change in Activation on addition of noise.*

		V	A	δV	δA
Environmental Noise					
NatSt		0.01	0.00		
NatdB (Co)	-10dB	0.01	0.00		
	Same	0.02	0.00		
	+10dB	0.03	0.01		
HumSt		0.01	0.00		
HumdB (Co)	-10dB	0.03	0.00		
	Same	0.02	0.00		
	+10dB	0.04	0.01		
IntSt		0.05	0.01		
IntdB (Co)	-10dB	0.02	0.01		
	Same	0.02	0.00		
	+10dB	0.04	0.01		
Signal Distortion					
SpeedSeg		0.01	0.0		
Fade	In	0.04	0.01		
	Out	0.04	0.00		
DropWord		0.01	0.00		
DropLetters		0.01	0.00		
Reverb		0.04	0.01		
Filler	L	0.10	0.06		
	S	0.06	0.03		
Laugh		0.16	0.17	+ .11	+ .26
Cry		0.20	0.22	- .20	- .43
SpeedUtt	1.25x	0.13	0.03	- .10	- .13
	0.75x	0.28	0.06	- .18	- .23
Pitch	1.25x	0.22	0.07	- .11	+ .19
	0.75x	0.29	0.10	- .07	- .15

Table 5.2: State of the art model performance in terms of UAR when using the general versions of traditional deep learning and end-to-end deep learning models. No noise refers to clean speech, all noises refers to the combined set of perception-retaining and perception-altering noise. The environmental and signal distortion categories shown include only the perception-retaining noises. As a reminder, samples in the all noises category have an uncertain ground truth, the row is marked with two stars (∗∗). *V: Valence, A: Activation, Clean: Training on clean dataset, Clean+Noise — Mismatch: Cleaning on noisy dataset and testing on mismatched noisy partition, Clean+Noise — Match: Cleaning on noisy dataset and testing on matched noisy partition.* Random chance UAR is 0.33.

			Traditional Deep Neural Network						End-To-End Deep Neural Network					
			Clean		Clean+Noise				Clean		Clean+Noise			
					Mismatch		Match				Mismatch		Match	
			A	V	A	V	A	V	A	V	A	V	A	V
	No Noise		0.67	0.59	-	-	-	-	0.70	0.63	-	-	-	-
	All Noises**		0.40	0.38	0.55	0.42	0.66	0.59	0.70	0.63	0.44	0.38	0.67	0.60
	Perception Retaining Noises		0.50	0.42	0.57	0.48	0.60	0.52	0.53	0.45	0.60	0.50	0.62	0.54
Environmental Category	**Nature** At Start		0.50	0.45	0.61	0.53	0.63	0.55	0.56	0.48	0.64	0.55	0.66	0.58
	Cont.	-5dB	0.45	0.39	0.55	0.44	0.59	0.50	0.49	0.42	0.58	0.47	0.62	0.53
		-10dB	0.42	0.35	0.56	0.46	0.59	0.49	0.47	0.38	0.57	0.48	0.61	0.51
		-20dB	0.40	0.35	0.51	0.44	0.55	0.47	0.47	0.39	0.52	0.46	0.56	0.50
	Interior At Start		0.53	0.44	0.61	0.52	0.64	0.57	0.57	0.47	0.64	0.56	0.66	0.59
	Cont	-5dB	0.46	0.36	0.55	0.44	0.58	0.49	0.49	0.39	0.59	0.49	0.62	0.51
		-10dB	0.44	0.36	0.54	0.43	0.57	0.48	0.49	0.39	0.56	0.45	0.59	0.52
		-20dB	0.40	0.35	0.52	0.44	0.55	0.49	0.46	0.37	0.56	0.45	0.58	0.51
	Human At Start		0.52	0.45	0.60	0.51	0.63	0.55	0.58	0.47	0.62	0.52	0.66	0.57
	Cont	-5dB	0.45	0.37	0.52	0.43	0.55	0.48	0.49	0.40	0.56	0.44	0.57	0.50
		-10dB	0.42	0.34	0.51	0.43	0.53	0.47	0.49	0.38	0.53	0.45	0.55	0.49
		-20dB	0.40	0.34	0.50	0.41	0.53	0.46	0.46	0.38	0.54	0.43	0.56	0.48
Signal Distortion	Speed Segment		0.61	0.52	0.63	0.53	0.64	0.55	0.63	0.55	0.64	0.55	0.67	0.58
	Fade	In	0.62	0.53	0.65	0.55	0.67	0.58	0.64	0.55	0.67	0.58	0.68	0.59
		Out	0.61	0.51	0.62	0.54	0.64	0.57	0.63	0.54	0.64	0.56	0.66	0.59
	DropWord		0.64	0.56	0.65	0.56	0.67	0.59	0.65	0.58	0.67	0.58	0.69	0.61
	DropLetters		0.65	0.58	0.69	0.60	0.71	0.62	0.66	0.59	0.72	0.60	0.74	0.63
	Reverb		0.43	0.37	0.50	0.43	0.53	0.45	0.35	0.34	0.51	0.42	0.55	0.46

Table 5.3: Noise-Robust (NR) state of the art model performance in terms of UAR when using the noise-robust versions of traditional deep learning and end-to-end deep learning models. No noise refers to clean speech, all noises refers to the combined set of perception-retaining and perception-altering noise. The environmental and signal distortion categories shown include only the perception-retaining noises. As a reminder, samples in the all noises category have an uncertain ground truth, the row is marked with two stars (∗∗). *V: Valence, A: Activation, Clean: Training on clean dataset, Clean+Noise — Mismatch: Cleaning on noisy dataset and testing on mismatched noisy partition, Clean+Noise — Match: Cleaning on noisy dataset and testing on matched noisy partition, NR: Noise Robust versions of the corresponding models* Random chance UAR is 0.33.

			NR-Traditional Deep Neural Network						NR-End-To-End Deep Neural Network					
			Clean		Clean+Noise				Clean		Clean+Noise			
					Mismatch		Match				Mismatch		Match	
			A	V	A	V	A	V	A	V	A	V	A	V
No Noise			0.67	0.59	-	-	-	-	0.70	0.63	-	-	-	-
All Noises**			0.44	0.40	0.58	0.44	0.68	0.60	0.50	0.40	0.50	0.40	0.72	0.61
Perception Retaining Noises			0.52	0.44	0.59	0.50	0.61	0.51	0.55	0.48	0.61	0.52	0.63	0.54
Environmental Category	**Nature** At Start		0.54	0.49	0.63	0.55	0.63	0.55	0.57	0.50	0.66	0.55	0.67	0.56
	Cont.	-5dB	0.50	0.42	0.58	0.50	0.60	0.52	0.49	0.42	0.61	0.51	0.63	0.53
		-10dB	0.48	0.38	0.59	0.48	0.61	0.51	0.50	0.46	0.63	0.52	0.65	0.53
		-20dB	0.44	0.38	0.55	0.49	0.59	0.51	0.50	0.43	0.53	0.46	0.58	0.47
	Interior At Start		0.53	0.44	0.61	0.52	0.65	0.54	0.58	0.52	0.63	0.56	0.66	0.58
	Cont	-5dB	0.46	0.36	0.55	0.44	0.59	0.47	0.51	0.42	0.58	0.48	0.62	0.52
		-10dB	0.44	0.36	0.54	0.43	0.57	0.43	0.49	0.44	0.57	0.48	0.61	0.50
		-20dB	0.40	0.35	0.52	0.44	0.55	0.46	0.46	0.40	0.55	0.48	0.58	0.49
	Human At Start		0.52	0.45	0.60	0.51	0.63	0.52	0.59	0.49	0.63	0.53	0.66	0.56
	Cont	-5dB	0.45	0.37	0.52	0.43	0.55	0.46	0.49	0.42	0.56	0.48	0.58	0.50
		-10dB	0.42	0.34	0.51	0.43	0.54	0.44	0.47	0.38	0.54	0.49	0.55	0.45
		-20dB	0.40	0.34	0.50	0.41	0.52	0.43	0.43	0.38	0.54	0.45	0.58	0.49
Signal Distortion	Speed Segment		0.63	0.55	0.65	0.57	0.67	0.58	0.66	0.58	0.67	0.59	0.67	0.60
	Fade	In	0.64	0.55	0.66	0.57	0.68	0.59	0.65	0.58	0.67	0.59	0.69	0.60
		Out	0.64	0.56	0.65	0.57	0.67	0.58	0.66	0.57	0.68	0.59	0.69	0.63
	DropWord		0.67	0.60	0.66	0.58	0.66	0.59	0.68	0.60	0.69	0.60	0.69	0.60
	DropLetters		0.67	0.60	0.65	0.62	0.66	0.60	0.68	0.64	0.69	0.60	0.69	0.60
	Reverb		0.48	0.41	0.56	0.47	0.58	0.48	0.52	0.40	0.55	0.45	0.60	0.45

Table 5.4: Success of misclassification attempts on different models with varying number of allowed attempts (lower is better). As a reminder, samples in the all noises category have an uncertain ground truth, the row is marked with two stars (∗∗). Reverberation (reverb) is a perception-retaining noise that is also analyzed separately. *Trad: Traditional Deep Learning Model, E2E: End to End deep learning model, NR: Noise Robust version of the deep learning model.*

	Noise Set	No. of Attempts	Activation				Valence			
			Trad	**E2E**	**NR-Trad**	**NR-E2E**	**Trad**	**E2E**	**NR-Trad**	**NR-E2E**
Performance Impact Per Noise Is Unknown	All Noises**	5	0.29	0.22	0.15	0.10	0.11	0.15	0.13	0.05
		15	0.31	0.33	0.31	0.32	0.23	0.24	0.22	0.22
		25	0.40	0.28	0.40	0.33	0.22	0.19	0.21	0.20
		inf	0.43	0.41	0.44	0.35	0.25	0.20	0.26	0.18
	Perception-Retaining	5	0.18	0.11	0.07	0.05	0.11	0.05	0.02	0.02
		15	0.25	0.12	0.25	0.15	0.14	0.08	0.18	0.10
		25	0.32	0.24	0.32	0.20	0.13	0.10	0.11	0.09
		inf	0.40	0.26	0.36	0.23	0.19	0.14	0.17	0.14
	Reverb	5	0.33	0.15	0.22	0.18	0.20	0.14	0.22	0.09
		15	0.40	0.23	0.30	0.21	0.30	0.13	0.34	0.16
Performance Impact Per Noise Is Known	All Noises**	5	0.33	0.28	0.24	0.22	0.15	0.14	0.12	0.12
		15	0.38	0.38	0.32	0.33	0.20	0.21	0.18	0.16
		25	0.52	0.32	0.44	0.37	0.25	0.16	0.18	0.16
		inf	0.54	0.42	0.46	0.41	0.24	0.20	0.22	0.22
	Perception-Retaining	5	0.29	0.16	0.22	0.13	0.14	0.10	0.15	0.08
		15	0.32	0.32	0.32	0.28	0.14	0.15	0.16	0.12
		25	0.47	0.30	0.47	0.29	0.22	0.18	0.23	0.16
		inf	0.51	0.36	0.50	0.32	0.22	0.19	0.25	0.17
	Reverb	5	0.38	0.22	0.33	0.22	0.28	0.20	0.31	0.21
		15	0.47	0.29	0.41	0.28	0.30	0.23	0.35	0.26

Table 5.5: Hyper-parameters used to select the best performing model on validation subset whilst training the traditional deep learning model.

Hyper-parameter	Values
Traditional	
No. of Convolution Kernels	{64, 128}
Convolution Kernels Width	{2}
Number of Convolution Layers	{5}
Number of GRU layers	{1, 2, 3}
Pooling Kernel Width	{2, 4}
GRU Layers Width	{32, 64}
Number of Dense Layers	{1, 2, 3}
End to End	
No. of Dense Layers	{1, 2}

5.6   Methods

We now describe the emotion recognition approaches, presenting two separate pipelines, one that relies upon direct feature extraction (Section 5.6.2) and the other that is end-to-end (Section 5.6.3). This allows us to investigate whether noise has a consistent effect. We discuss approaches to improve noise robustness by training models with noise-augmented data or denoised data (Section 5.6.4). Finally, we describe the setup and evaluation of the model robustness using an untargeted model misclassification test, which measures a model's fragility in terms of how likely it is that the model's decisions will change when specific types of noise are observed at test time (Section 5.6.5).

5.6.1   Creation of Data Partitions

We use a subject-independent five-fold cross validation scheme to select our train, test and validation sets. In the first iteration, sessions 1-3 are used for training, session 4 is used as validation, and session 5 is used for testing. This is repeated in a round-robin fashion, resulting in each session serving as a validation and a test fold. We also divide possible

---

[6]The sample name, code to create the paired noisy examples, and the resulting annotations will be made available for further research

noises in two different categories based on results of crowdsourcing study (see Section 5.7.1). The first category is *perception-altering*, those that changed perception of humans and hence cannot be used for model training or evaluation with the old annotations. The second category is *perception-retraining*, those that did not change human perception, and hence, the model should produce no change in predictions when using those noise categories for sample augmentation.

We use the noise categories (seeSection 5.4) in two varying circumstances. The first category is *matched*, where both the training and testing sets are augmented with same kinds of noise (e.g., both have nature-based sounds in them). The second category is *mismatched*, where the testing set is augmented with a noise *category* not used for augmenting the training set (e.g., only the test set is augmented with nature-based noise while the train set is augmented with human or interior noises).

5.6.2 Traditional Deep Learning Network

We first explore a common "traditional" deep learning network that is used in speech emotion recognition. In this method we extract Mel Filterbank (MFB) features as input to a model composed of convolutional and gated recurrent unit (GRU) layers.

5.6.2.1 Features

We extract 40-dimensional Mel Filterbank (MFB) features using a 25-millisecond Hamming window with a step-size of 10-milliseconds using python-speech-features [7]. Each utterance is represented as a sequence of 40-dimensional feature vectors. We *z*-normalize the acoustic features using parameters extracted from the training dataset. During each cross-validation fold, the parameters are chosen from the training data and are applied to both the validation and testing data.

---

[7]https://github.com/jameslyons/python_speech_features

5.6.2.2 Network

Our baseline network is a state-of-art single utterance emotion classification model which has been used in previous research [9, 116, 126]. The extracted MFBs are processed using a set of convolution layers and GRUs (see Table 5.5 for the hyperparameters used for these layers). The output of these layers is then fed through a mean pooling layer to produce an acoustic representation which is then fed into a set of dense layers to classify activation or valence.

5.6.2.3 Training.

We implement the models using the Keras library [51]. We use a cross-entropy loss function for each task (e.g., valence or activation). We learn the model parameters using the RMSProp optimizer. We train our networks for a maximum of 50 epochs and use early stopping if the validation loss does not improve after five consecutive epochs. Once the training process ends, we revert the network's weights to those that achieved the lowest validation loss. We repeat the experiment five times. We report the results in terms of Unweighted Average Recall (UAR, chance is 0.33), averaged over all test samples and five repetitions. We compare the performance of different models or the same model in different noisy conditions/partitions using a paired t-test using the Bonferroni correction, asserting significance when $p \leq 0.05$.

5.6.3 End-to-End Deep Learning Networks

Next, we explore a transformer-based model. In this method the raw audio signal is used as input to a pre-trained and fine-tuned network and the emotion prediction is directly obtained as an output. These models do not require us to perform manual or domain knowledge-based extraction of features. They instead have a feature encoder component inside the model, which is dynamic in nature, and hence, can change its output for the same signal based on the dataset and nature of the task.

5.6.3.1 Features

For the end-to-end deep learning models, we do not need to extract audio features. Instead we rely on the network itself to both normalize and extract features, that are later passed onto the deeper layers of the network. The feature set here is the original wav files that are not modified in any capacity. The eventual representations are of size 512, reproducing the setup in the state-of-the-art implementation [175].

5.6.3.2 Network

Our baseline network is the state-of-the-art wav2vec2.0 emotion recognition model [175]. The wav2vec model is comprised of three parts: (i) a convolutional neural network (CNN) that acts as feature encoder, (ii) a quantizier module, and (iii) a transformer module. The input to the model is raw audio data (16kHz) that is passed to a multi-block 1-d CNN to generate audio representations (25ms). The quantizer is similar to a variational autoencoder that encodes and extracts features using a contrastive loss. The transformer is used for masked sequence prediction and encodes the bi-directional temporal context of the features.

We use the base model, which has not been fine-tuned for ASR (wav2vec2.0-PT). We then fine-tune the base model to predict the binned emotion labels. We use the final representation of the output as an input to dense layers to produce the final output.

5.6.3.3 Training

We implement the model provided in the speech brain library [8]. As in the other pipeline (Section 5.6.2), we use cross-entropy loss for each task and learn the dense layer parameters. Reproducing the state of the art model [175] We run the model for a maximum of eight epochs. We revert the network's state to the one that achieved the lowest validation loss. We repeat this experiment five times. Again, we use UAR and report the results averaged over

---

[8]https://speechbrain.readthedocs.io/en/latest/API/
speechbrain.lobes.models.fairseq_wav2vec.html

both subjects and repetitions.

5.6.4 Noise Augmentation Overview

We will be assessing a model's ability to classify emotion given either environmental or signal distortion noise. We perform two kinds of analysis, one when using the set of noises that includes those that do alter human perception, and another when only using noises that are perception-retaining. We report overall model performances for both of these categories.

For a more thorough analysis, we then specifically focus on the categories of noise that do not significantly affect human perception. This allows us to evaluate a model's robustness, or its fragility, with respect to variations that wouldn't alter a human's perception of emotion. This is important because the overwhelming majority of the noise-augmented utterances in the IEMOCAP dataset were not included in the user study and, therefore, do not have perceptual labels (Section 5.5). We consider three types of environmental noise {Human (Hum), Interior (Int), Natural (Nat)} and three types of signal distortion noise {Speeding a segment (SpeedSeg), Fade, Reverberation (Reverb)}.

We use two separate testing paradigms: (i) matched testing, in which all noise types are introduced to the training, testing, and validation data and (ii) mismatched testing, in which n-1 types of noise are introduced to the training and validation sets and the heldout type of noise is introduced to the test set. In all cases, we analyze the test data in terms of specific noise categories. Therefore, the test sets are the same between the two paradigms.

We run both the matched and mismatched experiments twice, first with the noise-augmented data and second with a noise-robust/denoising pipeline. The first iteration will allow us to quantify the effect of the noise on the traditional and end-to-end classification pipelines. We then repeat the experiment with either denoised data for the traditional classifier (Section 5.6.4.1) or using the noise-robust implementation of wav2vec2.0 for the end-to-end classifier (Section 5.6.4.2). This allows us to investigate how, or if, noise-robust implementations can offset the effects of background noise.

5.6.4.1 Denoising

We implement denoising using the well-known Recurrent Neural Network Noise Suppression (RNNNoise, denoising feature space) approach, proposed in 2017 for noise suppression [218]. RNNNoise is trained on environmental noise, and these noises overlap considerably with those in our dataset. We use the algorithm's default parameters and use it on an 'as-is' basis for our experiments. We assume that the system does not have the knowledge of which noise, from the set of available noise categories, is introduced and, therefore, we do not compare with other denoising algorithms that assume a priori knowledge of noise category. The result is a set of 'noise-suppressed' samples in the training, validation and testing sets.

We pass all the data, including both the original and noise-augmented data, through a denoising algorithm. This allows us to ensure that acoustic artifacts, if any, are introduced to both the original and noise-augmented data. We then train the traditional deep learning model as described in Section 5.6.2.

5.6.4.2 Using a Noise-Robust Model

In the end-to-end model, we need to use a different denoising approach because the approach described in the previous section does not return a wav file, but instead is applied to the feature-space directly. Here, we enforce robustness to noise using a model trained to be noise-robust in an end-to-end fashion. We use the noise-robust version (Wav2Vec2-Large-Robust) of the aforementioned wav2vec2.0 model [98]. The noise-robust large model was pretrained on 16kHz sampled speech audio. Noisy speech datasets from multiple domains were used to pretrain the model: Libri-Light, CommonVoice, Switchboard, and, Fisher [98]. We then train the end-to-end model as described in Section 5.6.3.

5.6.5   Model Robustness Testing

Deployed emotion recognition models must be robust. One of the major scenarios that we robustness test any speech-based model for is the presence of noise. But the set of noises we choose to test robustness on can lead to different conclusions about the robustness of the models. In our case, we consider two different scenarios:

1. Robustness evaluation when using perception-retaining samples, noise samples that do not change human perception

2. Robustness evaluation when using any kind of noise (i.e., both perception-retaining and perception-altering)

We perform robustness evaluation of a model by using the model's output predictions to create new noise-enhanced samples that change the model's output, compared to the original clean sample. We do this using an untargeted model misclassification test, in which we add noise to the samples. The intentional misclassification algorithm assumes black-box model access. For our purposes, it needs to have access to: (i) a subset of the dataset, (ii) noises to add to create perturbed samples, and (iii) model input and output.

As in any other perturbation-based robustness testing, the goal is to introduce perturbations to the samples such that the resulting samples are as close to the original sample as possible. The minimally perturbed sample should be the one that causes a classifier to change its original classification. We measure the amount of perturbation using SNR, calculated using the logarithmic value of the ratio between the original signal and the noise-augmented signal's power. We note that the lower the decrease in SNR, the more minimally perturbed a sample is. The maximal decrease in SNR that we use in the algorithm is a difference of 10 dB. This condition ensures that the sample is not audibly judged as contaminated by humans [119].

The algorithm to choose this minimally perturbed sample has four major components:

1. Requirements: some labelled samples, noise files, model input and output access, unlabelled samples for testing, and, optionally, correlation between noise type and performance degradation for a given model.

2. Looping: The algorithm then loops over each noise category to figure out whether it can successfully force the model to misclassify. The noise category order is random if we do not have access to the optional performance degradation correlations.

3. Minimizing: The algorithm then aims to find the lowest decrease in dB, such that the model still misclassifies. This ensures that the resultant noisy sample is as imperceptible to humans as possible.

4. End Condition: The algorithm ends if a noise addition has been found, or if it runs out of number of tries allowed for model access.

Please see Algorithm 1 for more details. In the algorithm, *numAttempts* is the number of times the algorithm is allowed to access the model's input-output pairs. *Classifier output* refers to the prediction made by the model when the attack algorithm sends an input to the model to be classified. *Classifier output changes* is true when the model predicted the emotion label differently after noise was added to the sample, compared to the original clean sample. *Success* implies that the algorithm was successfully able to force the model to misclassify a particular sample in the allowed number of attempts. *Failure* implies that the algorithm could not force the model to misclassify in the allowed number of attempts and that the model can be considered robust for that sample.

We use the above algorithm in two different settings, under two different pre-known assumptions, with four levels of allowed queries, and four models (64 categories):

1. Settings {All vs. Not-Altering Human Perception}

2. Pre-Known Assumptions {No Knowledge vs. Knowledge About Noise Category Degradation Level}

3. Allowed Queries {5, 15, 20, inf}

4. Models {Traditional Deep Neural Network (T-DNN), End to End Deep Neural Network (E-DNN), Noise-Robust T-DNN (T-DNN-NR), Noise-Robust E-DNN (E-DNN-NR)}

For all the test samples, we execute five runs of the above algorithm to account for randomization in noise choices. These five runs are then averaged to obtain the average success of misclassification or average robustness for a given sample (1- average success of misclassification). We then average the robustness value over all the test samples. We report our obtained results for the above mentioned scenarios.

5.7   Analysis

5.7.1   Research Question 1 (Q1): Does the presence of noise affect emotion perception as evaluated by *human raters*? Is this effect dependent on the utterance length, loudness, the type of the added noise, and the original emotion?

We find that the presence of environmental noise, even when loud, rarely affects annotator perception, suggesting that annotators are able to psycho-acoustically mask the background noise in various cases, as also shown in prior work (e.g., [206]).

We find that the addition of signal distortion noise alters human perception. The reported change in valence and activation values is on a scale of -1 to 1 (normalized). The addition of laughter changes the activation perception of 16% of the utterances, with an average change of +22% (+.26). The valence perception is altered in 17% of the utterances, with an average change of +14% (+.11). Similarly for crying, valence is altered in 20% of the cases, with an average change of -21% (-.20). Crying changes activation perception in 22% of the cases, with an average change of -32% (-.43). Raises in pitch also alter the perception of emotion. In 22% of utterances, the perception of activation is changed. This contrasts with the perception of valence, which was altered only in 7% of utterances. In this scenario, activation increases by an average of 26% (+.19), and valence decreases by 12% (-.11). On

the other hand, decreases in pitch change the perception of activation in 10% of the cases and of valence in 29% of the cases. In this scenario, activation decreases by an average of 16% (-.15), and valence decreases by 7% (-.07). This ties into previous work [37], which looked into how changes and fluctuations in pitch levels influenced the perception of emotions. Changes in the speed of an utterance affect human perception of valence in 13% (average of -.13) of the cases when speed is increased, and 28% (average of -.23) when speed is decreased. On the other hand, changes in the speed of an utterance do not affect activation as often, specifically, 3% in case of increase and 6% in case of decrease.

We ensured that our crowsdourcing samples had an even distribution over gender of the speaker and the length of the sample (see Section 5.4.3). We performed paired t-test to evaluate whether these variables influenced the outcome of emotion perception change in presence of noise. We found that the changes in perception were not tied to characteristics of the speakers. For example, there was no correlation between changes in perception and variables such as, the original emotion of the utterance, the gender of the speaker, and the length of the utterance.

The human perception study provides insight into how emotion perception changes given noise. This also provides information about the potential effects of noise addition on model behavior. In the sections that follow, we will evaluate how machine perception changes given these sources of noise.

5.7.2 RQ2: Can we verify previous findings that the presence of noise affects the performance of *emotion recognition models*? Does this effect vary based on the type of the added noise?

We first assess the performance of the model on the original IEMOCAP data and find that the traditional model obtains a performance of 0.67 UAR on the activation and 0.59 UAR on the valence task. On the other hand, the end-to-end model obtains a performance of 0.73 on activation and 0.64 on the valence task. We hypothesize that the wav2vec2 model

has an added advantage of being trained to recognize word structures that can incorporate some paralinguistic/langauge information in the fine-tuned model.

Next, we augment the test samples of each fold with each of the noise types (Section 5.4) and investigate how the performance of the model changes. We include two cases: (i) only perception-retaining noises and,(ii) all noises.

