Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Xiaoyuan Huang April 12, 2022

Unsupervised, Context-Aware Emotion Classification of College-Related Reddit Posts

By

Xiaoyuan Huang

Jinho D. Choi
Advisor

Department of Mathematics

Jinho D. Choi
Advisor

Elizabeth Newman
Committee Member

Carl Yang
Committee Member

2022

Unsupervised, Context-Aware Emotion Classification of College-Related Reddit Posts

By

Xiaoyuan Huang

Jinho D. Choi
Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciencewith Honors

Mathematics

2022

Abstract

Unsupervised, Context-Aware Emotion Classification of College-Related Reddit Posts
By Xiaoyuan Huang

As emotion plays an important role in conversations, empathetic dialogue systems have been developed to be used in fields such as business and healthcare. However, a lack of such chatbots exists in the higher education sector. To develop such dialogue systems, emotion detection serves as the most important step. Sentiment analysis and emotion detection on social media has been a meaningful way to diagnose emotions, understand behaviors, and help improve empathetic agents. Current work has focused on machine learning and rule-based approaches, but the number of emotion labels of many existing models is limited. Therefore, inspired by the gap between higher education and emotion-related tasks in the Natural Language Processing field, the goal of this thesis is to develop a novel and well-performed emotion classifier specifically targeting college-related social media contents and producing more elaborated emotion labels than existing emotion classifiers. This thesis achieved this goal by three main steps. The first step was to generate a task-specific dataset for model development. The second step was to develop baseline models using Transformer trained on Empathetic Dialogues for basic emotion detection. The third part was to improve these baseline models by developing unsupervised models that overcome difficulties of detecting neutrality in the baseline models and target higher education contents with better model performances. This work would provide a meaningful tool for more fine-grained emotion detection in college-related textual data and future chatbot developments in higher education as an innovative solution for institutions.

Unsupervised, Context-Aware Emotion Classification of College-Related Reddit Posts

By

Xiaoyuan Huang

Jinho D. Choi
Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciencewith Honors

Mathematics

2022

Acknowledgments

My thanks to Dr. Jinho Choi, my advisor, who has guided my research since I joined the Emory NLP lab at the beginning of the Spring term in my junior year. As both my academic and life advisor, Dr. Choi greatly inspired my interests in computational linguistics and helped shape my future graduate study in Data Science. Also, I would like to thank my committee members, Dr. Elizabeth Newman and Dr. Carl Yang