In the first scenario, we do not include noise types that were found to affect human perception (e.g., *Pitch, SpeedUtt, Laugh*) because once these noises are added, the ground truth is no longer reliable. This lack of reliable ground-truth data hinders the evaluation of the model's performance on these samples because the majority of the utterances were not part of the original crowdsourcing experiment and are thus unlabeled. The remainder of this section focuses on the second scenario only.

We find that for matched train and test noise conditions, the traditional machine learning model's performance decreases by an average of 28% for environmental noise while it drops by 32% for signal manipulation. On the other hand, for end-to-end deep learning model, the model's performance decreases by an average of 22% and 26% for environmental and manipulated noises, respectively. In mismatched noise conditions, the models' performance decreases by an average of 33% for environmental noise, fading, and reverberation. There is also a smaller drop in performance for speeding up parts of the utterance and dropping words, showing the brittleness of these models. Table 5.2 reports the percentage change in performance when testing on noisy test data, compared to clean test data.

We see that the end-to-end deep learning model is less affected by environmental noise, but has a larger drop due to fading and reverberation. We observe a larger drop on performance when dropping words, which possibly can be attributed to the change in audio-structure and non-controllable feature extraction for this model.

In the second scenario, we observe a large drop in performance for both the traditional and the end to end machine learning model. For example, in the case of a traditional deep learning model, the valence prediction performance drops to a near-chance performance

when including all kinds of noises (see Table 5.2).

We specifically want to point out how the inclusion of all noises in the test conditions changes the observed model performance. Primarily, the models on an average seem to do 20% worse than they would if we only consider noises that do not alter human perception. We note the discrepancy between the results of the two noise addition scenarios and that results should be described with respect to the perceptual effects of noise, if noise augmentation is used.

5.7.3 RQ3a: Does dataset augmentation help improve the robustness of emotion recognition models to unseen noise?

We first report results for only perception-retaining noises. When the training datasets are augmented with noise, we observe an average performance drop of 26% and 10% for matched noise conditions when using the traditional and the end-to-end deep learning model, respectively. For the mismatched noise conditions, we observe an average performance drop of 31% and 16% for the traditional and end-to-end deep learning models, respectively.

Both models see improved performance when the training dataset is augmented with continuous background noise in the matched noise setting. We find that data augmentation improves performance on mismatched noisy test data over a baseline system trained only on the clean IEMOCAP data. For example, the end-to-end model tested on environmental noise-augmented dataset (as compared to traditional deep learning model), reduces the performance drop to nearly zero. This improvement is particularly pronounced (an increase of 22% as compared to when trained on the clean partition) when the environmental noise is introduced at the start of the utterance (e.g., when the test set is introduced with nature-based noises at the start of the utterance and the train set is introduced with human and interior noises at the start of the utterance). We speculate that the network learns to assign different weights to the start and end of the utterance to account for the initial noise.

However, we find that in both matched and mismatched conditions, it is hard to handle

utterances contaminated with reverberation, a common use case, even when the training set is augmented with other types of noise. We find that this improvement in performance is even more reduced when using the wav2vec model, alluding to the model's fragility towards data integrity/structural changes. This can be because reverberation adds a continuous human speech signal in the background delayed by a small period of time. None of the other kind of noise types have speech in them, and hence augmentation doesn't aid the model to learn robustness to this kind of noise.

Finally, we investigate the differences in model performance when we use all types of noise vs. those that are perception-retaining. Specifically, we focus on the perception-altering noises because samples augmented with noises in this category no longer have a known ground truth. We inquire as to whether the use of samples that alter perception may lead to the appearance of model performance improvement (note: appearance because the samples now have uncertain ground truth). To maintain equivalence, we ensure that the training and validation dataset sizes are equal even when they are augmented with more noise conditions. We observe that many cases of performance improvement occur when the noises include those that are perception-altering (see "Al noises" in Table 5.2). We observe a difference of 12% to 25% between the numbers that we obtain for the perception-retaining noises vs. when not distinguishing between the two noise categories. This supports our claim that the choice between types of noises used for data augmentation during model training and performance evaluation affects the empirical observations and should be carefully considered. We hypothesize that this improvement in performance may be due to the inherent nature of noises that change emotion perception, if they are perceptible enough to change emotion perception, then they may stand out enough that the model can adequately learn to separate them out and improve its prediction towards the original ground truth annotation. However, if the noise alteration truly does change perception, then the model is learning to ignore this natural human perceptual phenomenon. This may have negative consequences during model deployment.

75

5.7.4 RQ3b: Does sample denoising help improve the robustness of emotion recognition models to unseen noise?

In the *matched training testing condition*, we find that the traditional deep learning model has an average performance of 0.57 across all the datasets and testing setups, while the end-to-end models do substantially better at 0.61 UAR. See Table 5.3 for details.

In the *mismatched training testing condition*, we find that for the traditional deep learning model, adding a denoising component to the models leads to a significant improvement when the original SNR is high (e.g., after continuous noise introduction the SNR decreases only by 10dB). In this case, we see an average improvement of $23\% \pm 3\%$ across all environmental noise categories, compared to when there is no denoising or augmentation performed.

However, when the SNR decreases by 20dB, we observe a decline in performance when using the noise suppression algorithms. We believe that this decline in performance is reflective of the mismatch in goals: the goal of noise suppression is to maintain, or improve, the comprehensibility of the speech itself, not necessarily highlight emotion. As a result, it may end up masking emotional information, as discussed in [145].

We further show that the addition of a denoising component does not significantly improve performance in the presence of signal distortion noise (other than reverberation) as compared to the presence of environmental noise (noise addition). For example, when samples were faded in or out or segments were sped up, the performance is significantly lower (-28%) than when tested on a clean test set. However, we did see an improvement in the performance for *unseen* reverberation contaminated samples as compared to data augmentation (an average of $+36\%$). Finally, we observe a general trend of increase in emotion recognition performance for the combined dataset (noisy and non-noisy samples), as compared to when the model is trained on the clean training set, which supports the findings from previous dataset augmentation research [10].

For the end-to-end deep learning model, we use the noise-robust version. We find that the model is effective at countering environmental noise when trained on a dataset

augmented with environmental noise, even in the mismatched condition. The performance is equivalent to the model evaluated on the clean data. We further delve into the amount of noise augmentation needed to achieve this equivalency. We consider all of the original training data. We augment a percentage of the training data, starting by augmenting a random sample of 10% with perception-retaining noise and increasing by 10% each time. We find that we obtain equivalency after augmenting with only 30% of the training data. We compare this compares to the traditional model, in which all of the training data are noise-augmented and we still do not see equivalency.

We separately consider the signal distortion noise samples. These were not part of the training of the wav2vec2-Large-robust model. However, this model only sees a 6% loss in performance, where the traditional robust model saw a 20% loss in performance.

However, as discussed in the original traditional model, the end-to-end noise-robust model also fails on reverberation-based contamination even when trained on a similarly augmented dataset (note that the denoised traditional model could effectively handle reverberation). We believe that this may be because the wav2vec model is trained on continuous speech and relies on the underlying linguistic structure of speech. However, in reverberation, there is an implicit doubling of the underlying information, which is at odds with how this model was trained. This may explain why it is not able to compensate for this type of signal manipulation.

Next, we analyze whether the perception category of noise used for data augmentation of the samples, in both, the train and test dataset influences the reported results for noise-robust model improvement. We find that there is a significant difference in performance when the testing dataset is augmented with any kind of noise vs. when augmented with perception-retaining noise. Specifically, we observe that the maximal gains in performance when testing on matched noisy conditions are for samples for which we do not know whether or not the ground truth holds (i.e., both noise categories). For example, when using the noise robust traditional deep learning model, where the test and train dataset is augmented with any type

of noises, we observe a performance improvement, as compared to that using a clean train dataset, of 12%. Similarly for noise robust end to end models, the performance improvement difference when using all noises vs. only perception retaining ones is 15% for activation and 13% for valence. Again, this is a problem when we think about deploying models in the real world because although the perception of these emotion expressions may change, we are assuming that the system should think of these perception labels as rigid and unchanging.

5.7.5 RQ4: How does the robustness of a model to attacks compare when we are using test samples that with are augmented with perception-retaining noise vs. samples that are augmented with all types of noise, regardless of their effect on perception?

In this section, we aim to show the effects of noise augmentation in general and specifically highlight noise categories that do not alter human perception. We will show that if we are not careful with the selection of our noise types, moving from noise that we know not to alter perception to noise that may, the resulting noise sources can not only impact the brittleness of models, but also lead to inaccurate evaluation metrics. We also specifically report robustness performance when using reverberation-based contamination, as we observed that it is the most likely noise category to degrade the robustness of the model.

We allow the decision boundary attack algorithm a maximum of five queries to create a noise augmented sample that will change the model output. We find that if the attacker is also given perception-altering noises, compared to perception-retaining, it can more effectively corrupt the sample. It achieves an increase in success rate from 35% (only perception-retaining) to 48.5% (all noise categories). This increase in the success rate when perception-altering noises are included implies that the model does not remain robust when the effects of noise on human perception are not considered.

We next consider the type of noise (environmental vs. signal manipulation). We find that the success rate of flipping a model's output is 18% for noises belonging to the environmental

category, which is generally a category of perception-retaining noise. The success rate of flipping a model's output is 37% for all noises belonging to the signal manipulation category. When we constrain our possible noise choices to perception-retaining signal manipulations, we see that the success rate of the intentional misclassification algorithm drops to 24%. On the other hand, we observe that when we also consider the signal manipulations that are perception-altering, the success rate of flipping a model output is 39%. See Table 5.4 for more details.

We previously discussed the fragility of end-to-end models towards reverberation-based noise contamination, noise that is perception-retaining for human evaluators. Here, we specifically run an experiment to use only that particular noise category for the model fragility testing. If the attacker knows that the model is susceptible to reverberation-based prediction changes, the intentional misclassification algorithm can land on an optimal set of room and reverberation parameters in a maximum of five queries to be able to produce a flipped output for that particular sample. It achieves a success rate of 24%, compared to 12% for other perception-retaining noises. The traditional noise-robust deep learning model is even more challenged, compared to the end-to-end model. The number of queries required to flip the output is three, vs. five for the end-to-end model, suggesting that it is less robust. This empirical evaluation is performed primarily to demonstrate how such noise inclusions can not only invalidate the ground truth but also lead to inaccurate and fragile benchmarking and evaluation of adversarial efficiency and robustness.

5.7.6   RQ5: What are the recommended practices for speech emotion dataset augmentation and model deployment?

We propose a set of recommendations, for both augmentation and deployment of emotion recognition models in the wild, that are grounded in human perception. For augmentation, we suggest that:

1. Environmental noise should be added to datasets to improve generalizability to varied

noise conditions, whether using denoising, augmentation, or a combination of both.

2. It is good to augment datasets by fading the loudness of the segments, dropping letters or words, and speeding up small (no more than 25% of the total sample length) segments of the complete sound samples in the dataset. But it is important to note that these augmented samples should not be passed through the denoising component as the denoised version loses emotion information.

3. One should not change the speed of the entire utterance more than 5% and should not add intentional pauses or any background noises that elicit emotion behavior, e.g., sobs or laughter.

Regarding deployment, we suggest that:

1. Noisy starts and ends of utterances can be handled by augmentation, hence, if the training set included these augmentations, there should be no issue for deployed emotion recognition systems.

2. Reverberation is hard to handle for even augmented emotion recognition models. Hence, the samples must either be cleaned to remove the reverberation effect, or must be identified as low confidence for classification.

3. Deploy complementary models that identify the presence of noise that would change a human's perception.

5.8   Conclusion

In this work, we study how the presence of real world noise, environmental or signal distortion, affects human emotion perception. We identify noise sources that do not affect human perception, such that they can be confidently used for data augmentation. We look at the change in performance of the models that are trained on the original IEMOCAP

dataset, but tested on noisy samples and if augmentation of the training set leads to an improvement in performance. We conclude that, unlike humans, machine learning models are extremely brittle to the introduction of many kinds of noise. While the performance of the machine learning model on noisy samples is aided from augmentation, the performance is still significantly lower when the noise in the train and test environments does not match. In this chapter, we demonstrate fragility of the emotion recognition systems and valid methods to augment the datasets, which is a critical concern in real world deployment.

CHAPTER VI

Context in Crowdsourcing Emotion Annotations

6.1 Motivation and Contributions

Emotion recognition algorithms rely on data annotated with high quality labels. However, emotion expression and perception are inherently subjective. There is generally not a single annotation that can be unambiguously declared "correct". As a result, annotations are colored by the manner in which they were collected. In this chapter, we conduct crowdsourcing experiments to investigate this impact on both the annotations themselves and on the performance of these algorithms. We focus on one critical question: the effect of context. We present a new emotion dataset, Multimodal Stressed Emotion (MuSE), and annotate the dataset using two conditions: randomized, in which annotators are presented with clips in random order, and contextualized, in which annotators are presented with clips in order. We find that contextual labeling schemes result in annotations that are more similar to a speaker's own self-reported labels and that labels generated from randomized schemes are most easily predictable by automated systems.

6.2 Introduction

Emotion technologies, both recognition and synthesis, are heavily dependent on having reliably annotated emotional data, annotations that describe the observed emotional display.

The hope is often that these annotations capture the speaker's true underlying state. Yet, in practice, this true *felt sense* emotion is unknown, and researchers must resort to manual labeling of data. The hope is that these manual labels are sufficiently "correct" to enable the training and evaluation of emotion technologies. One method of ensuring quality labels has been to require the participation of expert raters. However, it can be both expensive and time consuming to hire expert raters. More recently, researchers have embraced crowdsourcing services (e.g., *Amazon Mechanical Turk*) to efficiently collect annotations from non-expert workers in a cost-effective and timely manner [201]. Once collected, annotations from non-expert workers are aggregated to form ground-truth labels that are used for training and evaluating automated systems. However, the method through which these annotations are collected can profoundly impact the behavior of the annotators. In this chapter, we study how the setup of a crowdsourcing task can influence both the collected emotion labels as well as the performance of classifiers trained using these labels.

The effective use of crowdsourcing for collecting reliable emotion labels has been an active research topic. Burmania et al. investigated the trade-off between the number of annotators and underlying reliability of the annotations [33]. Other work has looked at quality-control techniques to improve the reliability of annotations. For example, Soleymani et al. used qualification tests to filter out spammers and retain high-quality annotators [201]. Burmania et al. investigated the use of gold-standard samples to monitor annotators' reliability and fatigue [35].

However, variability also results from context, relevant past information that provides cues as to how to interpret an emotional display. Context, such as tone, words, expressions can affect how individuals perceive emotion [129]. Context is also implicitly included in the labeling schemes of many of the most common emotion datasets (e.g., IEMOCAP [36] and MSP-Improv [38]) because annotators rate each utterance (or time period) in order. That means that annotators are influenced by information that they recently observed [233]. However, emotion recognition systems are often trained over single utterances [10, 3, 157,

190], leading to a mismatch in the information available to annotators and to classification systems.

In this work, we study the difference between annotations obtained for audio clips when emotional displays are presented to annotators with context and when presented randomly. In both cases, annotators are affected by the emotion displays that they have recently observed [180, 188]. However, only in the contextual presentation there is also a cohesive story. We investigate the following research questions:

- Q1: Is there a significant difference between annotations obtained from random and contextual presentations?

- Q2: Are annotations obtained from contextual presentations more similar to a speaker's own self-reported labels than those from random presentations?

- Q3: Is there a significant difference between the inter-rater agreements obtained from random and contextual presentations?

- Q4: How does the performance of an emotion recognition system, operating on single utterances, vary given annotations obtained from random and contextual presentations?

- Q5: How does the performance gain of an emotion recognition system operating across multiple utterances vary given different amounts of context (defined as number of prior utterances) and labels obtained from random and contextual presentations?

The findings from this work will provide insight into performance implications of emotion recognition system given mismatches between the amount of context provided to the annotators generating the labels and the ultimate classification system.

The main aim of the work was to annotate the dataset described in the previous chapter. For all further experiments, we use the utterances created from the MuSE dataset.

6.2.1 Crowdsourcing

We posted our experiments as Human Intelligence Tasks (HITs) on *Amazon Mechanical Turk*. HITs were defined as sets of utterances in either the contextual or random presentation condition. In each condition, workers were presented with a single utterances and were asked to annotate the activation and valence values of that utterance using Self Assessment Manikins [30]. Once completed, the worker was presented with a new HIT and could not go back to revise a previous estimate of emotion. This annotation strategy is different than the one deployed in [48],where the workers could go back and re-evaluate utterances.

In the randomized experiment, each HIT is an utterance from any section, by any speaker, from any session and all HITs appear in random order. So, a worker might see the first HIT as *Utterance 10 from Section 3 of Subject 4's stressed recording* and see the second HIT as *Utterance 1 from Section 5 of Subject 10's non-stressed recording*. This setup ensured that the workers couldn't condition to any speaker's specific style or contextual information.

In the contextual experiment, we posted each HIT as a collection of ordered utterances from a section of a particular subject's recording. Because each section's question was designed to elicit a particular emotion, we still posted the HITs in a random order over sections from all subjects. This prevented workers from conditioning to the speaking style of an individual participant. For example, a worker might see the first HIT as *Utterance 1...N from Section 3 of Subject 4's stressed recording* and see the second HIT as *Utterance 1...M from Section 5 of Subject 10's non-stressed recording* where *N, M* are the number of utterances in those sections respectively.

We recruited from a population of workers in the United States who are native English speakers, to reduce the impact of cultural variability. We ensured that each worker had $> 98\%$ approval rating and number of HITs approved as > 500. We ensured that all workers understood the meaning of activation and valence using a qualification task that asked workers to rank emotion content. The workers were asked to select, given two clips, which clip had the higher valence and which had the higher activation. The options were chosen

from a set including: (1) a speaker in low activation, high valence state and (2) a speaker in high activation, low valence state.

We assigned each HIT to eight workers. All HIT workers were paid a minimum wage (\$9.25/hr), pro-rated to the minute. We removed and re-posted assignments where the worker completed the assignment in time shorter than the audio length. The ground-truth for each utterance was formed by taking the average of the eight annotations. Table 4.3 shows the data summary for the collected annotations and corresponding data points.

6.3   Experimental Setup

Acoustic Features.   We extract acoustic features using OpenSmile [71] with the eGeMAPS configuration [70]. The eGeMAPS feature set consists of 88 utterance-level statistics over the low-level descriptors of frequency, energy, spectral, and cepstral parameters. We perform speaker-level z-normalization on all features.

Static Network Setup (Hypothesis 4).   We train and evaluate four Deep Neural Networks (DNN) models: {random, contextual}$\times$ {valence, activation}. In all cases, we predict the continuous annotation using regression. For each network setup, we follow a five-fold evaluation scheme and report the average RMSE across the folds. For each test-fold, we use the previous fold for hyper-parameter selection and early stopping. The hyper-parameters include: number of layers $\{2, 3, 4\}$ and layer width $\{64, 128, 256\}$. We use ReLU activation and train the networks with MSE loss using Adam optimizer.

Dynamic Network Setup (Hypothesis 5).   We use Gated Recurrent Unit networks (GRU). The hyper-parameters are: number of layers $\{1, 2\}$ and layer width $\{64, 128, 256\}$. We pass the GRU output of the last time step through a regression layer to get the final outputs. We train the networks with MSE loss using Adam optimizer.

Network Training.   We train our networks for a maximum of 100 epochs and monitor the validation loss after each epoch. We stop the training if the validation loss does not improve for 15 consecutive epochs. We revert the network's weights to those that achieved

Table 6.1: Mean activation and valence values obtained from the two crowdsourcing labeling schemes (random and context) grouped by speaker condition (stress and non-stress).

	Activation		Valence	
	Random	Context	Random	Context
Stress	3.63	3.59	**5.27**	**5.36**
Non-Stress	**3.61**	**3.79**	**5.26**	**5.39**

the lowest validation loss during training. Finally we train each network five times and average the predictions to reduce variance due to random initialization.

6.4 Results and Analysis

6.4.1 RQ1: Differences in Obtained Annotations

Hypothesis: *Human annotations collected through randomized labeling are significantly different from those collected through contextualized labeling.* Prior work has shown context effects emotion perception [233], even when observers are explicitly asked not to take it under consideration [165, 43]. Hence, we believe that context provided by previous utterances would lead to a change in perception of a particular utterance. Tables 6.1 and 6.2 (sets of significantly different means are bolded (t-test, $p < 0.01$)) show the mean activation and valence, for the random and contextualized labeling schemes, grouped by condition and question, respectively. Table 6.1 shows that, for non-stress conditions, the mean of the activation ratings obtained through contextual labeling is significantly higher than that obtained through random labeling. The table also shows that, for both stress and non-stress conditions, the valence means obtained through contextual labeling are significantly higher than those obtained through random labeling. Table 6.2 shows that, although the mean valence and activation values were consistently different for the labelling schemes across all emotion elicitation techniques, the differences were significant in some elicitation techniques and not in others.

Table 6.2: Mean activation and valence values obtained from the two crowdsourcing labeling schemes (random and context) grouped by emotion elicitation question.

	Activation		Valence	
	Random	Context	Random	Context
Icebreaker	3.55	3.60	**5.41**	**5.61**
Positive	3.64	3.71	5.11	5.13
Negative	**3.57**	**3.67**	**5.40**	**5.55**
Intensity	**3.64**	**3.74**	5.17	5.31
Ending	3.69	3.71	**5.23**	**5.29**

6.4.2 RQ2: Self-Annotations and Crowdsourced Annotations

Hypothesis: *Annotations of outside observers are more similar to self-annotations in the contextual case, compared to the randomized case.* Path models [17] suggest that subjective voice variation, from the established mental baseline accounts for much of the variance in emotion inference. Hence, emotion inference is aided with more cues about the speech patterns that are more readily provided through context. Figure 6.1 shows the absolute differences between the mean crowdsourced labels (valence and activation, each for random and contextual schemes) and self-reported scores as a function of utterance position. The figure shows that contextual labels have consistently lower absolute differences, compared to self-reported labels, than the random labels. A paired t-test shows that these differences between the contextual and random labels are significant ($p < 0.01$) for both valence and activation.

Our results suggest that crowdsourced emotion labels collected with access to contextual information are closer to self-reported emotion labels. Our results further suggest that these differences are consistent across recording conditions (Table 6.3) and emotion elicitation questions (Table 6.4, sets of significantly different means are bolded, t-test, $p < 0.01$).

6.4.3 RQ3: Inter-Annotator Agreement

Hypothesis: *Individual annotators differ in annotation similarity in the contextual*

Table 6.3: Mean difference between the self-reported activation and valence ratings from the two labeling schemes (random and context) grouped by speaker condition (stress and non-stress).

	Activation		Valence	
	Random	Context	Random	Context
Stress	**2.03**	**1.96**	**1.20**	**1.14**
Non-Stress	**1.82**	**1.67**	**1.20**	**1.12**

Table 6.4: Mean difference between the self-reported activation and valence ratings from the two labeling schemes (random and context) grouped by emotion elicitation question.

	Activation		Valence	
	Random	Context	Random	Context
Icebreaker	1.81	1.80	**0.97**	**0.85**
Positive	**1.89**	**1.74**	1.14	1.11
Negative	**1.96**	**1.76**	**1.18**	**1.07**
Intensity	**2.19**	**2.08**	1.49	1.44
Ending	**1.81**	**1.73**	1.23	1.28

presentations, compared to the randomized presentation. Joseph et al. in [112] show that while insufficient context results in noisy and uncertain annotations, an overabundance of context may cause the context to outweigh other signals and lead to lower agreement. Further, contextual information biases different people differently on both temporal and intensity metrics [219, 223]. Our results highlight the impact of context: the agreement is significantly higher in the case of labels obtained from the randomized presentations, compared to the contextualized presentations: (1.55 vs. 1.62) for activation and (1.07 vs. 1.14) for valence. This trend holds true for all experimental design setups i.e. {random, contextual}× {valence, activation} and {random, contextual}× {icebreaker, positive, negative, intensity and ending}. As shown in Tables 6.1 and 6.2, the labels obtained in both cases are significantly different due to context-based conditioning. However, the conditioning may not impact the labels consistently across all workers, which may lead to lower inter-annotator agreement values. This suggests that it may be beneficial to consider the distribution of annotations as ground-truth, rather than averaging labels, which presumes that the impact of

conditioning is consistent across all workers [239].

6.4.4   RQ4: Non-Contextual Annotations and Static Classifiers

Hypothesis: *A static classifier will perform better when trained and evaluated using labels annotated with a randomized presentation, compared to a contextualized presentation.* Prior studies have shown that it is easier to classify data with less target variation [141] and matched classifier input, which in our case is labels obtained from the random labelling presentation (the classifier processes single utterances at a time, no context).

We test this hypothesis by training and evaluating classifiers for the four possible setups: $\{random, contextual\}$ x $\{valence, activation\}$. The classifier is described in Section 6.3. We find that the RMSEs are lower for the contextual labels in the case of activation (0.91 vs. 1.00) while the errors are lower for the random labels in the case of valence (1.13 vs. 1.20). Using a paired t-test, we find that the differences in errors are significant in the case of valence but not activation. These findings suggest that classification performance is impacted by the labelling methodology, but that this effect may depend on emotion dimension.

Prior work has demonstrated the importance of considering long-term context when predicting valence (the same effect has not been shown in activation) [116]. The contextual annotations provided the annotators with this information, but the classifier could not take advantage of this effect. This mismatch may have contributed to the relatively lowered performance of the valence classifier, compared to the activation classifier.

6.4.5   RQ5: Contextualized Annotations and Long-Context Classifiers

Hypothesis: *We anticipate that systems trained on contextualized labels will see greater increases in performance as the amount of provided context increases.* This finding would support results in the literature regarding the ordinal nature of emotion perception [233] and previous works in emotion recognition that have demonstrated that context can influence the performance of emotion classifiers [116].

90

Figure 6.1: Mean difference between the self-reported activation and valence ratings and the random and contextual presentations.
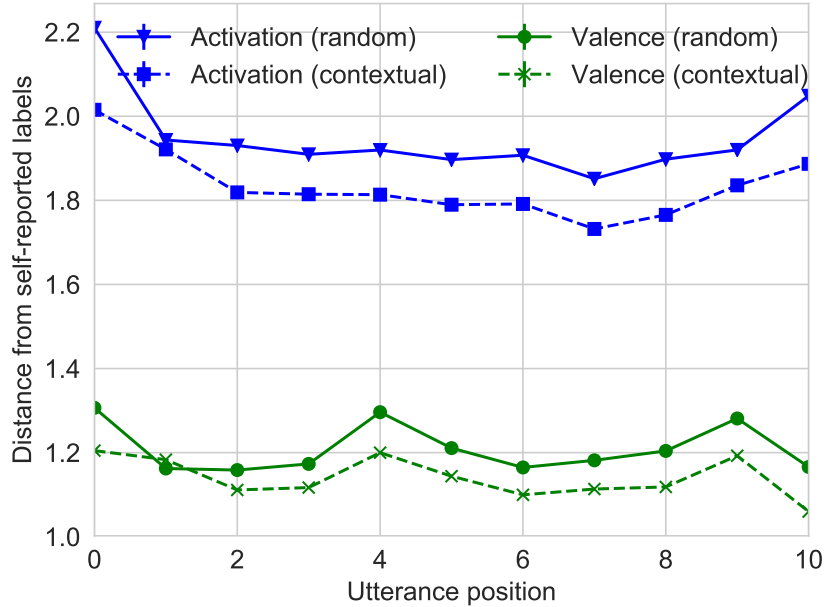
The classifier is described in Section 6.3. We test this hypothesis by using the contextual annotations in one classifier and the non-contextual (random) annotations for the other classifier. We select a subset of utterances in each section that have at least five consecutive utterances before them (59% of the original data). The initial classifier is trained without temporal context (but with the contextualized labels). We incrementally increase the number of past utterances (from zero to five). We run this for every task combination and report the results in Table 6.5.

Table 6.5 shows the performance gains after incrementally adding the past utterance, relative to the baseline performance. The addition of past utterances improves the performance over baseline for all setups. Where using contextual labels, however, the performance gains are generally higher than the gains obtained after using random labels. Our results suggest that it is necessary to consider the mismatch the amount of context provided to the annotators generating the labels and the ultimate classification system.

Table 6.5: Relative improvement in RMSE (%) obtained for each additional previous utterance, comparing random and contextual labels.

	Activation		Valence	
Past steps	Random	Context	Random	Context
0	-	-	-	-
1	+1.96%	+1.24%	+0.85%	+3.32%
2	+2.28%	+2.93%	+5.23%	+7.63%
3	+3.36%	+8.72%	+6.08%	+8.43%
4	+4.41%	+10.5%	+8.23%	+8.36%

6.5 Conclusion

In this work we showed that the amount of context provided to annotators when assigning emotion labels affects both the annotations themselves and the performance of classifiers using these annotations. We also studied the implications of a mismatch between annotation context and classifier context on classifier performance.

Controlling for Confounders in Emotion Recognition

7.1 Motivation and Contributions

Various psychological factors affect how individuals express emotions. Yet, when we collect data intended for use in building emotion recognition systems, we often try to do so by creating paradigms that are designed just with a focus on eliciting emotional behavior. Algorithms trained with these types of data are unlikely to function outside of controlled environments because our emotions naturally change as a function of these other factors. In this work, we study how the multimodal expressions of emotion change when an individual is under varying levels of stress. We hypothesize that stress produces modulations that can hide the true underlying emotions of individuals and that we can make emotion recognition algorithms more generalizable by controlling for variations in stress. To this end, we use adversarial networks to decorrelate stress modulations from emotion representations. We study how stress alters acoustic and lexical emotional predictions, paying special attention to how modulations due to stress affect the transferability of learned emotion recognition models across domains. Our results show that stress is indeed encoded in trained emotion classifiers and that this encoding varies across levels of emotions and across the lexical and acoustic modalities. Our results also show that emotion recognition models that control for stress during training have better generalizability when applied to new domains, compared to models that do not control for stress during training. We conclude that is is necessary to

consider the effect of extraneous psychological factors when building and testing emotion recognition models.

7.2 Introduction

Many extraneous psychological factors influence how individuals express and perceive emotions [171]. However, most emotion recognition algorithms, rely on data collected in controlled laboratory environments (e.g., [36, 38]) where influences from such factors are either not present, or kept constant. The performance of emotion recognition algorithms is likely to vary when applied to data where these external psychological factors are present. In this work, we study how an extraneous psychological factor, stress, affects multimodal (acoustic+lexical) emotion classifiers. Stress can affect how individuals produce and perceive emotion [171]. Yet, the effect of stress levels on the performance of state-of-the-art emotion recognition systems has not been explored.

Extraneous psychological factors can act as confounding factors, variables that influence both the output (e.g., emotion) and the input (e.g., acoustic and lexical features). Not controlling for confounding variables when training emotion classifiers can cause the classifiers to learn unintentional associations between the variables, associations that might not replicate in real world scenarios. For instance, consider a dataset where all the "sad" samples were unintentionally recorded from individuals who were experiencing stress at the time of recording. Not taking special care when building the models could cause a trained classifier to erroneously associate experiencing stress with being sad. In this work, we study how stress alters the performance of trained emotional classifiers in the context of neural networks. We then see how performance is affected when tested on samples out of domain, when we explicitly impede the network from learning such associations.

Previous research showed that controlling for confounding variables when training emotion recognition classifiers results in more robust models when compared to models trained without controlling for the same confounding variables. For instance, Abdelwahab

et al. [4] and Gideon et al. [83] showed that controlling for domain (i.e., data source), as a confounding factor, when training emotion recognition models results in improved cross-corpus generalization performance when compared to performance of models that were trained without controlling for domain as a confounding factor. Most of the above mentioned methods rely on samples obtained from the target domain to extract representations that retain information only about emotion and not domains. Our goal is to go beyond studying the effects of variations due to domain and background noise on the robustness of trained emotion recognition models, and instead focus on how stress affects the learned acoustic and lexical emotional representations. Unlike the commonly used methods for learning domain invariance, we aim to accomplish generalizing person specific behavior by proactively "unlearning" the modulations due to the presence of stress while still retaining emotion information in representations.

In particular, we seek to answer the following questions:

1. Can we recognize stress given representations trained solely for recognizing emotion? Is the stress recognition performance similar across the lexical and acoustic modalities?

2. Can we completely decorrelate emotion representations from stress representations? If so, how does this decorrelation impact the performance of emotion classifiers?

3. Does the impact of decorrelation on the performance of emotion classifiers vary given different levels of stress?

4. Does decorrelating these representations (i.e., emotion and stress) aid in model generalizability?

5. Can we proactively remove other types of confounders (e.g., spontaneity) to improve cross-dataset performance?

6. Are there identifiable lexical patterns in samples that are especially successfully classified by the adversarially trained model for emotion classification?
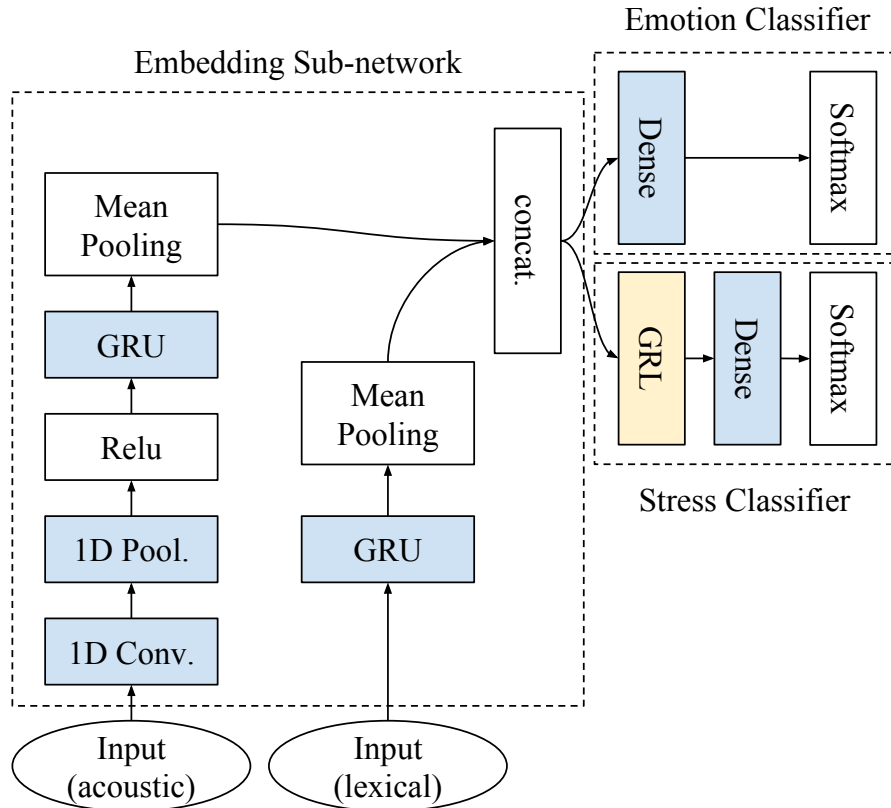
Figure 7.1: Adversarial multi-task network architecture.

To the best of our knowledge, this is the first work that studies the interplay between emotion and stress in the context of automatic emotion recognition and representation learning.

7.3 Experimental Setup

We use three datasets to study the effect of stress on emotion recognition: (1) Multimodal Stressed Emotion (MuSE) dataset [105]; (2) Interactive Emotional Dyadic MOtion Capture (IEMOCAP) dataset [36]; and (3) MSP-Improv dataset [38]. We use the acoustic and lexical features, MFBs and word2vec respectively, as defined in Section 3.3. We then describe the network architecture and the training recipe of the two emotion recognition models, one that controls for stress as a confound and one that does not.

7.3.1 Architecture

The network consists of three components (Figure 7.1): (1) embedding sub-network; (2) emotion classifier; and (3) stress classifier. The embedding sub-network induces fixed-size representations given the acoustic and lexical input streams. In Figure 7.1, the concatenation layer of acoustic and lexical stream shows the induced fixed-size representations. The emotion and stress classifiers perform their respective classification tasks given the fixed-size representations from the embedding sub-network. We use two variants of the embedding sub-network in this work: a unimodal and a multimodal variant. The unimodal embedding sub-network takes a single stream (acoustic or lexical) input while the multimodal embedding sub-network takes a two-stream (acoustic and lexical) input. The objective of the adversarial multi-task system is to maximize the performance of the emotion classifier while minimizing the performance of the stress classifier.

Stress-Invariance. The network is trained to unlearn stress. We achieve this goal using a Gradient Reversal Layer (GRL) [77]. The use of GRLs is a common approach that can be used to train networks that are invariant to specific properties [153, 196, 4, 79, 150, 67]. During the backward pass of the training phase, the GRL multiplies the backpropagated gradients by $-\lambda$. During the forward pass, the GRL acts as an identity function. To make the network invariant to stress, we place the GRL between the embedding sub-network and the stress classifier as shown in Figure 7.1.

Model Variations. We use 12 variants of the network shown in Figure 7.1 with the following combinations: {*normal-classification, adversarial-classification*} × {*activation, valence*} × {*uni-lexical, uni-acoustic, multimodal*}. The normal classification setup consists of the embedding sub-network (lexical, acoustic, or multimodal) and the emotion classifier (activation or valence) parts of the model. The adversarial classification setup adds the adversarial stress classifier.

7.3.2 Training

We implement models using the Keras library [52]. We use a weighted cross-entropy loss function for each task and learn the model parameters using the RMSProp optimizer [214]. We train our networks for a maximum of 50 epochs and monitor the validation loss from the emotion classifier after each epoch, stopping the training if the validation loss does not improve after five consecutive epochs. Once the training process ends, we revert the network's weights to those that achieved the lowest validation loss on the emotion classification task. For the adversarial classification model, we ensure that the chosen model yields a validation unweighted average recall (UAR) that is random (0.33) for the stress classification task. Finally, we train each setup three times with different random seeds and average the predictions over these runs to reduce variations due to random initialization.

We use validation samples for hyper-parameter selection and early stopping. The hyper-parameters that we use in our search include: number of convolutional layers $\{3, 4\}$, number of convolutional kernels $\{2, 3\}$, conv. layers width $\{32, 64, 128\}$, 1D max-pooling kernel width $\{2\}$, number of GRU layers $\{2, 3\}$, GRU layers width $\{32\}$, number of dense layers $\{1, 2\}$, dense layers width $\{32, 64\}$, GRL λ $\{0.3, 0.6, 0.8\}$. For the adversarial emotion classification setups, we use the hyper-parameters that maximize the validation emotion classification performance while minimizing the validation stress classification performance. We assess performance using UAR, given the imbalanced nature of our data [184].

7.4 Analysis

7.4.1 RQ1: Implicit Stress Encoding

QUESTION: *Can we recognize stress given representations trained solely for recognizing emotion?*

HYPOTHESIS: *We expect the performance of detecting stress from representations obtained by training emotion classifiers to vary based on the modality, and the emotion being modeled.*

Table 7.1: UAR (chance = 0.333) for predicting activation (top) and valence (bottom) in adversarial and non-adversarial (normal) setups. Bold signifies significantly different performance (paired *t*-test, $\alpha < 0.05$).

	Normal		Adversarial	
Setup	Act.	Stress	Act.	Stress
Unimodal (A)	**0.611**	0.412	**0.572**	0.305
Unimodal (L)	0.550	0.394	0.527	0.332
Multimodal (A+L)	**0.659**	0.425	**0.613**	0.322

(a) Activation

	Normal		Adversarial	
Setup	Val.	Stress	Val.	Stress
Unimodal (A)	0.460	0.396	0.431	0.332
Unimodal (L)	0.685	0.353	0.674	0.323
Multimodal (A+L)	**0.666**	0.397	**0.641**	0.328

(b) Valence

Stress has been shown to have varying effects on both the linguistic [32] and para-linguistic [171, 205] components of communication. Previous work has also demonstrated that the lexical part of speech carries more information about valence while the para-linguistic part carries more information about activation [116]. As a result, we expect the performance of stress classification to vary based on modality, and emotion dimension being modeled.

To test our hypothesis, we train the 12 model variants described in section 7.3.1 with five-fold speaker-independent cross-validation. We report the average across the five folds for the normal classification and the adversarial classification setups in Tables 7.1a and 7.1b for predicting activation and valence, respectively. Our results show that a network trained to only recognize emotion is generally discriminative for stress. For instance, we obtain a maximum UAR of 0.425 when using a multimodal network that was trained to only detect activation; and a maximum UAR of 0.397 when using multimodal network that was trained to only detect valence.

Our results in Table 7.1a suggest that the acoustic modality encodes information that is relevant for recognizing stress and activation. In contrast, the results show that the

Table 7.2: Confusion matrices for predicting activation (top) and valence (bottom) showing percentage change in classification performance of the multimodal setup for each emotion class after controlling for stress.

Act.	Low Stress			High Stress		
	(0)	(1)	(2)	(0)	(1)	(2)
(0)	−1.21	−0.22	+1.44	+1.30	−2.11	+7.35
(1)	+4.22	−2.73	+1.31	+18.40	−22.03	+14.80
(2)	−2.01	+6.66	−6.38	−8.81	+6.21	−3.14

(a) Activation

Val.	Low Stress			High Stress		
	(0)	(1)	(2)	(0)	(1)	(2)
(0)	−1.66	+1.12	+1.11	−8.22	+7.88	+1.33
(1)	−0.76	−2.22	+0.87	−2.31	−6.11	+4.26
(2)	−1.01	+0.31	+1.45	−1.15	−2.10	+0.79

(b) Valence

representations trained on lexical modality encode information that is relevant for detecting valence but not for stress. Our findings are consistent with previous research that demonstrated that stress is encoded in acoustic features [32, 75, 62].

7.4.2   RQ2: Decorrelate Emotion and Stress Representations

QUESTION: *Can we decorrelate emotion representations from stress representations? How does it impact performance of emotion classifiers?*

HYPOTHESIS: *Decorrelating the stress and emotion representations will cause a decrease in emotion classification performance on the source domain.*

Previous research demonstrated that controlling for confounders during the training process can cause the performance of the main task on the same dataset to decrease [149, 209]. For instance, Zhang et al. [149] showed that the performance of detecting sarcasm decreases when controlling for publication as a confounding variable in the training process, but the prediction accuracy increases on an unseen publication set. Similarly, Ganin et al. [79]

showed that controlling for domain while training a network for detecting sentiment can result in a performance reduction on the main task. The reduction in performance on the main task, after controlling for an extraneous confounding variable, can be attributed to the removal of information that the model can use as a "shortcut" for achieving the main task.

Our results show that (Tables 7.1a and 7.1b):

- Activation classification performance decreases given adversarial training. This decrease is statistically significant for the acoustic (6.382% drop in UAR) and multimodal (6.980% drop in UAR) setups.

- Valence classification performance decreases given adversarial training. This decrease is statistically significant for the multimodal setup (3.754% drop).

The reduction in performance in the main task after controlling for a confounding variable can also be caused by the removal of information that is equally beneficial for both detecting stress and detecting emotion. Our results in further sections , however, show that models that control for stress are better able to recognize emotion in new domains, compared to models that do not control for stress. This suggests that the process of "unlearning" stress does not come at the expense of the primary task of emotion detection.

7.4.3   RQ3: Impact of Stress Levels on Emotion Recognition

QUESTION: *Does the impact on the performance of emotion classifiers vary given different levels of stress*

HYPOTHESIS: *The valence and activation emotion classes (low, medium, and high) are impacted differently by stress.*

Prior research demonstrated that emotions produced by stressed individuals are not recognized in the same way as those by non-stressed individuals [171]. In particular, researchers found that speech patterns of negative emotions produced by stressed individuals are more difficult to recognize than negative emotions produced by non-stressed individuals [171]. We expect

similar patterns to hold in automatic emotion recognition systems. That is, we expect the presence of stress to have a varying effect on the performance of the classifier depending on the emotion class (for valence and activation), and the amount of stress induced.

To test our hypothesis, we study how the performance of the classifier varies for each emotion class when we control for stress. We report the changes in performance, after controlling for stress, for each emotion class, grouped by stress level (low, high), in Table 7.2. The results in the table show:

- High levels of stress impact classification more strongly (3.89% and 2.41% drop in UAR for activation and valence, respectively) than low levels of stress do (1.44% and 0.31% drop in UAR for activation and valence, respectively). This is generally true for all emotion classes (valence and activation).

- High levels of stress have the biggest impact on mid level of activation predictions (22.03% drop in accuracy for detecting neutral activation).

- High levels of stress have the biggest impact on low valence predictions (8.22% drop for low valence).

The results show that stress level effects emotion recognition, for both activation and valence. This drop in performance can be attributed to changes in the perceived emotions by the annotators. Researchers have demonstrated that stressed sentences are usually perceived by annotators to be more neutral than they were originally intended to be [171].

7.4.4 RQ4: Decorrelation and Model Generalizability

QUESTION: *Does the process of decorrelating these representations (i.e., emotion and stress) aid in model generalizability?*

HYPOTHESIS: *Removing the confounding factor stress would aid in creating models that are more generalizable across datasets.*

Previous research has shown that laboratory collected datasets are too small and often fail

to capture the complete distribution of the domain [100, 142] present in the real world. These datasets are often plagued with unintentional correlational factors [142, 128]. Hence we believe that removing modulations due to stress should aid the generalizability of the model to datasets, where this psychological factor is either unmeasured or the distribution is non-uniform between training and testing.

To answer if the models trained on MuSE dataset generalize better, we perform two sets of experiments: (a) self generalizability in artificially partitioned datasets with different stress distributions for evidence of concept and (b) cross-dataset generalizability.

Artificially Segmented Within-Dataset Performance. We run the first set of experiments by creating partitions of data by stress level. We do this to create artificially mismatched environments between training and testing. We reserve one set to be test set (target), while keeping the other two for training and validation combined (source). This is in similar vein to partitioning created across confounding factor for UCI Bike Rentals Dataset in [209]. To ensure speaker independent sets, we divide the training set using an 80:20 split (train and validation), ensuring no speakers overlap. We run these models n times where n is the number of speakers in test data, that are also present in train/validation data. For each run, we remove one speaker from the train/validation data and test on that speaker. We calculate the average test performance of all these runs as the performance of the model for that setup.

We report our results in Table 7.3. When we consider low levels of stress as our target, we see that adversarial classification significantly improves performance over normal classification for multimodal setup for activation and for both, lexical and multimodal setup for valence. Considering mid levels of stress as our target, adversarial classification significantly improves performance over normal classification for all setups for activation and for both, acoustic and multimodal setup for valence. Subsequently considering high levels of stress as our target, adversarial classification significantly improves performance over normal classification for acoustic setups for activation and for all setups for valence.

Table 7.3: Performance (UAR) predicting activation (left) and valence (right) in non-adversarial and adversarial (for best *lambda* value) for self-partition on MuSE. Bold signifies significantly different performance (paired *t*-test, $\alpha < 0.05$).

	Normal		Adversarial	
	Act	Val	Act	Val
Train: Stress (Medium + High) Test: Stress (Low)				
Unimodal (A)	**0.623**	0.451	0.582	0.453
Unimodal (L)	0.561	0.654	0.548	**0.691**
Multimodal (A+L)	0.650	0.673	**0.662**	**0.703**
Train: Stress (Low + High) Test: Stress (Medium)				
Unimodal (A)	0.610	0.420	**0.628**	**0.432**
Unimodal (L)	0.520	0.672	**0.545**	0.669
Multimodal (A+L)	0.602	0.621	**0.638**	**0.649**
Train: Stress (Medium + High) Test: Stress (Low)				
Unimodal (A)	0.582	0.384	**0.605**	**0.411**
Unimodal (L)	0.540	0.630	0.513	**0.652**
Multimodal (A+L)	0.642	0.621	0.647	**0.637**

Cross-dataset Performance. Now that we have evidence for concept for artificially mismatched distributions that removing stress as a confounding factor can aid generalizability, we ask if adversarially removing encoded stress from emotion representation improves cross-dataset performance. We assume that we previously do not have any samples from the target dataset to train our model, to test generalizability at deployment. We train a dataset on complete MuSE data, keeping 20% of speaker independent data for validation of hyper-parameters. Then we use the trained model to test on IEMOCAP for a combination of acoustic, lexical and multimodal inputs and on MSP-Improv for acoustic inputs.

We report our results of comparing the performance of the adversarial and normal models (Table 7.4):

- Activation: There is a significant increase in performance in all setups when the adversarial classification model is tested on IEMOCAP. We observe a significant increase in performance in acoustic setup of adversarial classification model (0.404 vs

Table 7.4: Performance (in UAR) predicting activation (left) and valence (right) in non-adversarial and adversarial (for best *lambda* value) setups across datasets when trained on MuSE. Bold signifies significantly different performance (paired *t*-test, $\alpha < 0.05$).

	Normal		Adversarial	
	Act	Val	Act	Val
MuSE to IEMOCAP				
Unimodal (A)	0.419	0.376	**0.448**	**0.401**
Unimodal (L)	0.401	0.433	**0.436**	0.447
Multimodal (A+L)	0.422	0.431	**0.459**	**0.472**
MuSE to MSP-Improv				
Unimodal (A)	0.404	0.368	**0.431**	0.372

0.421) when tested on MSP-Improv.

- Valence: There is a significant increase in performance using acoustic setup (0.376 vs 0.401) and multimodal setup (0.431 vs 0.472) of adversarial model when tested on IEMOCAP. We see no significant difference in performance when testing on MSP-Improv using adversarial classification model .

Based on these results, we understand that removal of a psychological confounding factor, stress, generally aids in the generalizability of the model on completely unseen data, where the distribution of this confounding factor is unknown.

7.4.5   RQ5: Spontaneity as Confounding Factors

QUESTION: *Can we proactively remove other types of confounders to improve cross-dataset performance?*

HYPOTHESIS: *Removing the confounding factor of spontaneity in IEMOCAP will improve cross-dataset performance.*

We observed in the last question that "unlearning" the confounder stress can aid generalizability. Now, we want to see if the same method can be used to make models trained using other datasets more reliable to change in target data distribution. We hypothesize that decorrelating

the effect of spontaneity on emotion representation will lead to models that generalize better. This is because, as shown in [146], the emotional content expression is different in scripted vs spontaneous speech, and hence should modulate the emotion representation in trained model. To this end, we use the IEMOCAP dataset which has utterances that come from sessions that are both scripted and improvised. We do not use MSP-Improv for similar analysis here, because the scripted sessions, by corpus design, have limited lexical content and hence wouldn't cover enough input representation space for generalizability. We train the same 12 model variants described in section 7.3.1 replacing the adversarial stress classifier sub-component with adversarial spontaneity classifier for this analysis.

We report our results in Table 7.5. We compare the performance of the adversarial and normal models:

- Activation: There is a significant increase in performance in lexical (0.401 vs 0.425) and multimodal setup (0.433 vs 0.467) when the adversarial model is tested on MuSE dataset. We see no significant difference in performance when the adversarially trained model is tested on MSP-Improv.

- Valence: There is a significant increase in the performance using all setups of the adversarial classification model when tested on MuSE. We observe a significant increase in the performance in the acoustic setup of the adversarial model (0.410 vs 0.438) when tested on MSP-Improv.

We see that the removal of modulations due to the data collection methodology improves generalizability for many cross-dataset cases. This suggests that this method can be extended to train stabler models by explicitly accounting for confounding variables in limited amounts of training data.

Table 7.5: Performance (in UAR) predicting activation (left) and valence (right) in non-adversarial and adversarial (for best *lambda* value) setups across datasets when trained on IEMOCAP. Bold signifies significantly different performance (paired *t*-test, $\alpha < 0.05$).

	Normal		Adversarial	
	Act	Val	Act	Val
IEMOCAP to MuSE				
Unimodal (A)	0.428	0.401	0.427	**0.431**
Unimodal (L)	0.401	0.423	**0.425**	**0.455**
Multimodal (A+L)	0.430	0.429	**0.463**	**0.468**
IEMOCAP to MSP-Improv				
Unimodal (A)	**0.414**	0.410	0.402	**0.439**

7.4.6 RQ6: Lexical Patterns in Samples That Benefit from Adversarial Training

QUESTION: *Are there identifiable lexical patterns in samples that are especially successfully classified by the adversarially trained model for emotion classification?*

HYPOTHESIS: *Certain properties of input features correlate with the increase in probability of successful classification in adversarially trained emotion recognition models.*

Our results in questions 4 and 5 of section 7.4.4 demonstrate that decorrelating the representations from modulations due to confounding factors can positively affect the classification performance of our trained emotion recognition models when applied to datasets whose properties differ from the data on which the models were trained.

In this section, we aim to understand what properties of input features in a given sample correlate with the probability of successful classification in trained emotion recognition models due to decorrelation of such modulations. Understanding the relationship between the properties of the input features and the likelihood of success can help us identify data points that are more likely to be correctly classified using adversarially trained models. This can help us identify samples in an unseen dataset for which we can trust the classification label obtained from the adversarial model as compared to the normal classification model. We analyze this relationship using word tokens, which are low-dimensional and human-interpretable.

7.4.6.1 Adjusted Probability of Success

We study the correlation between the lexical patterns of data samples and the probability that those samples are correctly classified. We focus on improvements in classifaction, moving from the normal model to the adversarial model. This allows us to focus on improvement and mitigates the challenge that certain samples may just be particularly easy or hard to classify. We define probability of success for a sample as the $P_{A,s}(Success)$ where A can either be a normal (*normal*) classification model or an adversarial classification model (*adv*) and *s* refers to the index of a particular sample. We calculate $P_A(Success)$ as the ratio of the number of times a model correctly classifies a given sample to the total number of fifteen runs. If a sample is correctly classified across all runs by adversarial model, the $P_{adv}(Success)$ for that sample would be 1. But we want to concentrate on the gain in performance of using adversarial over normal classification. It might be the case that this sample is correctly classified across all runs by normal classification model as well, the $P_{normal}(Success)$ for that sample would be 1. In this case, we do not see any betterment as a result of using adversarial training paradigm. To mitigate the above limitation, we define adjusted probability of success in the following manner: We define adjusted probability of success (APS) for sample *s* as: $P_{adv,s}(Success) - P_{norma,sl}(Success)$. When the APS is greater than 0, the sample is more accurately classified using the adversarial model. When the APS is less than zero, it is more accurately classifed using the normal model.

7.4.6.2 Features

Our goal is to correlate APS with interpretable lexical features. We use the Linguistic Inquiry and Word Count (LIWC) [174] tool. LIWC assigns a predefined category to a word based on its association with social, affective and cognitive process. These categories have been shown to be highly predictive of both emotion [114], spontatenity [53] and stress [226].

We form a twelve length feature vector for each utterance by counting the number of words that fall into each of the nine LIWC categories (adverb, pronoun, social process,

negation, positive and negative emotion, insight, tentative and certainty). We normalize the feature vector by how many words in the utterance. We augment this feature vector to include: (1) fillers (e.g., "uhh"), hesitation (e.g., "like"), and discourse markers (e.g., "so") and (2) content rate, defined as the number of words per unit length of time. The final feature vector comprises of all the above mentioned categories.

7.4.6.3 Discussion

Decorrelating Stress. We report the Pearson correlation coefficient and the resulting Benjamini-Hochberg adjusted [25] p-values that we obtain between each feature in the vector, and the APS for each sample. We perform this assessment for both the activation and valence normal and adversarial lexical models. We focus on the cross-dataset case in which the models were trained on MuSE and tested on IEMOCAP (in Table 7.6). A large positive correlation between a category and APS implies that samples with larger numbers of words in a given category are likely to be classified correctly more often given the adversarial model versus the normal model.

- Activation recognition: APS is significantly positively correlated with the presence of words that relate to: Adverb (0.217), pronoun (0.165), positive emotion (0.154), certainity (0.138), fillers (0.154), discourse markers (0.141), and content rate (0.196).

- Valence recognition: APS is significantly positively correlated with the presence of words that relate to: Adverb (0.177), negative emotion (0.143), fillers (0.182), content rate (0.178).

This finding is consistent with previous research [151], where the authors have shown that there is often an increase in usage of function words and intensifiers in stressed conditions. So, for example a sentence "I am really really sad about losing my pen" would have more likelihood of being correctly classified by the adversarial model compared to the normal emotion classification model. Hence, we can hypothesise that an increase in the likelihood

Table 7.6: Correlation between LIWC features and APS due to stress decorrelation, for activation and valence. *p*-values are Benjamini-Hochberg adjusted for multiple comparisons (α = 0.05). p-value codes: '**'<0.01; '*'<0.05; '-'<0.1

	Act.		Val.	
	r	p	r	p
LIWC				
Adverb	**0.217**	**	**0.177**	*
Pronoun	**0.165**	-	0.082	*
Social Process (social)	0.084	-	0.001	-
Negations (negate)	-0.018	-	0.005	-
Positive emotion (posemo)	**0.154**	*	0.093	-
Negative emotion (negemo)	0.086	-	**0.143**	*
Insight	-0.021	-	-0.012	-
Tentative (tentat)	0.074	-	0.101	-
Certainty (certain)	**0.138**	*	0.116	-
Hesitation				
Fillers	**0.154**	*	**0.182**	*
Discourse marker	**0.141**	*	0.111	-
Content Rate	**0.196**	**	**0.178**	**

of correct classification of samples containing these intensifiers occurs due to reduced weightage of adverbs in adversarial training paradigm.

There are fewer significant categories for valence than for activation. This is consistent with the results in Table 7.4 for the lexical modality. Although we see a significant correlation between filler words and APS for activation classification [65], we do not observe the same for discourse markers and presence of social process words. The absence of significance in these cases implies that though these values are markers of stress, the normal classifier is still able to learn reliable representations invariant of stress for predicting the correct target label, resulting in negligible impact on classification performance when decorrelating the representations.

Decorrelating Spontatenity. We do a similar analysis for analyzing lexical properties of samples that were aided by decorrelating spontaneity. We report the Pearson correlation coefficient and the resulting Benjamini-Hochberg adjusted [25] *p*-values that we obtain

Table 7.7: Correlation between LIWC features and APS due to spontaenity decorrelation, for activation and valence. *p*-values are Benjamini-Hochberg adjusted for multiple comparisons ($\alpha = 0.05$). p-value codes: '**'<0.01; '*'<0.05; '-'<0.1

	Act.		Val.	
	r	p	r	p
LIWC				
Adverb	0.121	**	0.088	-
Pronoun	**0.138**	-	0.016	*
Social Process (social)	0.132	-	**0.166**	*
Negations (negate)	-0.003	-	-0.011	-
Positive emotion (posemo)	0.122	-	**0.161**	*
Negative emotion (negemo)	**0.137**	*	**0.148**	*
Insight	0.017	-	0.099	-
Tentative (tentat)	**0.155**	*	0.112	-
Certainty (certain)	**0.191**	*	**0.172**	*
Hesitation				
Fillers	**0.221**	**	0.119	*
Discourse marker	**0.189**	*	0.122	-
Content Rate	**0.165**	*	**0.144**	-

from the LIWC features and APS for both emotion axes lexical-based classification models (trained on IEMOCAP; tested on MuSE) in Table 7.7.

- Activation recognition: APS is significantly positively correlated with the presence of words that relate to: Pronoun (0.138), negative emotion (0.137), tentativeness (0.155), certainty (0.191), fillers (0.221), discourse markers (0.189) and content rate (0.165).

- Valence recognition: APS is significantly positively correlated with the presence of words that relate to: Social Process (0.166), positive (0.161) and negative (0.148) emotion, certainty (0.172), and content rate (0.144).

The results suggest that there are identifiable linguistic properties of samples whose likelihood of correct classification benefits from the model trained adversarially to decorrelate spontaneity and emotion representation as compared to normal classification model. This is especially true for the use of words in the certainty category for both emotion dimensions

and all hesitation categories for activation. Spontaneous speech has been shown to have more of these words in [53]. Scripted content has been shown to have more exaggerated displays of emotion through words and facial expressions [113]. Controlling for the weights assigned to words in positive and negative emotion categories using the adversarial model, leads to better classification of samples that are comprised of these word tokens.

7.5 Conclusions

This work focused on the interplay between stress and emotion in the context of automatic emotion recognition. We first showed that the presence of stress affects the performance of emotion recognition models. We then observed that these effects vary depending on modality (acoustic or lexical) and task (activation or valence classification). We then showed how decorrelating stress modulations from emotion representations aids the generalizability of the model. Next, we showed how a similar method could be used to control for variations due to spontaneity; facilitating the generalizability of the model. Finally, we identified human interpretable lexical markers that correlate with successful generalization of the model; especially concentrating on samples that are aided by decorrelation of stress and emotion representation.

Our results suggest that an extraneous psychological factor, such as stress, can significantly impact the performance of emotion recognition systems both within and across datasets. As a result, extraneous psychological factors should be accounted for when collecting data for training emotion recognition systems, especially when being used to predict labels of data that may or may not be modulated by those same factors. We then show how proactive decorrelation of this confounder can improve generalization of the model to other dataset at time of deployment.

CHAPTER VIII

Reducing Leakage of Demographic and Membership Information Through Adversarial Networks

8.1 Motivation and Contributions

The rise of mobile applications and conversational agents harnessing emotion recognition capabilities has brought a crucial challenge to the forefront: the safeguarding of sensitive user information. This data, encapsulating detailed demographic aspects, is often extracted from user devices and stored on centralized servers. In worst-case scenarios, these details could be exploited without user consent or manipulated by harmful entities. Storing abstracted representations rather than raw data, while a potential solution, doesn't fully counter the problem. Indeed, these representations, intended for tasks like emotion recognition, are susceptible to inadvertent demographic leakage undermining user-defined settings (for example, to not receive gender targetted ads) and exposing sensitive variables from unimodal (textual or acoustic) or multimodal data.

The consequences of unintentional demographic leakage are significant. Continual exposure of sensitive attributes threatens the integrity of user data. Moreover, the inherent tie-in of demographics such as age, gender, and race with emotion recognition models makes these models potential reservoirs of sensitive user information—even when storing only representational data. As a result, measures to counter demographic leakage have become a

pressing need.

Addressing these issues, our research focuses on developing robust methods to mitigate demographic leakage. We propose to use an adversarial learning model, which is trained with the aim to remove the learnt sensitive information encoded in the generated representation from any emotion recognition system. Varying the intensity of this "removal" offers us insight into the effects on the primary task and the ability of an attacker to accurately predict demographic information from just the input to a model and the generated representations. Additionally, we introduce a metric aimed at quantifying the defense against sensitive information exposure and the probability of membership identification — an unauthorized person's ability to determine if a user was part of the training data set. Our work across multiple datasets yields encouraging results; enhancing protection of sensitive data without significantly hampering the performance of primary emotion-recognition tasks. Ultimately, our work represents a pivotal stride in bolstering user data security within emotion recognition systems, setting in place a strong foundation for future research and improvements in this critical field.

8.2   Introduction

Virtual conversational agents strive to emulate human-like interaction to have more naturally flowing conversation [154]. These agents often employ models that classify aspects of communication, including the classification of the emotional content of speech. [101]. The resulting predictions can then be used to bias response generation. Emotion classification is also used in mobile and web applications to identify heightened risk of suicidal ideation or mood fluctuations [117, 148, 84], for the purpose of tracking or intervention. Data are sent from users' devices, including mobile applications [117] and Alexa or Google home devices [177], and are stored on central servers for analysis.

However, data transmitted from users' devices are vulnerable to data hacking and re-identification [18]. Eavesdroppers can use these data for identification of an individual

and to gain access to sensitive information. A way to counter this issue in data collected by mobile or smart home applications is to generate a data representation on the device and then to transfer that representation to the server for additional processing. The benefit is that these representations can decrease leakage of content and sensitive information by partially obfuscating the actual content of the conversation [24]. However, they still contain sensitive demographic information.

The implications of sensitive information leakage is profound: research has shown that discrimination occurs across variables of age, race, and gender in hiring, policing and credit ratings [92]. [2] showed how adding random noise to aggregated dataset or individual samples can ensure defense against attacks that aim to classify sensitive information from the representation. But, previous research has shown that using additional noise can often be exploited if the adversary has access to the network used to generate anonymity [120]. Therefore to ensure robustness, we consider a scenario of the attacker having access to the same embedding sub-network to generate the representations that will be used to train its attack network.

In this work, we focus on sensitive information encoding in the context of emotion recognition. Emotion recognition provides an important test case because emotion production varies significantly across gender and race. As a result, the outputs of emotion recognition models are often highly correlated with these secondary demographic signals [45, 202]. We design approaches to first measure leakage and then to counteract this leakage. We measure sensitive information encoding in two ways: 1) using a sensitive information reduction metric, which we define as the incapability of an attacker to recover demographic information from representations, and 2) by determining an adversary's ability to perform membership identification [136], defined as the ability to determine if a given user was in a dataset from which the emotion recognition models were trained (this can be harmful if the training data are collected in a sensitive context, such as counselling or therapy). We ask the following seven questions:

115

1. Does demographic leakage differ in umimodal and multimodal emotion recognition models?

2. How does the sensitive information reduction metric change when a network is trained to not encode sensitive information?

3. How does emotion recognition performance change when networks are trained to not encode sensitive information?

4. How does the adversarial component's strength impact emotion recognition performance and the sensitive information reduction metric?

5. Focusing on gender, how does the performance of emotion recognition change when a network is trained to not encode sensitive information?

6. Does the location of the adversarial component within a network affect the sensitive information reduction metric and emotion recognition performance?

7. Does the sensitive information reduction paradigm that we used to reduce encoding of gender information also help defend against other attacks such as membership identification?

Our results show that representations obtained for emotion recognition can be exploited by an adversary to predict sensitive variables given unimodal information (either audio or lexical). We further show that multimodal models contain even more sensitive information, as lexical and audio each encode different aspects of demographic information. We show that we can increase the defense against this attack by adversarially training representations to be invariant to gender. The novelty of this work is two fold: (1) we analyze how the demographic variable encoded in a representation differs across modalities and how it can be increased using adversarial paradigm; and, (2) we obtain enhanced representations that defend against multiple sensitive information leakage or prediction attacks while still maintaining performance on emotion recognition.

Given most of the previous work on representations that aim to reduce encoded sensitive information concentrates on just lexical information, we tackle the questions that arise from
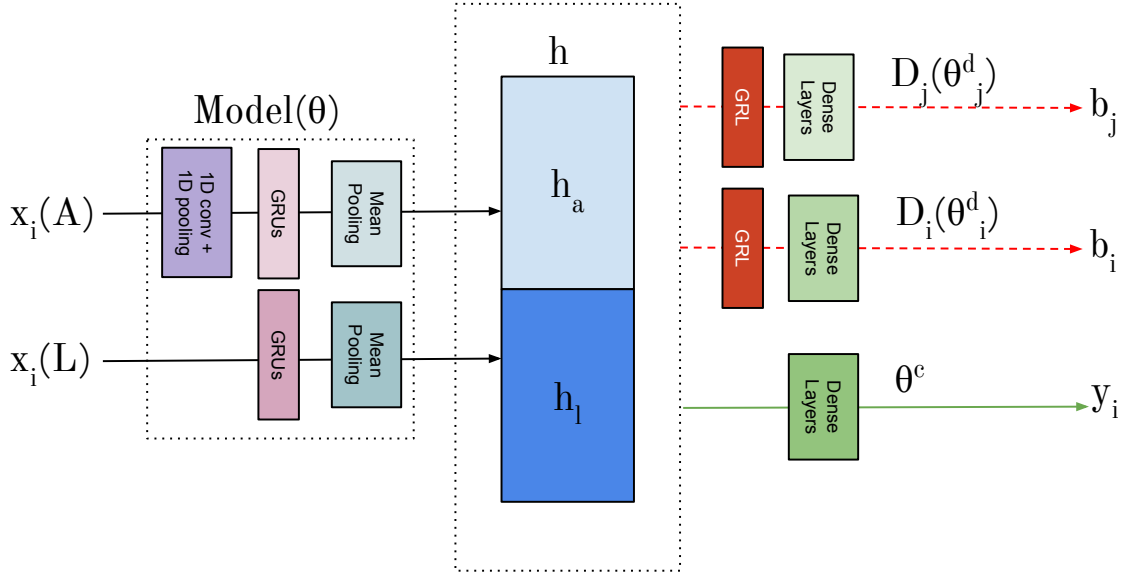
Figure 8.1: Privacy preserving network architecture.

desiring such preservation in multimodal representations for emotion recognition. While the primary goal of most previous works has been to avoid unintentional inference by the application itself, we concentrate on minimizing the potency of an attacker to deliberately recover sensitive attributes from an invariant representation.

8.3 Experimental Setup

We use four common emotion recognition datasets: MSP-Improv [38], MSP-Podcast [143], Interactive Emotional Dyadic MOtion Capture (IEMOCAP) dataset [36], and Multimodal Stressed Emotion (MuSE) dataset [105]. We use the acoustic and lexical features, MFBs and word2vec respectively, as defined in Section 3.3. Next, we describe the network architecture, the training recipe, and the metrics in consideration.

8.3.1 Architecture

The objective of this system is to maximize the performance of the emotion classifier while minimizing the performance of the gender classifier (see Figure 8.1). The main

network consists of three components: (1) embedding sub-network, $Model(\theta)$; (2) emotion classifier, θ^c and output y_i; and (3) gender classifier, $D_i(\theta_i^d)$, with output, b_i. We then disucss how an attacker network could maliciously use this information to obtain sensitive demographic information.

Main Network. The *embedding sub-network* uses a state-of-the-art multimodal approach in emotion recognition [9] in which the acoustic and lexical information are processed separately and then joined after the application of modality-specific global mean pooling. The acoustic input stream $(x_i(a))$, where i represents an input frame (40-dimensional) and a represents acoustic, is processed using a set of convolution layers and Gated Recurrent Units (GRU), which are fed through a mean pooling layer to produce an acoustic representation (h_a). The lexical input $(x_i(l))$, where i represents an input word (300-dimensional) and l represents lexical, is passed through GRUs and pooled to obtain a lexical representation (h_l). For the multimodal setup, these two representations, (h_a) and (h_l), are concatenated (h). The representations (h, h_a, h_l) are of fixed-length given acoustic and lexical inputs. The *emotion classifier* takes in the representation $(h, h_a,$ or $h_l)$ as input and estimates valence or activation using a set of dense layers. The *gender classifier* estimates the gender label (i.e., male or female) using a set of dense layers.

Gender-Leakage. The main network is trained to unlearn gender. To achieve this goal, we use a Gradient Reversal Layer (GRL) [77]. GRLs are a common multi-task approach to train networks such that they are invariant to specific properties [153, 107]. During the backward pass of the training phase, the GRL multiplies the backpropogated gradients by $-\lambda$ (i.e., the strength of the adversarial component). During the forward pass, the GRL acts as an identity function. To make the network invairant to gender, we place the GRL function between the embedding sub-network and the gender classifier. We obtain gender-invariant representations using the following loss function:

$$\widehat{\theta} = \min_{\theta_M} \max_{\{\theta_{D^i}\}_{i=1}^{N}} \chi(\widehat{y}(x; \theta_M), y) - \sum_{i=1}^{N}(\lambda_i . \chi(\widehat{b}(x; \theta_{D^i}), b_i))$$

where *N* is the number of targeted sensitive variables (here *N*=1). The loss function ensures that while the output components are trained to be good predictors, the representation is trained to be maximally good for the primary task (emotion) and maximally poor for the secondary task (gender).

Attacker Network. We assume that the attacker has access to a held-out dataset (either a different dataset or a section of the original dataset) with known gender labels. The attacker generates representations for this dataset using the previously described embedding sub-network. The network then learns to predict gender labels from the generated representations using a set of dense layers. Since the parameters used to construct the representation are fixed, the attacker only acts upon its own parameters to optimize the gender prediction linear loss. The purpose of the attacker's network is to recover gender information from representations whose labels are unknown. Though testing using a singular network isn't a guarantee of robustness of the representation to attacks that aim to predict or extract sensitive information from representations, for the scope of this chapter, we use a feed forward network, one of the powerful learning methods on a fixed size static representation.

Model Variations. We use 12 variants of the network shown in Figure 8.1. We train combinations of the following setups: {*general-classification-model (Gen), sensitive-information-reduction-classification-model (SIR)*} × {*activation, valence*} × {*uni-lexical, uni-acoustic, multimodal*}.

The general classification setup makes use of the embedding sub-network with text, speech, or both as input streams and the emotion classifier. The sensitive information reducing classification setup adds the gender classifier to the general setup.

Training. We implement models using the Keras library. [52]. We use a weighted cross-entropy loss function for each task and learn the model parameters using the RMSProp optimizer. [214]. We train our networks for a maximum of 50 epochs and monitor the validation loss for the emotion classifier after each epoch, stopping the training if the validation loss does not improve for five consecutive epochs. Once the training ends, we

revert the network's weights to those that achieved the lowest validation loss on the emotion classification task. For the classification model trained to not encode sensitive information, we ensure that the chosen model yields a chance unweighted average recall (UAR) for the gender classification task on the validation set. Finally, we train each setup three times with different random seeds and average the predictions over these runs to reduce variations due to random initialization.

We use validation samples for hyper-parameter selection and early stopping. The hyper-parameters that we use for the main network include: number of convolutional layers $\{3, 4\}$, width of the convolutional layers $\{2, 3\}$, number of convolutional kernels $\{32, 64, 128\}$, number of GRU layers $\{2, 3\}$, GRU layers width $\{32\}$, number of dense layers $\{1, 2\}$, dense layers width $\{32, 64\}$, GRL λ $\{0.3, 0.5, 0.75, 1\}$. For the adversarial emotion classification setups, we use the hyper-parameters that maximize the validation emotion classification performance while minimizing the validation gender classification performance. For the attacker's model, we use the following hyper-parameters: number of dense layers $\{2, 3, 4\}$, dense layers width $\{32, 64\}$. We report the UAR performance of our models, given the imbalanced nature of our data. [184].

8.3.2 Metrics

Performance. We define performance for emotion recognition as the ability of the model to correctly classify either activation or valence into 3 categories: low, medium, and high. We measure performance using UAR (chance is 0.33).

Demographic Leakage. Leakage is defined as the ability of a trained gender classifier to predict gender from the representations which are obtained when the network is simultaneously trained to perform the primary task.

Demographic Sensitive Information Reduction Metric. We define the sensitive information reduction metric as the inability of an attacker to be able to recover gender from the representations trained on a primary task. To test this, we use four phases of training.

1. We train the main network on a dataset (D1), represented by the pair (x_{D1}, y_{D1}), where x is the data input while y is the gender label. We obtain representations for this dataset $(h(x_{D1}))$.

2. We consider that the attacker has access to another dataset or unused subset of the same dataset (D2) represented by the pair (x_{D2}, y_{D2}). We generate representations $h(x_{D2})$ for the pairs in this dataset using the embedding sub-network of the main network.

3. We train a model (M_{att}) to predict gender labels using the representations obtained in step 2, represented as $M_{att}((h(x_{D2}), y_{D2}))$.

4. Using the model obtained previously (M_{att}), we choose $h(x_{D1})$ as inputs, and measure the gender prediction capability of the attacker $UAR(M_{att}((h(x_{D1}), y_{D1}))$. The Demographic Sensitive Information Reduction metric of an attacker is then quantitatively defined as $1 - UAR(M_{att})$.

The range of the sensitive information reduction metric goes from 0 (the attacker is always correct) to 0.5 (the attacker has a chance UAR).

Membership Identification. Membership identification is the possibility of an attacker being able to recognize if a speaker belongs to the training set. We assume that the adversary can obtain samples from a speaker from the same distribution as that for the training set. Consider that the adversary knows some speakers for whom representations definitely exist in the training set and some for whom they definitely don't. We test the possibility of membership identification using four steps:

1. We simulate the above using cross-validation. Given five speaker independent folds, we use three for the training set. From the remaining two folds, we add some samples of the selected speakers to the training set.

2. We consider that the attacker knows both, the speakers selected and not selected for training from set four $(s4)$, but has no information about this split for set five $(s5)$. The objective of the attacker is to predict whether speakers were selected for inclusion in

the training set from ($s5$).

3. The attacker trains a binary classification model comprised of dense layers (M_{att-mi}) using the representations obtained from dataset D1 as $M_{att-mi}(h(x_{D1}), \text{`}Yes\text{'})$. It obtains representations of the samples not used in training for the the selected speakers included in the training set and trains its model as $M_{att-mi}(h(x_{s4_selected}), \text{`}Yes\text{'})$ and for the speakers not included in training as $M_{att-mi}(h(x_{s4_selected}), \text{`}No\text{'})$. A speaker is saved from each label for validation.

4. We then define the UAR of the performance of $M_{att-mi}(h(x_{s5}))$ as membership identification.

Table 8.1: Results using general (left) and models trained to not encode sensitive information (right) for activation and valence prediction. U-UAR, U(M/F)-UAR for male/female, L-leakage, SIR-sensitive information reduction metric, MI-membership identification. ***Bold-Italic*** shows significant improvement in metrics as compared to general classification model and *Italic* shows significant difference in metrics as compared to the model trained to not encode sensitive information. Significance is established using paired t-test at adjusted p-value< 0.05.

(a) Prediction of activation using general (left) and models trained to not encode sensitive information (right)

		General Classification						Privacy Preserving Classification				
		L	U(M)	U(F)	U	SIR	MI	U(M)	U(F)	U	SIR	MI
Audio	Imp	0.69	0.65	*0.62*	*0.63*	0.35	0.71	0.64	0.57	0.60	***0.44***	0.68
	Pod	0.71	0.69	0.70	0.70	0.32	0.73	0.68	0.69	0.69	***0.44***	***0.68***
	Iem	0.73	0.66	0.69	0.67	0.30	0.72	***0.68***	0.70	***0.69***	***0.43***	***0.67***
	MuS	0.72	*0.61*	*0.64*	*0.63*	0.33	0.75	0.58	0.61	0.60	***0.45***	***0.69***
Lexical	Iem	0.62	0.51	0.52	0.52	0.39	0.59	***0.55***	***0.56***	***0.56***	***0.48***	***0.55***
	MuS	0.64	0.54	0.56	0.55	0.38	0.60	***0.58***	0.57	***0.58***	***0.47***	0.58
Multimodal	Iem	0.74	0.66	0.70	0.68	0.30	0.74	0.66	0.69	0.68	***0.41***	***0.67***
	MuS	0.73	0.65	0.66	0.66	0.31	0.76	0.65	0.64	0.65	***0.43***	***0.69***

(b) Prediction of valence using general (left) and models trained to not encode sensitive information (right)

		General Classification						Privacy Preserving Classification				
		L	U(M)	U(F)	U	SIR	MI	U(M)	U(F)	U	SIR	MI
Audio	Imp	0.56	0.53	0.49	*0.51*	0.44	0.70	0.51	0.49	0.48	***0.48***	0.68
	Pod	0.60	0.56	0.57	0.56	0.42	0.71	0.55	0.56	0.56	***0.47***	0.70
	Iem	0.62	0.60	0.61	0.60	0.39	0.70	0.60	0.62	0.61	***0.45***	0.68
	MuS	0.58	0.50	0.47	0.48	0.42	0.72	0.48	0.47	***0.47***	0.46	0.71
Lexical	Iem	0.61	0.64	0.65	0.65	0.41	0.62	***0.67***	***0.68***	***0.67***	***0.46***	0.62
	MuS	0.57	0.68	0.69	0.68	0.45	0.63	0.70	0.71	0.70	0.47	0.62
Multimodal	Iem	0.68	0.67	0.71	0.69	0.32	0.70	0.68	0.70	0.69	***0.45***	***0.68***
	MuS	0.64	*0.67*	0.66	*0.67*	0.38	0.71	0.64	0.65	0.65	***0.46***	0.71

8.4 Analysis

In all the tables, **U** is the unweighted average recall (UAR), and **U(M)** and **U(F)** represent the performance of the model for emotion recognition when gender is male and female respectively. Leakage in the model is represented by **L**, the lower the better, where chance leakage is 0.5. sensitive information reduction metric, represented by **P**, ranges from $[0, 0.5]$, and is the incapability of an attacker to obtain demographic information from the representation, the higher the better. Membership identification represented by **MI**, ranges from [0.5 (chance UAR),1], and is the capability of an attacker to identify if the subject belongs in the training set, for which the lower the value, the better. We code identify the datasets as follows: **Imp**-MSP-Improv; **Pod**-MSP-Podcast; **Iem**-IEMOCAP; and **MuS**-MuSE. All significance tests are paired t-test, with significance established (shown in bold) when **Benjamini-Hochberg adjusted** (FDR = 5%) p-value< 0.05.

8.4.1 RQ1: Unimodal vs Multimodal Leakage

Q: *Does demographic leakage differ in umimodal and multimodal emotion recognition models?*

HYPOTHESIS: *Multimodal representations leak more gender information than unimodal representations.*

Previous research has shown that different modalities have varying capabilities of capturing demographic information, such as age or gender [134] information.The authors showed that audio, as compared to lexical, is used more successfully to predict gender. Hence, we hypothesize that a combination of these modalities leads to an increase in the leakage of the sensitive variable.

We train the six setups separately for each dataset described in Section 8.3.1 for activation and valence. We report the average across five-fold speaker-independent cross-validation in Table 8.1a and Table 8.1b. We find that:

- A network trained to only recognize emotion is generally discriminative for gender as well. For instance we obtain a leakage of 0.73 when training a multi-modal network for activation and of 0.64 when trained for valence on MuSE.

- In unimodal systems, leakage is higher when systems are trained using only audio streams compared to lexical.

- Leakage of gender in learned representation is higher for multimodal systems than that for the unimodal systems for both, MuSE and IEMOCAP (the two datasets with both audio and lexical information).

Our results suggest that models that aren't explicitly trained for gender recognition, or, that don't use gender as an input feature, still learn representations that are discriminative to identify gender. This leakage is more prominent when the input stream is audio as compared to lexical, but the leakage compounds in multimodal systems.

8.4.2 RQ2: Privacy Preservation Performance

Q: *How does the sensitive information reduction metric change when a network is trained to not encode sensitive information?*

HYPOTHESIS: *Representations that are gender-invariant are less prone to leakage when attacked by an adversary, leading to reduced leakage of sensistive information.*

Previous research has shown that obtaining a representation from a model trained to be invariant to gender, age, or location leads to better protection from an attacker who tries to recover this information [55]. Previous research [67] has also shown that while the representations might be trained such that leakage of sensitive variable is reduced to chance, the attacker might still be able to recover this information. Hence, we concentrate on using this incapability as our primary metric. To test our hypothesis, we train the adversarial variants of the six models as mentioned above, while making sure that the leakage in the models is reduced to chance performance and compare our results to those in Table 8.1a and Table 8.1b. We train the multimodal models only for MuSE and IEMOCAP. Our results in

Table 8.1a and Table 8.1b show that:

- The sensitive information reduction metric is always higher when the representations are trained adversarially, compared to generally.

- Even when leakage is adversarially reduced to chance, the attacker is still able to recover information about gender.

- The sensitive information reduction metric is in general always lower for audio than for lexical based unimodal systems.

- Multimodal systems often have the lowest sensitive information reduction metric.

Our results suggest that, though the sensitive information encoded in the learned representation is reduced using the proposed method, the attacker can still recover that information. This effect is especially compounded for multimodal systems. While previous work has concentrated on text (Section II), our work shows how audio is the major culprit and that models involving audio as input are easier to exploit, even when trained adversarially for not encoding sensitive information.

8.4.3 RQ3: Privacy Preserving Emotion Recognition Performance

Q: *How does emotion recognition performance change when networks are trained to not encode sensitive information?*

HYPOTHESIS: *There is a minor drop in emotion recognition performance when models are trained to not encode sensitive information.*

Previous research has shown that training a model invariant to a dataset variable might lead to drop in performance on the primary task, especially when there exists known correlations or biases in the datasets between the target label for the primary task and the secondary task [153]

We compare the performance for predicting activation and valence of the models trained just to predict emotion (Act: Table 8.1a, Val: Table 8.1b) versus the model trained

to not encode sensitive information while still predicting emotion in (Act: Table 8.1a, Val: Table 8.1b). Our results suggest that, in general there is no significant effect on the performance on the primary task when we train networks to not encode sensitive information. We find that the performance is either maintained, e.g., Act: multimodal-MuSE; Val: multimodal-IEMOCAP, or there is a slight decrease in performance for some setups, e.g., Act: unimodal-acoustic-MuSE; Val: multimodal-MuSE. In multiple cases, such as Act/Val:unimodal-lexical-MuSE/IEMOCAP, contrary to some previous work, we also see a significant increase in performance, implying that making the model invariant to gender increases its robustness by not learning replicable associations between gender and emotion label.

8.4.4 RQ4: Privacy Preservation Strength

Q: *How does the adversarial component's strength impact emotion recognition performance and the sensitive information reduction metric?*

HYPOTHESIS: *As the strength of the adversarial component increase, the sensitive information reduction metric increases and the performance on the pimary task is unchanged.*

Our results in Section 8.4.2 suggest that while the leakage of the model was reduced to chance performance, the attacker is still capable of recovering this information. We analyze the effect of the strength of the adversarial component on the performance of the primary task and the sensitive information reduction metric.

We find that the emotion recognition performance is generally unaffected with a change in λ, as expected from the results in Section 8.4.3. We observe that the the attacker is usually less capable of inferring gender from learned representations when $\lambda = 0.50$ as compared to when $\lambda = 0.75$. For example, the sensitive information reduction metric for the unimodal audio system trained on MuSE increases from 0.39 to 0.45. But contrary to our expectation, we often see a decrease in the sensitive information reduction metric when we move from $\lambda = 0.75$ to $\lambda = 1.00$ for both activation and valence. For example, the sensitive

information reduction metric for the unimodal audio system trained on MuSE decreases from 0.45 to 0.41. The decrease in the sensitive information reduction metric as $\lambda \to 1$ could be attributed to overfitting of data [194] when being trained for invariance to the sensitive variable which the attacker network is able exploit. This suggests that an increase in the strength of the adversarial component doesn't necessarily correlate to an increase in the sensitive information reduction metric.

8.4.5   RQ5: Privacy Preserving Emotion Recognition Performance Change & Gender

Q: *Focusing on gender, how does the performance of emotion recognition change when a network is trained to not encode sensitive information?*

HYPOTHESIS: *Learning representations invariant to gender will affect performance on the primary task in an imbalanced manner across subgroups.*

Previous research [14] has shown that training models invariant to race or gender can harm performance for one group more than others. This may be worrying when the prediction is used for sensitive application such as intervention or policing. Hence, we analyze if the performance on emotion recognition is affected in an imbalanced way for the models trained to not encode sensitive information.

We compare the performance for predicting activation and valence of the models trained just to predict emotion (Act: Table 8.1a, Val: Table 8.1b) versus the model trained to not encode sensitive information while still predicting emotion in (Act: Table 8.1a, Val: Table 8.1b). We find that while the performance is affected differently for the subgroups, the effect is not consistent across multiple setups and datasets. For example, the unimodal-acoustic system trained on MSP-Improv for activation classification decreases in performance for both the male and female groups, but the effect on the female group is greater. But the pattern isn't consistent across other datasets for the same model setup. Our takeaway from this analysis is cautionary, that though the sensitive information reduction metric increases when a model is adversarially trained to not encode sensitive information, we need to ensure

that the performance of the model on that dataset doesn't harm one subgroup more than the other.

8.4.6   RQ6: Location of Adversarial Component

Q: *Does the location of the adversarial component within a network affect the sensitive information reduction metric and emotion recognition performance?*

Hʏᴘᴏᴛʜᴇsɪs: *Unlearning the demographic variable in separate pooled streams will improve the sensitive information reduction metric.*

Previous work has shown that curtailing a variable on intermediate layers often leads to a difference in the performance of the classifier [44]. As seen in Section 8.4.1, audio is more prone to leakage than lexical information, hence, a multimodal system's sensitive information reduction metric might benefit from curtailing audio separately. Our initial multimodal model (Fig 8.1) only allows for the same strength and parameters of the adversarial component to be applied for both audio and lexical streams. To test our hypothesis, we place the same adversarial component after the mean pooling layer of both input streams, allowing us separate control of gender invariance for both modalities, before concatenation of representation.

We show our results in Table 8.2. We find that, using adversarial component separately for each input stream improves sensitive information reduction metric for emotion recognition models trained on both datasets, as compared to using one adversarial component. This suggests that a granular control of invariance over modalities leads to better defense of representations against gender identification.

8.4.7   RQ7: Defending Set-Based Membership Identification

Q: *Does the sensitive information reduction paradigm that we used to reduce encoding of gender information also help defend against other attacks such as membership identification?*

Hʏᴘᴏᴛʜᴇsɪs: *Membership identification will decrease when models are trained to be invariant to speaker.*

Table 8.2: Results for activation (Act) and valence (Val) prediction using multimodal input, when adversarially unlearning gender in each input (SIR-E) [left] stream separately. U-UAR, U(M/F)-UAR for male/female, P-sensitive information reduction metric, MI-membership identification. ***Bold-Italic*** shows significant improvement in the sensitive information reduction metric as compared to model trained to not encode sensitive information by maximizing loss on the concatenated representation (SIR-C) [right]. Significance is established using paired t-test, adjusted p-value< 0.05.

		SIR-E			SIR-C		
		U	P	MI	U	P	MI
Act	Iem	0.66	***0.43***	0.73	0.68	0.41	0.67
	MuS	0.65	0.44	0.74	0.65	0.43	0.69
Val	Iem	0.67	***0.46***	0.70	0.69	0.45	0.68
	MuS	0.66	0.47	0.74	0.65	0.46	0.71

Membership identification is defined as an attack that tries to identify if samples from a speaker 'x' were present in the training set [136]. [168] showed that removing identifying factors from learned representations reduces the probability of membership leakage. For this analysis, we ask two questions: (a) can we defend against membership identification using a proxy task and, (b) can we defend against both, gender and membership identification?

We train an attack model for membership identification as specified in Section 8.3.2. We find that while adversarial removal of gender in the learned representation (Act: Table 8.1a and Val: Table 8.1b) does lead to reduced membership identification, as compared to a model trained solely for emotion recognition (Act: Table 8.1a and Val: Table 8.1b), the membership identification is still far higher than chance.

Our goal is to be unable to identify whether samples from speaker 'x' exist in the training set. This is different from the usual membership defense that prevents prediction of presence of a data-point pair $(input_x, output_x)$ is in the training set. As a result, we require a proxy task, because our model cannot use samples from the speakers not in the training set even to induce invariance. We hypothesize that given randomly chosen speakers from the population, speaker-invariant training leads to representations that are less likely to encode speaker-specific information. This will make it harder for the attacker to identify membership of a particular speaker in the training set. We train the emotion recognition models specified

Table 8.3: Results for activation and valence prediction, for general classification (General), and, when adversarially unlearning subject identity (SIR-SubjectID) and both subject identity and gender (SIR-Multiple). U-UAR, U(M/F)-UAR for male/female, SIR-sensitive information reduction metric, MI-membership identification. ***Bold-Italic*** shows significant improvement in metrics as compared to general classification model and *Italic* shows significant difference in metrics as compared to the models trained to not encode sensitive information. Significance is established using paired t-test at adjusted p-value< 0.05.

		General			SIR-SpeakerID			SIR-Multiple		
		U	SIR	MI	U	SIR	MI	U	SIR	MI
		Activation								
Audio	Imp	*0.63*	0.35	0.71	0.59	***0.40***	***0.58***	0.59	***0.45***	***0.58***
	Pod	0.70	0.32	0.73	0.67	***0.37***	***0.60***	0.69	***0.46***	***0.59***
	Iem	0.67	0.30	0.72	0.66	***0.35***	***0.58***	0.67	***0.43***	***0.57***
	MuS	*0.63*	0.33	0.75	0.61	***0.36***	***0.62***	0.59	***0.44***	***0.60***
Lexical	Iem	0.52	0.39	0.59	0.51	0.40	0.52	0.53	***0.48***	***0.52***
	MuS	0.55	0.38	0.60	0.52	0.39	0.53	0.54	***0.47***	***0.52***
Multimodal	Iem	0.68	0.30	0.74	0.67	0.33	***0.58***	0.66	***0.40***	***0.57***
	MuS	0.66	0.31	0.76	0.65	0.33	***0.60***	0.65	***0.40***	***0.58***
		Valence								
Audio	Imp	*0.51*	0.44	0.70	0.47	0.45	***0.54***	0.47	0.48	***0.53***
	Pod	0.56	0.42	0.71	0.55	0.43	***0.56***	0.54	***0.48***	***0.56***
	Iem	0.60	0.39	0.70	0.61	0.41	***0.59***	0.60	***0.47***	***0.57***
	MuS	*0.48*	0.42	0.72	0.45	0.42	***0.60***	0.46	***0.46***	***0.58***
Lexical	Iem	0.65	0.41	0.62	0.67	0.41	***0.52***	0.66	***0.47***	***0.53***
	MuS	0.68	0.45	0.63	0.68	0.44	***0.53***	0.68	0.46	***0.53***
Multimodal	Iem	0.69	0.32	0.70	0.69	0.34	***0.57***	0.65	***0.43***	***0.56***
	MuS	*0.67*	0.38	0.71	0.64	0.37	***0.58***	0.62	***0.44***	***0.58***

in Section 8.3.1 and replace the gender invariance sub-network with speaker invariance and use the same membership attack network.

We show our results in Table 8.3. We find that models trained to be invariant to speaker identity have significantly lower UAR for membership identification than those trained solely to recognize emotion, or trained invariant to gender, which matches our hypothesis.

Extension towards multi-attribute invariance. We train our emotion recognition model using both the adversarial components (speaker id and gender) and the primary classification task i.e., emotion recognition. This ensures that the model can defend against both, gender and membership identification attacks. We report our results in Table 8.3. We

find that we can successfully train models that are safer against both, gender and membership identification attacks, while still maintaining similar performance on the primary task, as an evidence towards multi-attribute invariance.

8.5   Conclusion

In this work, we show how sensitive information preserving networks trained for emotion recognition can be used to protect against gender and membership identification. This provides a compelling case for separating the process of data processing on user devices and of task-specific training on central servers. While in this chapter we concentrate on a single primary task i.e., emotion recognition, this method can be extended to maximize utility on multiple primary tasks that are loosely related to each other and are benefited from a multi-task setup as shown for dialogue act and turn detection, and sentiment and topic classification [187].

CHAPTER IX

Sociolinguistics Based Human-Centered Metric for Emotion Recognition

9.1 Motivation and Contribution

Large language models have limitations in subjective tasks like emotion recognition, partly due to insufficient annotation diversity and inadequate data coverage. A diverse range of annotations is crucial for capturing the complex expressions of human emotions, while extensive data coverage renders it more likely that the many possible variations of these expressions are adequately represented. However, obtaining a comprehensive and diverse set of annotations in emotion datasets is both time-consuming and expensive, involving the recruitment of numerous human annotators to provide their unique perspectives on emotional expressions.

These annotation costs are further compounded by the need to evaluate models against diverse emotions, which often require not only human annotation for the collected data but also for the generated emotion label on unseen data for validation purposes. Current strategies rely heavily on costly, time-consuming human-based feedback for model evaluation, which, while effective in their own right, may not always be economically viable or efficient.

To address these challenges, we propose using metrics that are based off sociological literature to capture the nuances of human behavior and emotions in a more cost-effective manner.

We propose model-agnostic evaluation templates can be instantiated with sociological word lists, and these templates can be generalized to any task with existing word lists. We introduce two metrics: one for emotion generalization and another for intentional reduction of learnt sensitive information. These metrics aim to comprehensively assess model performance, while reducing dependency on expensive human-based feedback. By effectively evaluating the robustness and sensitive variable leakage of models, the proposed metrics provide valuable insights into model performance, resulting in more efficient utilization of resources.

Evaluating their effectiveness in enhancing model performance and reducing leakage of sensitive demographic, we find that these metrics are significantly correlated with model performance and can improve cross-corpus results. The proposed method empowers us to leverage the benefits of diverse and comprehensive emotion recognition models, while mitigating the costs associated with the traditional, labor-intensive feedback processes. Ultimately, this approach leads to the development of more accurate and relevant emotion recognition models in a cost-effective manner.

9.2   Introduction

Recent advances in natural language processing have led to the development of large and very large language models that can be trained on massive amounts of labeled data. However, despite their impressive performance on various natural language processing tasks, these models do not necessarily learn behavior that is similar to that of humans. Data diversity is particularly important in subjective, paralinguistic tasks, such as those seen in emotion recognition and other behavior modeling areas. But paralinguistic datasets are often relatively small, compared to the size of datasets seen in fields such as speech recognition. Further, it can be difficult to identify appropriate metrics for these subjective tasks and, as a result, those common in emotion recognition tend to be metrics such as cross-entropy loss between the predicted and original labels.

The challenge with the common metrics is that they do not capture the full complexity of

the problem, leading to models that have limitations at their core. Models trained to optimize for cross-entropy are often suffused with biases. Reducing learnt sensitive information or generalizing these models can be extremely difficult because sensitive information reduction processes generally rely on these same data and labels. Therefore, limitations in the original dataset are still present when the model attempts to unlearn these problematic associations [192, 81].

Researchers have looked at countering these issues in several different ways: from data generation, to data augmentation, and, from annotation diversification to comprehensive data collection [46]. Each of these methods comes with its own pros and cons. For example, generating data samples that are not perceived differently by humans is not only challenging, but also, requires extensive post-hoc evaluations [109]. Researchers have also looked at diverse data collection methods, focusing on active learning optimizing for data diversity [182]. While this method generally improves performance by a significant measure, we can still end up with datasets that reinforce unwanted stereotypes and prejudices [49].

This has led to a growing need for feedback that is based on human evaluation to ensure that the models produce more accurate and relevant results. One of the main challenges of using human-based feedback though is that it can be expensive to obtain, particularly for tasks that involve emotions. This is because emotions are subjective and can vary widely depending on demographic and situational factors. As a result, there is a need for alternative methods that can provide metrics for evaluating the performance of language models that are centered around human behavior.

To address this issue, we propose leveraging metrics that are based on sociological literature, which can capture the nuances of human behavior and emotions. By drawing on insights from sociology, we can develop more nuanced and context-sensitive metrics that can better capture the complexities of human behavior. This approach has the potential to provide a more comprehensive optimization of language models, while also reducing the dependence on expensive and time-consuming human-based feedback.
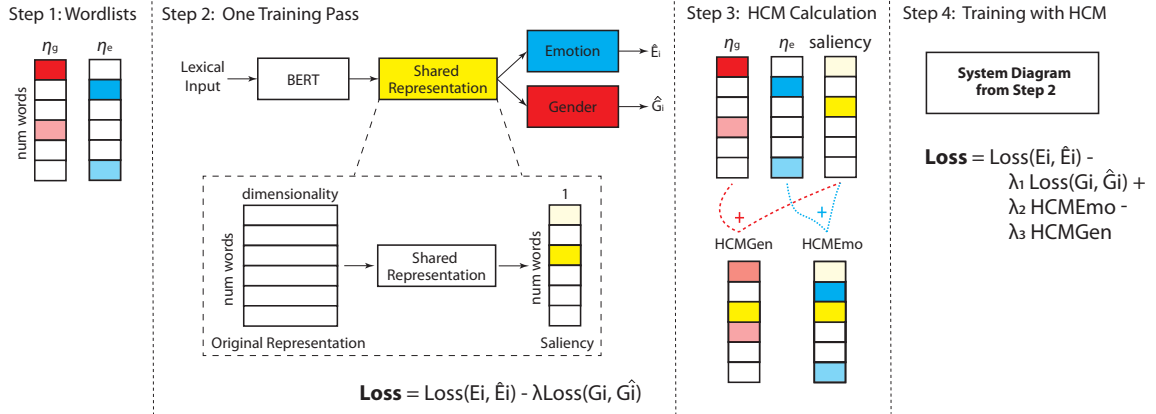
Figure 9.1: Diagram of the approach. In step 1, the wordlists are created (e.g., for gender, w_g, and for emotion, w_e. The length is the number of words. The brightness corresponds to the strength of the relationship between the word and the category (gender or emotion). In step 2, the model learns a shared representation with the goal of learning emotion, $\hat{E}_i$, and unlearning gender, $\hat{G}_i$. The shared representation is used to create a word-level saliency vector. In step 3, the HCM metrics are calculated by combining the salience with either w_g or w_e. In step 4, the model is trained with a loss that includes the original target (maximizing emotion accuracy and minimizing gender accuracy) in addition to the two HCM metrics.

9.3   Intuition Behind Human-Centered Metric Design

In this chapter, we propose a novel approach to developing task-invariant metric templates that are centered around human perception. To begin with, we define metric blueprints, called as metric templates. These templated metrics that have generic components that can be adapted for specific tasks. Our aim is to leverage these templates to evaluate a model, and to add a training objective that relies on sociological literature in addition to human-obtained labels. This is a significant departure from the traditional approach of using labeled data, which can be costly and time-consuming to obtain.

Ideally, important words identified by models should align well with human's interpretations. To achieve this, we use model interpretability; any method that provides us some intuition as to the words that the model pays attention to for its prediction to define the metric templates. This allows us to have a training objective that optimizes for importance assigned to certain words based on sociolinguistic literature. This approach is different from using labels because it enables us to incorporate sociolinguistic knowledge into the metric

templates.

The metric templates we propose consist of three components: directionality (sign), input interpretation saliency weights, and word lists. The combination of the three components results in a task-invariant metric template.

The sign component specifies whether the metric is meant to be maximized or minimized. By setting the sign, we can control the direction of the correlational relationship between the input and output variables, and thus, improve the accuracy of our predictions.

The input interpretation saliency weights component specifies how the input data should be interpreted. They are saliency values obtained from the model for each word token in the input. These values are highly correlated with the predicted label and can be positive, negative, or zero. A positive value indicates that the presence of a word correlates with the prediction of that label. A negative value indicates that the presence of that word correlates with "not predicting the given label." A zero value means that the word has no influence on the prediction.

The wordlists component provides a list of words that should be given special consideration when computing the metric. These lists can be grouped by class labels and may have associated weights with each word, but both of these features are optional. Wordlists are an essential tool for training machine learning models, as they allow us to focus on specific words or concepts that are relevant to our task. By using wordlists, we can improve the accuracy and efficiency of our models, as well as gain valuable insights into the relationships between words and concepts in our data.

In summary, our metric template allows us to focus on three separate aspects, each aligned with one of the components. Sign allows us to encourage the model to learn the correlation between a particular word and the output (sign of loss function). The input interpretation saliency weights allow us to measure how well the model captures the correlation between all words in a sentence and the output. The wordlists allow us to impact the learned correlations to align with our expectations about the relationshp between a given category and word

choice.

We now ground these aspects, focusing on a given task-related wordlist focusing on the task of emotion recognition. For our purposes, we use the NRC-VAD [159] and a self curated, Emotion-Gender wordlists. We first calculate the overlap between the words in the input sentence and the task-related wordlist(s). This overlap allows us to identify words whose correlations (learned by a model) may need to change. For class-based wordlists, we use the signed correlation that corresponds to the predicted class label, which is controlled by the sign component of our metric template. By setting the sign, we can encourage the model to learn the correlation between a particular word and the output. In cases where we want to unlearn a task completely, we aim to move the correlation of those words to zero to remove any influence on the model output. The input interpretation saliency weights component of our metric template specifies how well the model captures the correlation between all words in a sentence and the output. By leveraging the overlap between task-related words and the input sentence, we can effectively optimize the correlations and improve the model's performance. Our proposed metric is designed to help models learn the correlations that are relevant to the task at hand while minimizing the impact of irrelevant words. By using the metric template, we can focus on three separate aspects, each aligned with one of the components: sign, iiw, and wordlists, allowing us to impact the learned correlations to align with our expectations about the relationship between a given category and word choice.

The focus of this chapter is on the extensive testing of a metric template designed for the task of emotion recognition. We had the option to choose between task coverage or usability. Task coverage involves testing multiple tasks on a single state-of-the-art (SoTA) dataset and model, while usability involves extensively evaluating the proposed metric for one subjective task of emotion, but testing it on multiple datasets and different model combinations. We chose the latter approach to validate the usability of the metric rather than the task coverage. Our approach involves selecting one task for learning, which is emotion, and one subtask for unlearning, which is demographics. Through this chapter, we iterate

over various combinations of this setup, thoroughly validating the utility of the metric for both cases, including emotion classification improvement and reducing encoded sensitive information in emotion classification models, while varying the datasets and other factors. Our results demonstrate the effectiveness of the proposed metric in improving emotion classification and producing models with reduced sensitive information leakage.

To ensure that our proposed metric aligns with annotator choices, we conduct a crowd-sourcing survey. The survey results indicate that our metric aligns well with annotator choices, thereby validating the effectiveness of our approach. The metric can be similarly applied to any other human-centered subjective tasks, such as, deception detection, opinion analysis etc.

9.3.1 Requirements

The proposed templates has three inputs: input interpretation saliency weights, direction of learning (sign) and relevant wordlists.

9.3.1.1 Input Interpretation Saliency Weights

The goal of input interpretation saliency weights is to understand what a trained black-box neural network is doing. This is achieved by examining the learned correlations from the model, which are obtained through a post-hoc method. By analyzing these correlations, we can gain insights into how the model is making its predictions and identify any potential leakage of sensitive information or limitations.

We calculate the reliance of a model on any given word in the input using the Captum interpretability library for PyTorch [124]. Captum provides state-of-the-art algorithms to identify how input features contribute to a model's output.

We use the attribution algorithms implemented via integrated gradients [211]. Integrated gradients represent the integral of gradients with respect to inputs along the path from a given baseline (absence of the cause) to input sample. The output is a set of words for each

model instance that contribute towards the prediction along with their attribution weights.

The sign of the attribution weight shows the direction of the correlation, and the value is the strength of the correlation. Words that do not influence prediction should have zero correlation.

9.3.1.2 Wordlists

A wordlist here refers to linguistic indicators extracted from the literature for a given task. The lists can either have singular words or phrases. A wordlist has three main properties: the variable it is created for, existence of label associations, and associated weights or importances. For example, if we want to measure 'emotion' then words such as 'excited', or 'lonely' are strong linguistic indicators, the former for positive emotion, the latter for negative emotion.

Similarly, if our measurement variable is 'age', then phrases such as 'cancel culture' or 'groovy' have been found influencing perception of age. If the aim is to evaluate 'unlearning' of information in a given model, we do not necessarily need to have a pairwise list (e.g., words related to females or males). Instead, a combined list will still allow us to mitigate the possibility of encoding sensitive variables in the learned representations. A wordlist can also have weights associated with a particular entry that correlate to strength of distinguishing ability of perception. For example, a wordlist associated with 'emotion' can have a weight of 4 for 'amazing' vs. 2 for 'good', the former being perceived as a stronger positive emotion as compared to the latter.

9.3.2 Terminology

First, let's consider a model trained for for a binary classification task ($[T]$), using dataset (D) of size $|D|$ with two output labels (L), A and B ($\{A, B\}$), i.e., ($[T] \rightarrow L_{\{A,B\}}$). We refer to samples in the dataset as s and individual words as x, and if a word occurs in both as, $x \in s$.

We create a wordlist (η) that has human-centered linguistic indicators corresponding to [T]. Depending on the nature of the task, the wordlist can either be combined, for both output labels, ($\eta_{L_{A+B}}$), or can have pairwise association with only one output label (e.g., η_{L_A} or η_{L_B}). While this does not necessarily need to be the case, combined wordlists can be used for reduced sensitive information leakage tasks (**HCM$^{\text{SIR}}$**), i.e., where the aim is to minimize the presence of information relevant to the sensitive variable. On the other hand, pairwise associative wordlists can be used for generalizability (**HCM$^{\text{gz}}$**) tasks, i.e., we need to learn and investigate associations between learned model correlations and the task outputs. The wordlist can also have weights, ω, associated with the individual entries (e.g., individual words, ω_x). If the associated weights are absent, all entries are assigned a weight of 1.

During prediction, the model outputs a saliency value (λ_x) for each word. We group words that have a positive ($\lambda_x > 0.05$), negative ($\lambda_x < -0.05$), or negligible ($|\lambda_x| \leq 0.05$) saliency attribution value. The set of words in a sample with positive saliency will be noted $+_x = \{x | \lambda_x > 0.05, for\ x \in s\}$ and negative saliency $-_x = \{x | \lambda_x < -0.05, for\ x \in s\}$ [1].

We describe how this information can be used at a sample-level (**HCM** : **s**) or of the dataset (**HCM**) as a whole to either promote generalizability or reduced sensitive variable information encoding:

1. **HCM$^{\text{gz}}$Emo** is a metric calculated over a dataset that is focused on generalizability (gz) in the task of emotion recognition (Emo). It can be evaluated in either a **standalone fashion** (stn, **HCM$^{\text{gz}}_{\text{stn}}$Emo**), in terms of performance on a dataset. It can also be calculated as the difference in performance between two models (relative improvement), (ri, **HCM$^{\text{gz}}_{\text{ri}}$Emo**).

2. **HCM$^{\text{sir}}$Dem** is a metric calculated over a dataset that is focused on sensitive information reduction (sir) with respect to a demographic variable (Dem, e.g., gender). It can also be evaluated in terms of either standalone performance or relative improvement.

---

[1]We do not specifically call out words with negligible saliency.

9.3.3 Metric Types and Calculation

The template enables two different types of model evaluations: (a) standalone (*stn*), and, (b) relative improvement (*ri*).

9.3.3.1 Standalone (HCM$_{stn}$)

We refer to standalone evaluation as benchmarking a model in isolation. We will first consider the case where we have combined wordlists (e.g., $\eta_{L_{A+B}}$). This condition would be used when we **do not have the class labels for a specific task in our setup**, as in the case for the sensitive variable of age. It allows us to identify words that are broadly associated with age in general, rather than a specific age grouping (e.g., older adults).

We define $\{\lambda_x\}_{x \in s}$ as the set of saliency values for all words in a given sample s. We then define an intersection set, which identifies words that occur both in the sample ($x \in s$) and in a wordlist ($\eta_{L_{A+B}}$). The intersection set is defined as:

$$int = \{x | x \in s\} \cap \eta_{L_{A+B}} \tag{9.1}$$

We can then derive the standalone metric (**HCM$_{stn}$**) from this intersection set for a given sample. We consider each word (x) in the intersection set defined by Equation 9.1. We weigh the saliency attribution for that word (λ_x) by the wordlist weight (ω_x) and sum this value over all words in intersection set for the sample. We normalize by the number of words in the intersection set (n_s). This gives us the sample-wise generalizability, **HCM$_{stn}^{gz}$[T] : s**, for sample, s (Equation 9.2).

$$\textbf{HCM}_{\textbf{stn}}^{\textbf{gz}}[\textbf{T}] : \textbf{s} = \frac{1}{n} \sum_{x \in int} (|\omega_x| \times |\lambda_x|) \tag{9.2}$$

We calculate the generalizability metric for a dataset by averaging over the number of samples in the dataset (Equation 9.3).

$$\mathbf{HCM^{gz}_{stn}[T]} = \frac{1}{|D|} \sum_{s \in D} \mathbf{HCM^{gz}_{stn}[T]} : \mathbf{s} \qquad (9.3)$$

We use $\mathbf{HCM^{gz}_{stn}[T]}$ to anticipate the generalizability of the model for a given task [T]. Higher values of λ_x for words with higher wordlist weights ω_x suggest that the model is relying on words that are known to be related to a given task (e.g., the word "thrilled" and positive valence).

We can use the same metric to anticipate the degree to which a model reduces encoded sensitive information by replacing $|\lambda_x|$ with $1 - |\lambda_x|$ (Equation 9.4) for sample s. Lower values of λ_x (and therefore high weights of $1 - |\lambda_x|$) for words with higher wordlist weights ω_x suggest that the model is not relying upon words that are known to be related to a demographic variable (e.g., the word "she" and the task of gender classification).

$$\mathbf{HCM^{sir}_{stn}[T]} : \mathbf{s} = \sum_x |\omega_x| \times (1 - |\lambda_x|) \qquad (9.4)$$

We again calculate the sensitive information reduction metric for a dataset by averaging over the number of samples in the dataset (Equation 9.5).

$$\mathbf{HCM^{sir}_{stn}[T]} = \frac{1}{|D|} \sum_{s \in D} \mathbf{HCM^{sir}_{stn}[T]} : \mathbf{s} \qquad (9.5)$$

Next, we consider the case where we **do have labels for a given sensitive variable** (e.g., gender). In this case, we calculate the value of the metric with respect to the pairwise association between the wordlist for a given label (e.g., η_{L_A}) and the model's output saliency attribution value for words within the sample ($\{\lambda_x\}_{x \in s}$).

We must consider four separate intersections, each of which evaluates the alignment between the directionality of the saliency attribution for a given word (i.e., positive vs. negative, $\lambda_{x,x \in s}$) and whether or not the word is in a given class A. For simplicity, we will discuss the metric applied to a single class[2], class A.

---

[2]The metric can be extended to multiclass problems.

We first define four intersections, using the notation: $(int_i, i \in \{1, 2, 3, 4\})$. We will discuss sets of words with certain properties:

1. Words with positive saliency and in wordlist, η_{L_A}:

   $int_1 = +_x \cap \{x | x \in \eta_{L_A}\}$

2. Words with negative saliency and *not* in wordlist, η_{L_A}:

   $int_2 = -_x \cap \{x | x \notin \eta_{L_A}\}$

3. Words with negative saliency and in wordlist, η_{L_A}:

   $int_3 = -_x \cap \{x | x \in \eta_{L_A}\}$

4. Words with positive saliency and *not* in wordlist, η_{L_A}:

   $int_4 = +_x \cap \{x | x \notin \eta_{L_A}\}$

Intuitively, we would like to encourage the model to maximize the number of words in the first and second intersection sets (int_1, int_2) and minimize the number of words in the third and fourth intersection sets (int_3, int_4). When this occurs, the model's behavior is aligned with the sociolinguistic wordlists.

We can then calculate the generalizability and sensitive information reduction metrics. For a given sample, we calculate either $\mathbf{HCM_{stn}^{gz}[T]} : \mathbf{s}$ (Equation 9.2) or $\mathbf{HCM_{stn}^{sir}[T]} : \mathbf{s}$ (Equation 9.4) for each of the four interaction sets. We calculate the metric by summing over weighting saliency values for matching sets and subtracting out the sum of weighted saliency values for the mismatched sets. We show the calculation of the generalizability metric here:

$$\mathbf{HCM_{stn}^{gz}[T]} : \mathbf{s} = \sum_{i=1}^{2} \sum_{x \in int_i} \omega_x \times \lambda_x - \sum_{i=3}^{4} \sum_{x \in int_i} \omega_x \times \lambda_x \tag{9.6}$$

We average over the samples within the dataset to obtain the dataset-level metric for generalizability metric as in Equation 9.3 and for the sensitive information reduction metric as in Equation 9.5.

9.3.3.2 Relative Improvement (HCM_{ri})

We refer to relative improvement evaluation as measuring how a model improves in a given task after adding an additional component (e.g., to reduce sensitive information encoded in a model). In aggregated level metrics such as accuracy, the relative improvement is measured by difference in performance aggregated over all sample's predictions. In our proposed method, we instead of focus on relative improvement per sample, as measured by the weights assigned to words/phrases. This is to measure the shift in the learnt associations of the model as compared to just the output predictions. As in the standalone case, we first calculate the relative improvement at the sample-level (e.g., $\textbf{HCM}_{\textbf{ri}}^{\textbf{gz}}[\textbf{T}] : \textbf{s}$ or $\textbf{HCM}_{\textbf{ri}}^{\textbf{sir}}[\textbf{T}] : \textbf{s}$). For example, we can capture the relative improvement in the generalizability of the new model, by calculating the standalone values of both the new and original models (each using Equation 9.2) and dividing by the standalone of the original:

$$\textbf{HCM}_{\textbf{ri}}^{\textbf{sir}}[\textbf{T}] : \textbf{s} = \frac{\textbf{HCM}_{\textbf{stn}}^{\textbf{sir}}[\textbf{T}]_{\textbf{new}} : \textbf{s} - \textbf{HCM}_{\textbf{stn}}^{\textbf{sir}}[\textbf{T}]_{\textbf{orig}} : \textbf{s}}{\textbf{HCM}_{\textbf{stn}}^{\textbf{sir}}[\textbf{T}]_{\textbf{orig}} : \textbf{s}} \tag{9.7}$$

We can then average this value over all samples within the dataset, to calculate the dataset-level relative improvement metric for generalizability as in Equation 9.3 and for the sensitive information reduction metric as in Equation 9.5. The final sensitive information reduction metric for relative improvement, moving from model *orig* (the original) to model *new* can be written as:

$$\textbf{HCM}_{\textbf{ri}}^{\textbf{sir}}[\textbf{T}] = \frac{1}{|D|} \sum_{s \in D} \textbf{HCM}_{\textbf{ri}}^{\textbf{sir}}[\textbf{T}] : \textbf{s} \tag{9.8}$$

9.3.4 Metrics for Measuring and Improving Emotion Recognition Model Generalizability

We use the sample-wise metric ($\textbf{HCM}_{\textbf{stn}}^{\textbf{gz}}\textbf{Emo} : \textbf{s}$, Equation 9.2) for model training. We use the dataset-level metric ($\textbf{HCM}_{\textbf{stn}}^{\textbf{gz}}\textbf{Emo}$, Equation 9.3) to quantify emotion recognition

performance. We empirically evaluate both the correlation and impact of including this metric on the performance of the model, both in within and cross-dataset model analyses.

We create wordlists using the valence annotations from National Research Council Canada - Valence, Arousal, and Dominance (NRC-VAD) Lexicon [159] consisting of 20,000 annotated unigrams. We normalize the annotated valence values between *-1,1* and bin those into 3 quantile categories *low, med, high* to match our model training label schema. This wordlist has information about the class label and the associated word. Therefore, we use the metric that leverages this information (Eq. 9.6).

9.3.5 Metrics for Relative Measurement and Improvement of Reducing Sensitive Information Encoded in Emotion Recognition Models

We use the sample-wise metric ($\mathbf{HCM_{stn}^{sir}Dem : s}$, Equation 9.5) for model training. We use the dataset-level metric ($\mathbf{HCM_{ri}^{sir}Dem}$, Equation 9.8) to quantify the sensitive information reduction efficacy.

We use various sociological studies to create wordlists (Section 9.3.1.2) and derive sensitive information reduction metrics corresponding to three demographic variables: gender ($\mathbf{HCM_{stn}^{sir}Gen : s}$), age ($\mathbf{HCM_{stn}^{sir}Age : s}$), and race ($\mathbf{HCM_{stn}^{sir}Race : s}$). For gender and race, we use (a) word lists that have been tested using the Implicit Association Test (IAT) [166] to measure population-level bias between the genders and (b) sociological studies [163] and the corresponding category based words from LIWC [212]. For age, we consider sociolinguistic literature for both, biological and social age [66], using word trends dataset [197], dictionary updates [222] and vocabulary grade levels [236].

9.4 Research Questions

RQ0-emo: Can we accurately recognize emotion, here labeled in terms of valence (positive vs. negative)? This baseline question demonstrates that our system is capable of recognizing the information of interest.

RQ0-SIR: Can we use representations trained in the context of emotion recognition to recognize a demographic variable, here gender[3]? We use this second research question to demonstrate that emotion representations encode information beyond the categories of interest, as shown in our prior work [108].

RQ1: How do metrics that capture sensitive information reduction ($\mathbf{HCM_{stn}^{sir}\,Dem}$) and generalizability ($\mathbf{HCM_{stn}^{gz}\,Emo}$) relating to reliance on emotionally-unrelated words vary across different sensitive information reduction methods? This research question provides a validation of the metrics and a way of measuring the effects of sensitive information reduction by itself.

RQ2: How can the $\mathbf{HCM_{stn}^{sir}\,Dem}$ metric be used in model training to decrease the leakage of gender information? This research question provides evidence that the metric can be effectively leveraged as a sensitive information reduction technique in model training.

RQ3: How does the inclusion of $\mathbf{HCM_{stn}^{gz}\,Emo}$ increase cross-dataset performance? This research question provides evidence that increasing reliance on emotional words, and decreasing reliance on non-emotional words, promotes better performance in unknown environments.

RQ4: How does cross-dataset performance change when we include both $\mathbf{HCM_{stn}^{gz}\,Emo}$ and $\mathbf{HCM_{stn}^{sir}\,Dem}$? Here, we investigate generalizable metrics that are based on knowledge of emotion expression and on spurious correlations between the perception of emotion and demographic categories (i.e., gender, age, race).

RQ5: For each of these tasks, emotion recognition and sensitive information reduction,

---

[3]In this chapter, we rely on binary markers of gender because these are the labels available within existing emotion corpora. Binary labels of gender to not align with how many individuals identify.

do humans' model preferences correlate with $\mathbf{HCM}_{\mathbf{stn}}^{\mathbf{gz}}\mathbf{Emo}$ and $\mathbf{HCM}_{\mathbf{ri}}^{\mathbf{sir}}\mathbf{Dem}$, respectively, when considering the sample, model's prediction and salient explanations?

9.5   Methods

In this section, we provide an overview for the all the baseline models that we will evaluate with respect to **HCM**.

9.5.1   Emotion Recognition

9.5.1.1   Baseline Emotion Recognition Model: Base

We use the base version of Bidirectional Encoder Representations from Transformers (BERT) model due to the prevalence of this approach [63]. We use a pre-trained BeRT tokenizer for the model. We implement and fine-tune the model using the HuggingFace library [230].

9.5.1.2   Multi-Dataset Training: Multi-D

Our goal is to create an emotion recognition model that is not biased by dataset. To do this, we start with the same model described above (**Base**). We introduce an additional adversarial task, recognizing the dataset (e.g., IEMOCAP, Section 3.1), following the domain adversarial networks method suggested in [78]. We train our model on combinations of two datasets and test on the third.

9.5.2   Sensitive Information Reduction

Our goal is to reduce a given model's ability to detect a sensitive attribute (e.g., gender) from embeddings learned for the task of emotion recognition. In all cases, the models are initialized with the **Base** model.

We measure sensitive information reduction success using Sensitive Information Leakage Measurement [152]. It assumes that an adversary has black-box access to the representations learned from the emotion recognition algorithm and access to a gender labeled subset of the dataset. The adversary uses these representation to train an auxiliary model to predict gender for any representation from that model, as described in our prior work [108].

9.5.2.1 Sensitive Information Reduction by Adversarial Training: SIRAdv

SIRAdv uses an adversarial training paradigm to reduce the sensitive information encoded in the generated embeddings with respect to gender. The main network is trained to unlearn gender using a Gradient Reversal Layer (GRL) [77], a multi-task approach to train models invariant to specific properties [153]. We place the GRL between the embedding sub-network and gender classifier to obtain gender-invariant representations.

9.5.2.2 Sensitive Information Reduction by Data Augmentation: SIRAug

SIRAug is is trained using an augmented data set using gender-swapping to compare our proposed method and resultant metric to other successful approaches [104]. We use a pronoun-based word list and create a gender-swapped equivalent for each sentence, e.g., replacing "he" with "she", "his" with "hers", and so on. Data augmentation has its own issues: it doubles the training data size with no added label information, has expensive list creation of gender-based words in the dataset to be replaced, and nonsensical sentence creation [22].

9.5.2.3 Sensitive Information Reduction by Bias Fine-Tuning: SIRBias

SIRBias leverages an additional outside dataset that is has reduced correlations with respect to the sensitive variable, which is used to train an initial model. We use the improvised turns subset of MSP-Improv (see Section 3.1). We chose this subset because both actors (the male and female actor) perform improvisations using the same initial set of scenarios,

controlling the word choice due to topic variations. Therefore, the manner and speaking and word choice is more likely to be related to gender differences, rather than topic differences. We then fine tune a BERT model to be unable to distinguish between genders for a particular improvisational target. This clusters samples together by prompt, accounting for topic-based word variations. The result of this fine-tuning should ideally be a model with reduced leakage of gender information. This resulting model is then finetuned for emotion recognition on IEMOCAP (see Section 3.1).

9.5.3 Helper Model

9.5.3.1 Gender Control: GenControl

Gender Control is a sanity check for verifying the reliability of our template. We hypothesize that if our proposed method indeed captures relevant information, then a model trained specifically to recognize gender should have the lowest $\mathbf{HCM_{ri}^{sir}Gen}$ value. We train a multi-task model for predicting both gender and emotion.

9.5.3.2 Artificially Noisy Model: ArtNoise

We create a parallel corpus of IEMOCAP to use as an attention check for the crowdsourced task. We add six artificial "noisy" features, {'zq0', 'zq1', 'zq2', 'zx0', 'zx1', 'zx2'}, such that they correlate specifically with both emotion and gender and train a classification model to predict emotion on this dataset. The added signals are 100% correlated to the sub-classes, ensuring that the model always learns these added nonsensical tokens as salient features.

9.6 Experimental Setup

9.6.1 Models for Metric Correlation To Performance

We perform all training experiments three times, using one dataset for training at a time, and testing on the same dataset (within-corpus testing) as well as the other two (cross-corpus

testing). We train six models as described in Section 9.5.1 and Section 9.5.2: two solely for emotion recognition {**Base**, **Multi − D**}, three for reduced sensitive information encoded emotion recognition {**SIRAdv**, **SIR**, **Aug**, **SIRBias**}, and one control {**GenControl**}.

9.6.2 Using the Metrics to Train Generalizable and Sensitive Information Reduced Models

We evaluate whether including the proposed metric in the training process improves emotion recognition and sensitive information reduction performance. We initialize the model with the baseline emotion recognition model (**Base**, Section 9.5.1.1). We retrain the model with the approaches discussed in Section 9.5.2. Please see Figure 9.1 for a pictorial example for how metrics are integrated within the training process. We consider five different training conditions:

1. Only $\mathbf{HCM_{stn}^{gz}Emo}$: **s**

2. Cross-entropy + $\mathbf{HCM_{stn}^{gz}Emo}$: **s**

3. Only $\mathbf{HCM_{stn}^{sir}Dem}$: **s**

4. Adversarial component + $\mathbf{HCM_{stn}^{sir}Dem}$: **s**

5. Cross-entropy + adversarial component + $\mathbf{HCM_{stn}^{gz}Emo}$: **s** + $\mathbf{HCM_{stn}^{sir}Dem}$: **s**

We also perform experiments with each demographic metric on its own and other permutations of the setup.

9.6.3 Crowdsourcing

Finally, we explore whether user preference aligns with our proposed metrics. We select a representative sample population by considering three levels of: 1) valence (low, medium, and high, see Section 3.2.1) and 2) the metrics of $HCM_{stn}^{sir}Gnd$: s (low, medium, and high, measured via quantiles). This results in nine bins.
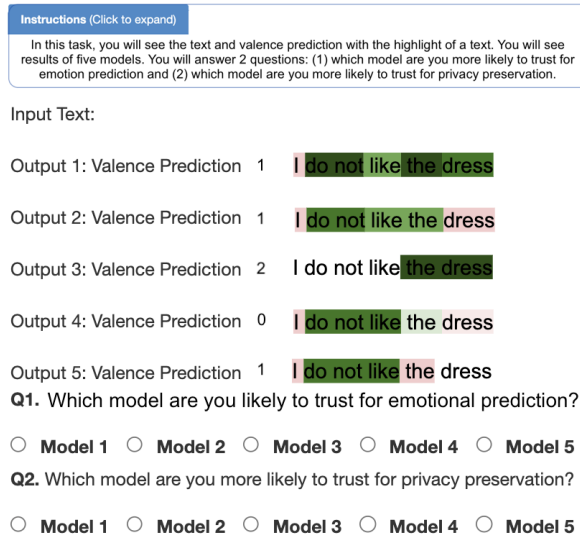
Figure 9.2: The crowdsourcing interface. Users are asked to indicate their model preference given visualizations of the relative importance of specific words in the prediction of emotion. The evaluators were not presented with the names of the models and the order of model outputs was randomized for each viewing. Note that the sentence was chosen to clearly convey gender, and potentially, gendered information. In the figure, green indicates that the model is using the word in the prediction, while red indicates that it is not. The outputs correspond to: 1) GenControl - note the heavy focus on the word "dress", 2) SIRBias, 3) ArtNoise, 4) SIRAug, 5) SIRAdv (for model details, see Section 9.5.2). Observe that the three sensitive information reduction methods are no longer focusing on the word "dress".

We select 60 random samples from each bin for a total 540 samples. We classify each sample using our five models (**SIRAdv**, **SIRAug**, **SIRBias**, **GenControl**, **ArtNoise**, described above). We extract sample-level predictions. In addition, we extract sample-level explanations using Captum [211], Section 9.3.1.1).

ArtNoise is used as a attention check baseline for human evaluation, ensuring that the evaluators are paying attention to the task at hand. We consider this option as an attention check and discard any response where the crowdsourced worker preferred the result and/or the explanation from this model (7.31% samples were discarded in total).

We presented the predictions from the models and heat map-based explanations. We asked the workers to choose between the five models, focusing on the following questions: (1) which model are you more likely to trust for emotion prediction and (2) which model are you more likely to trust for not being able to predict sensitive information.

We recruited annotators using Prolific from a population of workers with the following characteristics: 1) in the United States, and 2) native English speakers. Each task was annotated by three workers. We used qualification tests to check that all workers understood valence. Each task took an average of one-minute. The compensation was $9.45/hr.

9.7 Results and Discussion

9.7.1 RQ0-emo: Can we accurately recognize emotion, here labeled dimensionally for valence?

The baseline emotion classification model (**Base**) for within-corpus valence classification has 0.65 UAR for IEMOCAP, 0.68 for MuSE, and, 0.54 for MSP-Podcast (chance is 0.33 UAR), which is comparable to existing approaches on this domain [9].

In the cross-corpus setting, we see an average drop of at least 15% on all the models and datasets permutations. For example, a model trained on IEMOCAP and tested on MuSE has a UAR of 0.58 for **Base**, and, 0.53 for the **SIRAdv** model, which is significantly lower (paired t-test adjusted for multiple comparisons, p=0.0017) than the performance of these models when trained and tested on MuSE (0.68 and 0.71 respectively).

The multi-dataset emotion classification model has an overall higher performance in the cross-corpus setting. For example, the **Multi-D** has 0.64 UAR for IEMOCAP as compared to 0.61 and 0.57 obtained using **Base**.

We find that while reducing sensitive information leakage reduces within-corpus emotion recognition performance in many cases, these sensitive information leakage reduced models, especially **SIRAdv** usually have equivalent or better performance as compared to **Base** in cross-corpus setting, e.g., **Base** and **SIRAdv** trained on MuSE and tested on IEMOCAP has a UAR of 0.61 and 0.63, respectively.

9.7.2 RQ0-SIR: Can we use representations trained in the context of emotion recognition to recognize a demographic variable, here gender?

Gender classification within dataset is highest for models using the emotion representations learned from the baseline model and the control gender-addition model, supporting prior work [108]. The multi-dataset model usually has a lower or equivalent gender UAR as compared to the baseline model. Gender classification operating on the representations learned from all three sensitive information reduction models have lower UAR, compared to the multi-dataset model.

The leakage of gender information is higher for all models in the cross-corpus condition, compared to within corpus testing, excepting the multitask emotion-gender model. This suggests some loss of efficacy of the sensitive information reduction methods in cross-corpus settings. In almost all the cases, **SIRAdv** has the lowest cross-corpus gender leakage while almost retaining emotion recognition performance. Hence, we use **SIRAdv** as our main model for retraining purposes (please see results tables in appendix for full details).

9.7.3 RQ1: How do metrics that capture sensitive information reduction, $\textbf{HCM}_{\textbf{ri}}^{\textbf{sir}}\textbf{Dem}$, and reliance on words unrelated to emotion, $\textbf{HCM}_{\textbf{stn}}^{\textbf{gz}}\textbf{Emo}$, correlate to a model's emotion recognition and sensitive information reduction performance?

We expect that models that effectively reduce leakage of sensitive information the model have high $\textbf{HCM}_{\textbf{ri}}^{\textbf{sir}}\textbf{Dem}$ values, while models that do not will have low $\textbf{HCM}_{\textbf{stn}}^{\textbf{gz}}\textbf{Emo}$ values.

We initially validate the assumption that the $\textbf{HCM}_{\textbf{ri}}^{\textbf{sir}}\textbf{Gen}$ metric is very low for the GenControl model (-0.21). This aligns with our expectation that the relative sensitive information reduction should be low. The goal of the GenControl model is to *learn*, rather than ignore, gender identity.

For within corpus settings, our expectation is that the SIRAug and SIRAug models will have the highest $\textbf{HCM}_{\textbf{ri}}^{\textbf{sir}}\textbf{Gen}$ values. We find that across all the datasets, and models, gender leakage is significantly negatively correlated (paired t-test adjusted for multiple comparisons)

154

with $\mathbf{HCM^{sir}_{ri}Gen}$ for both within corpus (-0.95^*) and cross-corpus (-0.89^*) evaluation. ($*$ refers to significant value when using paired t-test adjusted for multiple comparisons)

We expect that models that generalize well will have high $\mathbf{HCM^{gz}_{stn}Emo}$ values, while models that generalize poorly will have low $\mathbf{HCM^{gz}_{stn}Emo}$ values. We find that in two out of three cases, **SIRAdv** has a higher $\mathbf{HCM^{gz}_{stn}Emo}$ than **Base**. We find that across all the datasets, and models, emotion recognition performance is significantly positively correlated (paired t-test adjusted for multiple comparisons) with $\mathbf{HCM^{gz}_{stn}Emo}$ for both within corpus (0.72^*), and cross-corpus (0.80^*) evaluation.

We analyze the intersections between word-lists and saliency attribution. We consider instances in which $\mathbf{HCM^{gz}_{stn}Emo}$ positively impacts model performance, when a sample moves from being incorrectly to correctly classified after $\mathbf{HCM^{gz}_{stn}Emo}$ is included in model training. We observe that in 87.3% of the words in the intersections (Equation 9.1), in the better performing model have medium valence label (from the set: low, medium, and high). This suggests that the metric enables the model improve in the recognition of subtle displays of emotion.

9.7.4   RQ2: How does additionally optimizing for $\mathbf{HCM^{sir}_{stn}Gen}$ affect gender information leakage?

We observe that when using the $\mathbf{HCM^{sir}_{stn}Gen}$ metric as the only optimizer, we fall short of the sensitive information reduction performance of **SIRAdv**, indicating need for composite metrics. We find that augmenting **SIRAdv** training with $\mathbf{HCM^{gz}_{stn}Gen}$: **s** improves valence and decreases gender classification performance (0.64 in **SIRAdv** vs 0.61 with **SIRAug**) in cross-corpus settings (Table 9.1). In within corpus settings, we find that the addition of metric still benefits sensitive information reduction but can also lead to a non-negligible drop in emotion recognition performance. We also find that while age ($\mathbf{HCM^{sir}_{stn}Age}$) and race ($\mathbf{HCM^{sir}_{stn}Race}$) metrics on their own do not signify any positive change in cross-corpus dataset performance.

Table 9.1: Training models with only $\mathbf{HCM_{stn}^{gz}\,Gen:s}$

| Training | Testing | | |
	IEMOCAP	MuSE	MSP-Improv
IEMOCAP	0.64	0.60	0.45
MuSE	0.63	0.67	0.41
MSP-Podcast	0.60	0.62	0.52

We find that the metrics can be jointly considered in a simple multi-task setup between the primary task (i.e., emotion) and an adversarial component ($\mathbf{HCM_{stn}^{sir}\,Gen}$). The multi-task setup results in the best performing gender information reduction (we do not have labels to measure age/race) of emotion recognition model (Table 9.2).

9.7.5 RQ3: How does additionally optimizing for $\mathbf{HCM_{stn}^{gz}\,Emo}$ affect emotion recognition?

We find that optimizing for alignment with the human-centered metric leads to better cross-corpus performance across every setup (Table 9.1). For example, when we augment the **Base** model trained on IEMOCAP with $\mathbf{HCM_{stn}^{gz}\,Emo}$. The UAR increases to 0.6 and 0.45 for Muse and MSP-Podcast respectively, as compared to 0.58 and 0.43. Additionally, we find that, training the model with these metrics give us better or equivalent performance to when using the combined dataset method, which was the best solely emotion recognition model. This shows the effectiveness of the metric, and lowers the required computation time and power.

We also see that while we see a performance increase for cross-dataset testing, some within dataset performances are lower than the generally trained model (e.g., **Base** has UAR 0.54 for MSP-Podcast, but **Base+$\mathbf{HCM_{stn}^{gz}\,Emo}$** has a UAR 0.52). We hypothesize that this occurs because the models were relying on spurious correlations (sprurious, with respect to emotion recognition), which, when removed, decreased the model's performance.

Table 9.2: Training models with composite metrics: $\mathbf{HCM_{stn}^{gz}\,Gen}$: $\mathbf{s}$ + $\mathbf{SIRAdv}$(Gender) + $\mathbf{HCM_{stn}^{sir}\,Dem}$: s ($\mathbf{HCM_{stn}^{sir}\,Gen}$: s + $\mathbf{HCM_{stn}^{sir}\,Age}$: s + $\mathbf{HCM_{stn}^{sir}\,Race}$: s)

| Training | Testing | | |
	IEMOCAP	MuSE	MSP-Improv
IEMOCAP	0.71	0.59	0.45
MuSE	0.67	0.70	0.46
MSP-Podcast	0.57	0.56	0.50

9.7.6 RQ4: Does emotion recognition and sensitive information reduction performance change when including both $\mathbf{HCM_{stn}^{gz}\,Emo}$ and $\mathbf{HCM_{stn}^{sir}\,Dem}$ during training?

We include all sensitive information reduction components ($\mathbf{HCM_{stn}^{sir}\,Gen}$, $\mathbf{HCM_{stn}^{sir}\,Age}$, $\mathbf{HCM_{stn}^{sir}\,Race}$) and the emotion generalizability component ($\mathbf{HCM_{stn}^{gz}\,Emo}$) in multi-task setup method in model training. We find that this composite metric (combination of all metrics: $\mathbf{HCM_{stn}^{sir}\,Gen}$: s + $\mathbf{HCM_{stn}^{sir}\,Age}$: s + $\mathbf{HCM_{stn}^{sir}\,Race}$: s) leads to better model performance across the board. For example, models trained with the composite metric improves the UAR to 0.7 on MuSE within-corpus, as compared to 0.68 from the **Base** model).

We find that the two components interact to give better performance on both cross-corpus emotion recognition and sensitive information reduction performance. For example, we see an improved UAR of 0.67 on IEMOCAP (cross-corpus), compared to 0.62 from the **Base** method). Further, we also see that the gender leakage universally reduces when training the models on this combination (cross-corpus testing on MSP-Podcast when using composite metric training on Muse improves performance by 6%). We hypothesize that encouraging certain generalizable correlations while discouraging spurious correlations leads to models that are more robust and have reduced sensitive information leakage across the spectrum, but the enforcement of specific 'unlearning' often leads to negligible drops in within-corpus sensitive information reduction performance.

Finally, we found that inclusion of age and race, to gender, as sensitive information

reduction components improves emotion recognition performance in cross corpus evaluation (Please see the full results table in the appendix).

9.7.7 RQ5: For each of these tasks, emotion recognition and sensitive information reduction, do humans' model preferences correlate with $\text{HCM}^{gz}_{stn}\text{Emo}$ and $\text{HCM}^{sir}_{ri}\text{Dem}$ respectively, when considering the sample, model's prediction and salient explanations?

We show that user preference depends on what the user was asked to consider the model's behavior for, with respect to emotion or to sensitive information leakage [103]. The user preference for emotion recognition models is significantly correlated with within-corpus valence recognition performance (0.78) and with the $\text{HCM}^{gz}_{stn}\text{Emo}$ metric (0.71).

The correlation between user preference for emotion recognition and the $\text{HCM}^{sir}_{ri}\text{Gen}$ metric is negligible (0.09). This suggests that when asked to focus on emotion recognition performance with models that have the same prediction but different highlighted words, both performance and the potential for generalizability (when grasped using the highlighted words/phrases for attributed by the model for the prediction) dominate user preference.

When choosing sensitive information leakage reducing model, we find that user's preference is moderately correlated with emotion recognition performance (0.46) and strongly negatively correlated with the ability to predict gender (-0.86). This is also supported by the correlation between user preference and $\text{HCM}^{sir}_{ri}\text{Gen}$ (0.64) and $\text{HCM}^{gz}_{stn}\text{Emo}$ (0.81). This support the hypothesis that, human perception can be encoded into models through the proposed metric template ($\text{HCM}[\text{T}]$).

9.8 Conclusion

In this chapter, we present a novel set of emotion generalizability metrics focused on reducing sensitive information encoding, and reliance on emotionally-unrelated words. We evaluate these metrics across the tasks of emotion recognition and classification of

158

demographic variables (here, gender). We demonstrate that these metrics are correlated with the performance of algorithms, both within and across corpus, and that these metrics can be used during model training to improve classification performance and leakage of gender information. In future work, we aim to look at finer grained controls over reducing learnt sensitive information when evaluated by the end user, and methods to incorporate user-specific personalization of the baseline metric.

Please see the table on the following page that provides a detailed accounting for the results discussed in the rest of the chapter.

Table 9.3: Within and cross dataset performance of models across various datasets, correlation with these metrics and using metrics for training. Note that for space, the column headings refer to H, rather than HCM. We use the notation "Dem" to refer to the combined demographic categories of Gender, Age, and Race. SIR refers to sensitive information reduction.

Train	Model	Valence			$H_{stn}^{gz}E$	Gender			$H_{ri}^{sir}G$	$1- H_{ri}^{sir}G$
		IEM	MuSE	MSP-P		IEM	MuSE	MSP-P		
IEMOCAP	Baseline	0.65	0.58	0.43	0.64	0.71	0.74	0.68	-	
	SIRAdv	0.67	0.53	0.38	0.62	0.56	0.62	0.59	0.65	0.35
	SIRAug	0.64	0.59	0.44	0.68	0.55	0.64	0.55	0.68	0.32
	SIRBias	0.62	0.53	0.38	0.6	0.62	0.69	0.7	0.56	0.44
	GenControl	0.63	0.52	0.37	0.59	0.82	0.76	0.76	-0.21	1.21
	HCM_{stn}^{sir} Gen	0.64	0.56	0.4	0.61	0.58	0.61	0.6	0.6	0.4
	HCM_{stn}^{sir} Age	0.65	0.51	0.38	0.63	-	-	-	0.32	0.68
	HCM_{stn}^{sir} Race	0.63	0.52	0.39	0.65	-	-	-	0.51	0.49
	HCM_{stn}^{sir} Dem	0.65	0.54	0.41	0.64	-	-	-	0.63	0.37
	SIRAdv+**HCM_{stn}^{sir} Gen**	0.69	0.56	0.43	0.67	-	-	-	0.74	0.26
	SIRAdv+**HCM_{stn}^{sir} Dem**	0.67	0.58	0.44	0.66	0.54	0.6	0.52	0.72	0.28
MuSE	Baseline	0.61	0.68	0.38	0.65	0.67	0.57	0.72	-	
	SIRAdv	0.63	0.71	0.36	0.67	0.58	0.53	0.56	0.67	0.33
	SIRAug	0.57	0.62	0.34	0.6	0.6	0.51	0.59	0.61	0.39
	SIRBias	0.56	0.6	0.34	0.59	0.62	0.55	0.63	0.59	0.41
	GenControl	0.4	0.53	0.3	0.52	0.8	0.88	0.72	-0.17	1.17
	HCM_{stn}^{sir} Gen	0.58	0.62	0.34	0.6	0.65	0.54	0.58	0.58	0.42
	HCM_{stn}^{sir} Age	0.6	0.62	0.35	0.58	-	-	-	0.21	0.79
	HCM_{stn}^{sir} Race	0.57	0.61	0.34	0.59	-	-	-	0.33	0.67
	HCM_{stn}^{sir} Dem (all)	0.6	0.64	0.38	0.62	-	-	-	0.55	0.45
	SIRAdv+**HCM_{stn}^{sir} Gen**	0.66	0.7	0.44	0.73	-	-	-	0.7	0.3
	SIRAdv+**HCM_{stn}^{sir} Dem**	0.64	0.69	0.4	0.7	0.55	0.52	0.55	0.69	0.31
MSP Pod-cast	Baseline	0.57	0.58	0.54	0.53	0.65	0.55	0.72	-	
	SIRAdv	0.54	0.48	0.5	0.42	0.54	0.51	0.57	0.68	0.32
	SIRAug	0.52	0.43	0.48	0.38	0.59	0.52	0.59	0.64	0.36
	SIRBias	0.44	0.42	0.48	0.38	0.61	0.54	0.61	0.53	0.47
	GenControl	0.37	0.38	0.4	0.32	0.78	0.7	0.79	-0.19	1.19
	HCM_{stn}^{sir} Gen	0.44	0.46	0.45	0.37	0.59	0.6	0.58	0.5	0.5
	HCM_{stn}^{sir} Age	0.46	0.48	0.47	0.38	-	-	-	0.45	0.55
	HCM_{stn}^{sir} Race	0.47	0.5	0.47	0.39	-	-	-	0.5	0.5
	HCM_{stn}^{sir} Dem	0.47	0.51	0.48	0.4	-	-	-	0.55	0.45
	DemAdv+**HCM_{stn}^{sir} Gen**	0.54	0.56	0.57	0.55	-	-	-	0.69	0.31
	DemAdv+**HCM_{stn}^{sir} Dem**	0.52	0.55	0.53	0.51	0.53	0.52	0.52	0.63	0.37

CHAPTER X

Concluding Remarks

The objective of this dissertation is to investigate the effect of implicit decisions made during the machine learning (ML) process on the performance of ML models trained for emotion recognition. Implicit decisions refer to the choices made by practitioners during the various stages of the ML pipeline, such as data preprocessing, feature selection, and model selection. These decisions are often made based on intuition, experience, or convenience, rather than a rigorous analysis of their impact on the final model. Chapters IV and Chapter VI of this dissertation focus on identifying key areas for dataset creation and annotation that can significantly impact the evaluation and performance of trained machine learning (ML) models. Chapters VI and VII of this dissertation are dedicated to investigating the impact of confounding factors in data on the performance and reliability of machine learning (ML) models. Chapters VIII and IX of this dissertation are dedicated to addressing two critical challenges in the development and deployment of machine learning (ML) models: unintentional encoding of demographic and memership information in model representations and generalizability of these models.

10.1   Summary of Contributions

- **Chapter IV** presented the following contributions:

- We unveiled a unique dataset, Multimodal Stressed Emotion (MuSE), purposed for studying the interaction between stress and emotion in speech.

- We elucidated the procedure for data collection, the potential applications, and the methodology for annotating emotional content.

- We showcased the dataset's potential in the development and evaluation of models focused on emotion recognition and stress detection.

- **Chapter V** brought forth:

  - Augmenting IEMOCAP with realistic noisy samples through the use of various types of environmental and synthetic noise.

  - An evaluation of how noise influences both the ground truth and predicted labels with associated guidance for noise-based augmentation within speech emotion datasets.

  - A strong emphasis on the need to consider the influence of noise on human emotion perception and the necessity of robustly testing models.

- **Chapter VI** provided insights on:

  - How context influences the annotation of emotional content within the MuSE dataset.

  - The comparison between two labeling methods, randomized and contextualized, and the discovery that contextualized labeling produces annotations closer to self-reported labels.

  - The finding that labels generated using the randomized method can be more efficiently predicted by automated systems.

- **Chapter VII** offered:

- A new method for separating stress modulations from emotion representations utilizing adversarial networks.

- Evidence showing that controlling for stress during training boosts the generalizability of emotion recognition models across new domains.

- An effective demonstration of our unique approach's usability using the MuSE dataset and its potent applicability to other confounding variables present in emotion datasets.

- **Chapter VIII** achieved:

  - A comprehensive analysis of demographic leakage in representations obtained from textual, acoustic, and multimodal data when trained for emotion recognition.

  - The introduction of an adversarial learning paradigm to reduce demographic leakage from generated representations.

  - A valid demonstration of our novel approach's efficacy on numerous datasets and its potential use to defend against set-based membership identification.

- **Chapter IX** proposed:

  - Automatic and quantifiable metrics for measuring generalizability and sensitive information encoding in the representation of machine learning models within the context of speech emotion recognition.

  - A verification method of the utility of the proposed metrics using crowdsourcing, showcasing their adaptability for evaluating cross-corpus generalization.

  - A cost-efficient and dependable way of appraising the effectiveness of machine learning models within speech emotion recognition.

These points, in sum, illustrate the specific contributions this dissertation makes to improve understanding, methodologies, and tools in the area of emotion recognition research.

10.2 Future Work

Building upon the contributions of this dissertation, there are two primary areas where future research could focus.

Firstly, when using interpretation and explanation, whether for emotion recognition or other human centered machine learning tasks, future research should focus on developing a standardized evaluation checklist for deep learning model explanations. Currently, the field is marked by inconsistent evaluation methodologies in various research papers, making it challenging to compare results. It is essential to address these variations by designing a checklist that encapsulates evaluation criteria applicable across different studies. The introduction of such a tool will help simplify the evaluation process and foster a clearer understanding and comparison of diverse research outcomes. The current environment lacks a standardized approach, and individual studies often adopt unique methods, which detracts from the comparability of research. Creating a universal evaluation framework is thus a crucial step for future advancements in the field.

Secondly, in terms of evaluation metrics for any human centered tasks, the use of closed language models (LLMs) poses certain challenges when it comes to testing their performance. If we do not know what went into the training dataset of a closed LLM, it is very likely that the test dataset we use or generate will be contaminated. This can result in inaccurate performance metrics and difficulty in verifying the correctness of the model's responses. Prompt-based benchmarking can exacerbate these challenges further by introducing another layer of complexity. If we are not directly testing for an "input sample" to "output" correspondence, the way we write the prompt and what we ask for as an output will significantly change how we benchmark the ability of any model. This can make evaluating the performance of a closed LLM more difficult and error-prone, and increase the risk of producing biased results. Future work can look into Prompt Based Experimental Design Schemas. These schemas can be thought of as templates, but with a key distinction. They not only focus on varying the input sample, but also on modifying the prompt specifications.

164

By doing so, they allow us to closely monitor and analyze the changes that occur from one prompt to another. This approach provides a valuable way to study and understand the impact of different prompts on the overall outcome of an experiment.

In essence, these proposed trajectories, in tandem with this dissertation's contributions, provide a roadmap for further research in the field of models trained for human centered tasks such as emotion recognition. The continuance of these research undertakings will augment our understanding and enhance the effectiveness of the techniques used in the recognition and evaluation of human centered predictive models.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Phonological history of english consonant clusters= `https://bit.ly/3l08zbu`.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proc. 2016 ACM Conference CCS*.

[3] Mohammed Abdelwahab and Carlos Busso. Incremental adaptation using active learning for acoustic emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5160–5164. Ieee, 2017.

[4] Mohammed Abdelwahab and Carlos Busso. Domain adversarial for acoustic emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2423–2435, 2018.

[5] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin RB Butler, and Joseph Wilson. Practical hidden voice attacks against speech and speaker recognition systems. *arXiv preprint arXiv:1904.05734*, 2019.

[6] Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Analyzing thermal and visual clues of deception for a non-contact deception detection approach. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, page 35. Acm, 2016.

[7] Mohamed Abouelenien, Veronica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. Deception detection using a multimodal approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, Icmi '14, pages 58–65, New York, NY, USA, 2014. Acm.

[8] Ana Aguiar, Mariana Kaiseler, Mariana Cunha, Hugo Meinedo, Pedro R Almeida, and Jorge Silva. Voce corpus: Ecologically collected speech annotated with physiological and psychological stress assessments. In *LREC 2014, Ninth International Conference on Language Resources and Evaluation*, 2014.

[9] Zakaria Aldeneh, Soheil Khorram, Dimitrios Dimitriadis, and Emily Mower Provost. Pooling acoustic and lexical features for the prediction of valence. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 68–72. Acm, 2017.

[10] Zakaria Aldeneh and Emily Mower Provost. Using regional saliency for speech emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2741–2745. Ieee, 2017.

[11] Arthur Aron, Edward Melinat, Elaine N Aron, Robert Darrin Vallone, and Renee J Bator. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, 23(4):363–377, 1997.

[12] Nicolas Audibert, Véronique Aubergé, and Albert Rilliard. Prosodic correlates of acted vs. spontaneous discrimination of expressive speech: a pilot study. In *Speech Prosody 2010-Fifth International Conference*, 2010.

[13] James R Averill. A constructivist view of emotion. In *Theories of emotion*, pages 305–339. Elsevier, 1980.

[14] Eugene Bagdasaryan and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *arXiv preprint arXiv:1905.12101*, 2019.

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

[16] Ntombikayise Banda and Peter Robinson. Noise analysis in audio-visual emotion recognition. In *Proceedings of the 11th International Conference on Multimodal Interaction (ICMI)*, 2011.

[17] Tanja Bänziger, Georg Hosoya, and Klaus R Scherer. Path models of vocal emotion communication. *PloS one*, 10(9):e0136675, 2015.

[18] Michael Barbaro, Tom Zeller, and Saul Hansell. A face is exposed for aol searcher no. 4417749. *New York Times*, 2006.

[19] Murray R Barrick and Michael K Mount. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26, 1991.

[20] Ellen Elizabeth Bartolini. Eliciting emotion with film: Development of a stimulus set. 2011.

[21] Anton Batliner, Ralf Kompe, A Kießling, E Nöth, and H Niemann. Can you tell apart spontaneous and read speech if you just look at prosody? In *Speech Recognition and Coding*, pages 321–324. Springer, 1995.

[22] Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

[23] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[24] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

[25] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[26] Rebecca H Berger, Alison L Miller, Ronald Seifer, Stephanie R Cares, and Monique K LeBourgeois. Acute sleep restriction effects on emotion responses in 30-to 36-month-old children. *Journal of sleep research*, 21(3):235–246, 2012.

[27] Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1042–1047, 2016.

[28] Nadia Bianchi-Berthouze and Christine L Lisetti. Modeling multimodal expression of user's affective subjective experience. *User modeling and user-adapted interaction*, 12:49–84, 2002.

[29] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 2016.

[30] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[31] Pascal Bruno, Valentyna Melnyk, and Franziska Völckner. Temperature and emotions: Effects of physical temperature on responses to emotional advertising. *International Journal of Research in Marketing*, 34(1):302–320, 2017.

[32] Tony W Buchanan, Jacqueline S Laures-Gore, and Melissa C Duff. Acute stress reduces speech fluency. *Biological psychology*, 97:60–66, 2014.

[33] Alec Burmania, Mohammed Abdelwahab, and Carlos Busso. Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5190–5194. Ieee, 2016.

[34] Alec Burmania and Carlos Busso. A stepwise analysis of aggregated crowdsourced labels describing multimodal emotional behaviors. In *Interspeech*, pages 152–156, 2017.

[35] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388, 2016.

[36] 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, 2008.

[37] Carlos Busso, Murtaza Bulut, and Sungbok Lee. Shrikanth narayanan fundamental frequency analysis for speech emotion processing. *The role of prosody in affective speech*, 97:309, 2009.

[38] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2017.

[39] Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and RBV Subramanyam. Benchmarking multimodal sentiment analysis. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 166–179. Springer, 2017.

[40] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Cvpr*, 2017.

[41] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX$}$*, 2019.

[42] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*.

[43] Richard T Cauldwell. Where did the anger go? the role of context in interpreting emotion in speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[44] Hervé Chabanne, Amaury de Wargny, Jonathan Milgram, Constance Morel, and Emmanuel Prouff. Privacy-preserving classification on deep neural network. *IACR Cryptology ePrint Archive*, 2017.

[45] Tara M Chaplin. Gender and emotion expression: A developmental contextual perspective. *Emotion Review*, 2015.

[46] Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, and Shrikanth Narayanan. Data augmentation using gans for speech emotion recognition. In *Interspeech*, pages 171–175, 2019.

[47] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao K. Huang, and Lun-Wei Ku. Emotionlines: An emotion corpus of multi-party conversations. *CoRR*, abs/1802.08379, 2018.

[48] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*, 2018.

[49] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021.

[50] Farah Chenchah and Zied Lachiri. Speech emotion recognition in noisy environment. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 788–792. Ieee, 2016.

[51] François Chollet. keras. `https://github.com/fchollet/keras`, 2015.

[52] François Chollet. keras. `https://github.com/fchollet/keras`, 2015.

[53] Herbert H Clark and Jean E Fox Tree. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111, 2002.

[54] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. What makes a good conversation?: Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 475. Acm, 2019.

[55] Maximin Coavoux, Shashi Narayan, and Shay B Cohen. Privacy-preserving neural representations of text. *arXiv preprint arXiv:1808.09408*, 2018.

[56] Sheldon Cohen. Perceived stress in a probability sample of the united states. 1988.

[57] Sheldon Cohen, T Kamarck, R Mermelstein, et al. Perceived stress scale. *Measuring stress: A guide for health and social scientists*, pages 235–283, 1994.

[58] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. A global measure of perceived stress. *Journal of health and social behavior*, pages 385–396, 1983.

[59] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[60] Julien Dalpé, Martin Demers, Jérémie Verner-Filion, and Robert J Vallerand. From personality to passion: The role of the big five factors. *Personality and Individual Differences*, 138:280–285, 2019.

[61] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*, 2019.

[62] Daniel J DeDora, Joshua M Carlson, and Lilianne R Mujica-Parodi. Acute stress eliminates female advantage in detection of ambiguous negative affect. *Evolutionary Psychology*, 9(4):147470491100900406, 2011.

[63] 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.

[64] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–e1462, 2014.

[65] Emily D. Duvall, Alan Stuart Robbins, Thomas R Graham, and Scott Divett. Exploring filler words and their impact. 2014.

[66] Penelope Eckert. Age as a sociolinguistic variable. 2017.

[67] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*, 2018.

[68] Alex Engler. Auditing employment algorithms for discrimination. *Brookings Institute, Center for Technology Innovation*, 2021.

[69] Alexandre Evfimievski. Randomization in privacy preserving data mining. *ACM Sigkdd Explorations Newsletter*, 2002.

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

[71] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. Acm, 2010.

[72] José-Miguel Fernández-Dols and Carlos Crivelli. Emotion and expression: Naturalistic studies. *Emotion Review*, 5(1):24–29, 2013.

[73] Andrea Ferrario, Michele Loi, and Eleonora Viganò. In ai we trust incrementally: a multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology*, pages 1–17, 2019.

[74] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 2017.

[75] Thomas W Freeman, John Hart, Tim Kimbrell, and Elliott D Ross. Comprehension of affective prosody in veterans with chronic posttraumatic stress disorder. *The Journal of neuropsychiatry and clinical neurosciences*, 21(1):52–58, 2009.

[76] Susan R Fussell. The verbal communication of emotion: Introduction and overview. In *The verbal communication of emotions*, pages 9–24. Psychology Press, 2002.

[77] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

[78] Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. In *J. Mach. Learn. Res.*, 2016.

[79] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[80] Marc Garbey, Nanfei Sun, Arcangelo Merla, and Ioannis Pavlidis. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE transactions on Biomedical Engineering*, 54(8):1418–1426, 2007.

[81] Ismael Garrido-Muñoz, A. Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. A survey on bias in deep nlp. volume 11, page 3184, 2021.

[82] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, et al. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. Ieee, 2017.

[83] John Gideon, Melvin G McInnis, and Emily Mower Provost. Barking up the right tree: Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog). *arXiv preprint arXiv:1903.12094*, 2019.

[84] John Gideon, Heather T Schatten, Melvin G McInnis, and Emily Mower Provost. Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation. In *The 20th Annual Conference of the International Speech Communication Association INTERSPEECH 2019*, 2019.

[85] Tom Giraud, Mariette Soury, Jiewen Hua, Agnes Delaborde, Marie Tahon, David Antonio Gomez Jauregui, Victoria Eyharabide, Edith Filaire, Christine Le Scanff, Laurence Devillers, et al. Multimodal expressions of stress during a public speaking task: Collection, annotation and global analyses. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*. Ieee, 2013.

[86] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. Ieee, 2011.

[87] Lewis R Goldberg. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26, 1992.

[88] Jose Maria Gomez-Hidalgo, Jose Miguel Martin-Abreu, Javier Nieves, Igor Santos, Felix Brezo, and Pablo G Bringas. Data leak prevention through named entity recognition. In *IEEE Second International Conference on Social Computing*, 2010.

[89] Yuan Gong and Christian Poellabauer. Crafting adversarial examples for speech paralinguistics applications. *arXiv preprint arXiv:1711.03280*, 2017.

[90] Paul E Griffiths. Iii. basic emotions, complex emotions, machiavellian emotions 1. *Royal Institute of Philosophy Supplements*, 52:39–67, 2003.

[91] Antonio Gulli and Sujit Pal. *Deep Learning with Keras*. Packt Publishing Ltd, 2017.

[92] Sara Hajian and Josep Domingo-Ferrer. A study on the impact of data anonymization on anti-discrimination. In *2012 IEEE 12th ICDM*.

[93] Lasse Hansen, Yan-Ping Zhang, Detlef Wolf, Konstantinos Sechidis, Nicolai Lade-gaard, and Riccardo Fusaroli. A generalizable speech emotion recognition model reveals depression and remission. 2021.

[94] Katherine A Herborn, James L Graves, Paul Jerem, Neil P Evans, Ruedi Nager, Dominic J McCafferty, and Dorothy EF McKeegan. Skin temperature reveals the intensity of acute stress. *Physiology & behavior*, 152:225–230, 2015.

[95] Jet Hoek and Merel Scholman. Evaluating discourse annotation: Some recent insights and new approaches. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (isa-13)*, 2017.

[96] Frank Horvath. An experimental comparison of the psychological stress evaluator and the galvanic skin response in detection of deception. *Journal of Applied Psychology*, 63(3):338, 1978.

[97] Frank Horvath. Detecting deception: the promise and the reality of voice stress analysis. *Journal of Forensic Science*, 27(2):340–351, 1982.

[98] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, et al. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027*, 2021.

[99] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourc-ing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35. Association for Computational Linguistics, 2009.

[100] Guosheng Hu, Xiaojiang Peng, Yongxin Yang, Timothy M Hospedales, and Jakob Verbeek. Frankenstein: Learning deep face representations using small data. *IEEE Transactions on Image Processing*, 27(1):293–303, 2017.

[101] Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. Automatic dialogue generation with expressed emotions. In *Proceedings of NAACL*, 2018.

[102] Z. Huang and F. S. Cohen. 3-d motion estimation and object tracking using b-spline curve modeling. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 748–749, June 1993.

[103] Giovanni Iachello and Jason Hong. *End-user privacy in human-computer interaction*, volume 1. Now Publishers Inc, 2007.

[104] Vasileios Iosifidis and Eirini Ntoutsi. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, 24, 2018.

[105] Mimansa Jaiswal, Zakaria Aldeneh, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee, 2019.

[106] Mimansa Jaiswal, Zakaria Aldeneh, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. *CoRR*, abs/1903.11672, 2019.

[107] Mimansa Jaiswal, Zakaria Aldeneh, and Emily Mower Provost. Controlling for confounders in multimodal emotion classification via adversarial learning. *arXiv preprint arXiv:1908.08979*, 2019.

[108] Mimansa Jaiswal and Emily Mower Provost. Privacy enhanced multimodal neural representations for emotion recognition. In *Aaai*, pages 7985–7993, 2020.

[109] Mimansa Jaiswal and Emily Mower Provost. Best practices for noise-based augmentation to improve the performance of emotion recognition "in the wild". volume abs/2104.08806, 2021.

[110] Mimansa Jaiswal, Sairam Tabibu, and Rajiv Bajpai. The truth and nothing but the truth: Multimodal analysis for deception detection. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 938–943. Ieee, 2016.

[111] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[112] Kenneth Joseph, Lisa Friedland, William Hobbs, Oren Tsur, and David Lazer. Constance: Modeling annotation contexts to improve stance classification. *arXiv preprint arXiv:1708.06309*, 2017.

[113] Rebecca Jürgens, Annika Grass, Matthis Drolet, and Julia Fischer. Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. *Journal of nonverbal behavior*, 39(3):195–214, 2015.

[114] Jeffrey H Kahn, Renee M Tobin, Audra E Massey, and Jennifer A Anderson. Measuring emotional expression with the linguistic inquiry and word count. *The American journal of psychology*, pages 263–286, 2007.

[115] Divyansh Kaushik, Douwe Kiela, Zachary Chase Lipton, and Wen tau Yih. On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study. volume abs/2106.00872, 2021.

[116] Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, Melvin McInnis, and Emily Mower Provost. Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition. *Proc. Interspeech*, 2017.

[117] Soheil Khorram, Mimansa Jaiswal, John Gideon, Melvin McInnis, and Emily Mower Provost. The priori emotion dataset: Linking mood to emotion detected in-the-wild. In *Interspeech 2018*.

[118] Soheil Khorram, Mimansa Jaiswal, John Gideon, Melvin McInnis, and Emily Mower Provost. The priori emotion dataset: Linking mood to emotion detected in-the-wild. *arXiv preprint arXiv:1806.10658*, 2018.

[119] Gerald Kidd Jr, Christine R Mason, Jayaganesh Swaminathan, Elin Roverud, Kameron K Clayton, and Virginia Best. Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America*, 2016.

[120] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the ACM SIGMOD International Conference on Management of data*, 2011.

[121] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[122] Cara Kingston and James Schuurmans-Stekhoven. Life hassles and delusional ideation: Scoping the potential role of cognitive and affective mediators. *Psychology and Psychotherapy: Theory, Research and Practice*, 89(4):445–463, 2016.

[123] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993.

[124] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[125] Jáchym Kolář. *Automatic Segmentation of Speech into Sentence-like Units*. PhD thesis, University of West Bohemia in Pilsen, 2008.

[126] Kalpesh Krishna, Liang Lu, Kevin Gimpel, and Karen Livescu. A study of all-convolutional encoders for connectionist temporal classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5814–5818. Ieee, 2018.

[127] Hindra Kurniawan, Alexandr V Maslov, and Mykola Pechenizkiy. Stress detection from speech and galvanic skin response signals. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 209–214. Ieee, 2013.

[128] Virgile Landeiro and Aron Culotta. Robust text classification in the presence of confounding bias. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[129] Debi Laplante and Nalini Ambady. On how things are said: Voice tone, voice intensity, verbal content, and perceptions of politeness. *Journal of Language and Social Psychology*, 22(4):434–441, 2003.

[130] Amandine Lassalle, Delia Pigat, Helen O'Reilly, Steve Berggen, Shimrit Fridenson-Hayo, Shahar Tal, Sigrid Elfström, Anna Råde, Ofer Golan, Sven Bölte, Simon Baron-Cohen, and Daniel Lundqvist. The eu-emotion voice database. *Behavior Research Methods*, 51(2):493–506, Apr 2019.

[131] Richard S Lazarus and Judith Blackfield Cohen. Environmental stress. In *Human behavior and environment*, pages 89–127. Springer, 1977.

[132] Margaret Lech and Ling He. Stress and emotion recognition using acoustic speech analysis. In *Mental Health Informatics*, pages 163–184. Springer, 2014.

[133] Mihee Lee, Seokwon Bang, and Gyunghye Yang. Apparatus and method for inducing emotions, June 7 2011. US Patent 7,955,259.

[134] Sarah Ita Levitan, Taniya Mishra, and Srinivas Bangalore. Automatic identification of gender from speech. In *Proceeding of Speech Prosody*, pages 84–88, 2016.

[135] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014.

[136] Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu, and Weining Yang. Membership privacy: a unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. Acm.

[137] Tao Li, Shan Yang, Liumeng Xue, and Lei Xie. Controllable emotion transfer for end-to-end speech synthesis. pages 1–5, 2021.

[138] Wei Li, Farnaz Abtahi, Christina Tsangouri, and Zhigang Zhu. Towards an "in-the-wild" emotion dataset using a game-based framework. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1526–1534. Ieee, 2016.

[139] Li-Mei Liao and Mary G Carey. Laboratory-induced mental stress, cardiovascular response, and psychological characteristics. *Reviews in cardiovascular medicine*, 16(1):28–35, 2015.

[140] JEROEN Lichtenauer and MOHAMMAD Soleymani. Mahnob-hci-tagging database, 2011.

[141] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

[142] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

[143] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 2017.

[144] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.

[145] Weiyi Ma and William Forde Thompson. Human emotions track changes in the acoustic environment. *Proceedings of the National Academy of Sciences*, 112(47):14563–14568, 2015.

[146] Karttikeya Mangalam and Tanaya Guha. Learning spontaneity to improve emotion recognition in speech. *arXiv preprint arXiv:1712.04753*, 2017.

[147] David Matsumoto. Cultural influences on the perception of emotion. *Journal of Cross-Cultural Psychology*, 20(1):92–105, 1989.

[148] Katie Matton, Melvin G McInnis, and Emily Mower Provost. Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. *Proc. Interspeech 2019*, pages 1438–1442, 2019.

[149] Robert McHardy, Heike Adel, and Roman Klinger. Adversarial training for satire detection: Controlling for confounding variables. *CoRR*, abs/1902.11145, 2019.

[150] Robert McHardy, Heike Adel, and Roman Klinger. Adversarial training for satire detection: Controlling for confounding variables. *arXiv preprint arXiv:1902.11145*, 2019.

[151] Matthias R Mehl, Charles L Raison, Thaddeus WW Pace, Jesusa MG Arevalo, and Steve W Cole. Natural language indicators of differential gene regulation in the human immune system. *Proceedings of the National Academy of Sciences*, 114(47):12554–12559, 2017.

[152] Michael Mendelson and Yonatan Belinkov. Debiasing methods in natural language understanding make bias more accessible. *arXiv preprint arXiv:2109.04095*, 2021.

[153] Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gang, and Biing-Hwang Juang. Speaker-invariant training via adversarial learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5969–5973. Ieee, 2018.

[154] Katherine Metcalf, Barry-John Theobald, Garrett Weinberg, Robert Lee, Ing-Marie Jonsson, Russ Webb, and Nicholas Apostoloff. Mirroring to build trust in digital assistants. *arXiv preprint arXiv:1904.01664*, 2019.

[155] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[156] Caitlin Mills and Sidney D'Mello. On the validity of the autobiographical emotional memory task for emotion induction. *PloS one*, 9(4):e95837, 2014.

[157] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2227–2231. Ieee, 2017.

[158] Lyndon A Mitchell. The relationship between emotional recognition and personality traits. 2006.

[159] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, 2018.

[160] Saif M Mohammad. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278, 2022.

[161] Joan K Monin, Richard Schulz, Edward P Lemay Jr, and Thomas B Cook. Linguistic markers of emotion regulation and cardiovascular reactivity among older caregiving spouses. *Psychology and aging*, 27(4):903, 2012.

[162] Md Nasir, Brian Robert Baucom, Panayiotis Georgiou, and Shrikanth Narayanan. Predicting couple therapy outcomes based on speech acoustic features. *PloS one*, 12(9):e0185123, 2017.

[163] Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236, 2008.

[164] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675, 2003.

[165] Nhi Ngo and Derek M Isaacowitz. Use of context in emotion perception: The role of top-down control, cue type, and perceiver's age. *Emotion*, 15(3):292, 2015.

[166] Brian A Nosek, Anthony G Greenwald, and Mahzarin R Banaji. Understanding and using the implicit association test: Ii. method variables and construct validity. *Personality and Social Psychology Bulletin*, 31(2):166–180, 2005.

[167] A Oikonomopoulos, M Pantic, and I Patras. Human gesture recognition using sparse b-spline polynomial representations. In *Proceedings of Belgium-Netherlands Conf. Artificial Intelligence (BNAIC 2008), Boekelo, The Netherlands*, pages 193–200, 2008.

[168] Nicolas Papernot. A marauder's map of security and privacy in machine learning. *arXiv preprint arXiv:1811.01134*, 2018.

[169] Raghavendra Pappagari, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez, and Najim Dehak. Copypaste: An augmentation method for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[170] Emilia Parada-Cabaleiro, Alice Baird, Anton Batliner, Nicholas Cummins, et al. The perception of emotions in noisified nonsense speech. In Francisco Lacerda, editor, *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. Isca, 2017.

[171] Silke Paulmann, Desire Furnes, Anne Ming Bøkenes, and Philip J Cozzolino. How psychological stress affects emotional prosody. *Plos one*, 11(11):e0165022, 2016.

[172] Ioannis Pavlidis and James Levine. Thermal image analysis for polygraph testing. *IEEE Engineering in Medicine and Biology Magazine*, 21(6):56–64, 2002.

[173] Ioannis Pavlidis, James Levine, and Paulette Baukol. Thermal imaging for anxiety detection. In *Proceedings IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (Cat. No. PR00640)*, pages 104–109. Ieee, 2000.

[174] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[175] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek, editors, *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*. Isca, 2021.

[176] Karol J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*.

[177] Kurt Wesley Piersol and Gabriel Beddingfield. Pre-wakeword speech processing, January 29 2019. US Patent App. 14/672,277.

[178] 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.

[179] Danish Pruthi, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary Chase Lipton, Graham Neubig, and William W. Cohen. Evaluating explanations: How much do explanations from the teacher aid students? volume 10, pages 359–375, 2022.

[180] Emilie Qiao-Tasserit, Maria Garcia Quesada, Lia Antico, Daphne Bavelier, Patrik Vuilleumier, and Swann Pichon. Transient emotional events and individual affective traits affect emotion recognition in a perceptual decision-making task. *PloS one*, 12(2):e0171375, 2017.

[181] Jan Querengässer and Sebastian Schindler. Sad but true?-how induced emotional states differentially bias self-rated big five personality traits. *BMC Psychology*, 2(1):14, 2014.

[182] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

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

[184] Andrew Rosenberg. Classifying skewed data: Importance weighting to optimize average recall. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[185] Leon JM Rothkrantz, Pascal Wiggers, Jan-Willem A Van Wees, and Robert J van Vark. Voice stress analysis. In *International conference on text, speech and dialogue*, pages 449–456. Springer, 2004.

[186] Marius Rubo and Matthias Gamer. Social content and emotional valence modulate gaze fixations in dynamic scenes. *Scientific reports*, 8(1):3804, 2018.

[187] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[188] James A Russell. Emotion recognition: Is it universal? 2017.

[189] Andrea C Samson, Sylvia D Kreibig, Blake Soderstrom, A Ayanna Wade, and James J Gross. Eliciting positive, negative and mixed emotional states: A film library for affective scientists. *Cognition and emotion*, 30(5):827–856, 2016.

[190] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak. Emotion identification from raw speech signals using dnns. *Proc. Interspeech 2018*, pages 3097–3101, 2018.

[191] Odette Scharenborg, Sofoklis Kakouros, and Jiska Koemans. The effect of noise on emotion perception in an unknown language. In *Speech Prosody 2018*. Isca, 2018.

[192] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. volume 9, pages 1408–1424, 2021.

[193] Wolff Schlotz, Ilona S Yim, Peggy M Zoccola, Lars Jansen, and Peter Schulz. The perceived stress reactivity scale: Measurement invariance, stability, and validity in three countries. *Psychological assessment*, 23(1):80, 2011.

[194] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in NIPS*, 2018.

[195] Philipp Schmidt and Felix Biessmann. Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558*, 2019.

[196] Yusuke Shinohara. Adversarial multi-task learning of deep neural networks for robust speech recognition. In *Interspeech*, pages 2369–2372. San Francisco, CA, USA, 2016.

[197] Anna L. Shparberg. Google books ngram viewer. 2021.

[198] Ewa Siedlecka and Thomas F Denson. Experimental methods for inducing basic emotions: A qualitative review. *Emotion Review*, 11(1):87–97, 2019.

[199] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Cvpr*, 2017.

[200] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[201] Mohammad Soleymani and Martha Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. *SIGIR-Workshops*, 2010.

[202] Jose Angel Soto and Robert W Levenson. Emotion recognition across cultures: The influence of ethnicity on empathic accuracy and physiological linkage. *Emotion*, 9(6):874, 2009.

[203] Fabian Sperrle, Udo Schlegel, Mennatallah El-Assady, and Daniel Keim. Human trust modeling for bias mitigation in artificial intelligence. In *ACM CHI 2019 Workshop: Where is the Human? Bridging the Gap Between AI and HCI*, 2019.

[204] Arthur W Staats and Georg H Eifert. The paradigmatic behaviorism theory of emotions: Basis for unification. *Clinical Psychology Review*, 10(5):539–566, 1990.

[205] Herman JM Steeneken and John HL Hansen. Speech under stress conditions: overview of the effect on speech production and on system performance. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 4, pages 2079–2082. Ieee, 1999.

[206] Victoria Stenbäck. *Speech masking speech in everyday communication: The role of inhibitory control and working memory capacity*, volume 1559. Linköping University Electronic Press, 2016.

[207] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*, Prague, Czech Republic, 2007.

[208] Harald Strömfelt, Yue Zhang, and Björn W Schuller. Emotion-augmented machine learning: Overview of an emerging domain. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 305–312. Ieee, 2017.

[209] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Learning predictive models that transport. *arXiv preprint arXiv:1812.04597*, 2018.

[210] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. Activity-aware mental stress detection using physiological sensors. In Martin Gris and Guang Yang, editors, *Mobile Computing, Applications, and Services*, pages 282–301, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[211] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

[212] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. volume 29, pages 24–54, 2010.

[213] E Heather Thompson, Phyllis Robertson, Russ Curtis, and Melodie H Frick. Students with anxiety: Implications for professional school counselors. *Professional School Counseling*, 16(4):2156759x150160402, 2013.

[214] T. Tieleman and G. Hinton. Lecture 6.5—rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[215] JHM Tulen, P Moleman, HG Van Steenis, and F Boomsma. Characterization of stress reactions to the stroop color word test. *Pharmacology Biochemistry and Behavior*, 32(1):9–15, 1989.

[216] Matthew T Tull, Heidi M Barrett, Elaine S McMillan, and Lizabeth Roemer. A preliminary investigation of the relationship between emotion regulation difficulties and posttraumatic stress symptoms. *Behavior Therapy*, 38(3):303–313, 2007.

[217] Francisco Valente, Simão Paredes, and Jorge Henriques. Improving the compromise between accuracy, interpretability and personalization of rule-based machine learning in medical problems. pages 2132–2135, 2021.

[218] Jean-Marc Valin. A hybrid dsp/deep learning approach to real-time full-band speech enhancement. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. Ieee, 2018.

[219] Leaf Van Boven, Katherine White, and Michaela Huber. Immediacy bias in emotion perception: Current emotions seem more intense than previous emotions. *Journal of Experimental Psychology: General*, 138(3):368, 2009.

[220] Ashwini Ann Varghese, Jacob P Cherian, and Jubilant J Kizhakkethottam. Overview on emotion recognition system. In *2015 international conference on soft-computing and networks security (ICSNS)*, pages 1–5. Ieee, 2015.

[221] Daniel Västfjäll. The subjective sense of presence, emotion recognition, and experienced emotions in auditory virtual environments. *CyberPsychology & Behavior*, 6(2):181–188, 2003.

[222] Nicholas Lo Vecchio. Updating the oed on the historical lgbtq lexicon. 2021.

[223] W Richard Walker and John J Skowronski. The fading affect bias: But what the hell is it for? *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 23(8):1122–1136, 2009.

[224] Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. Allennlp interpret: A framework for explaining predictions of nlp models. *arXiv preprint arXiv:1909.09251*, 2019.

[225] Manjie Wang and Kimberly J Saudino. Emotion regulation and stress. *Journal of Adult Development*, 18(2):95–103, 2011.

[226] Wei Wang, Ivan Hernandez, Daniel A Newman, Jibo He, and Jiang Bian. Twitter analysis: Studying us weekly trends in work stress and emotion. *Applied Psychology*, 65(2):355–378, 2016.

[227] Martin Wegrzyn, Maria Vogt, Berna Kireclioglu, Julia Schneider, and Johanna Kissler. Mapping the emotional face. how individual face parts contribute to successful emotion recognition. *PloS one*, 12(5):e0177239, 2017.

[228] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Cvpr*, 2016.

[229] Kelly H Werner, Philippe R Goldin, Tali M Ball, Richard G Heimberg, and James J Gross. Assessing emotion regulation in social anxiety disorder: The emotion regulation interview. *Journal of Psychopathology and Behavioral Assessment*, 33(3):346–354, 2011.

[230] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910, 2019.

[231] Mingke Xu, Fan Zhang, and Wei Zhang. Head fusion: Improving the accuracy and robustness of speech emotion recognition on the iemocap and ravdess dataset. *IEEE Access*, 9:74539–74549, 2021.

[232] Yi-Hsuan Yang and Homer H Chen. Prediction of the distribution of perceived music emotions using discrete samples. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2184–2196, 2011.

[233] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. The ordinal nature of emotions. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*, pages 248–255. Ieee, 2017.

[234] Habib Yaribeygi, Yunes Panahi, Hedayat Sahraei, Thomas P Johnston, and Amirhossein Sahebkar. The impact of stress on body function: A review. *EXCLI journal*, 16:1057, 2017.

[235] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[236] Paul Yovanoff, Luke Duesbery, Julie Alonzo, and Gerald A. Tindal. Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. volume 24, pages 4–12, 2005.

[237] Jiahong Yuan and Mark Liberman. Automatic detection of "g-dropping" in american english using forced alignment. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. Ieee.

[238] Xiaobu Yuan. An approach to integrating emotion in dialogue management. In *International Conference in Swarm Intelligence*, pages 297–308. Springer, 2015.

[239] Biqiao Zhang, Georg Essl, and Emily Mower Provost. Predicting the distribution of emotion perception: capturing inter-rater variability. In *International Conference on Multimodal Interaction*, pages 51–59, 2017.

[240] Xuan Zhang, Hui W Yu, and Lisa F Barrett. How does this make you feel? a comparison of four affect induction procedures. *Frontiers in psychology*, 5:689, 2014.

[241] Jingwen Zhao, Yunfang Chen, and Wei Zhang. Differential privacy preservation in deep learning: Challenges, opportunities and solutions. *IEEE Access*, 7:48901–48911, 2019.

[242] Sicheng Zhao, Guiguang Ding, Jungong Han, and Yue Gao. Personality-aware personalized emotion recognition from physiological signals. In *Ijcai*, pages 1660–1667, 2018.

[243] Xin Zuo and Pascale Fung. A cross gender and cross lingual study on acoustic features for stress recognition in speech. In *ICPhS*, pages 2336–2339, 2011.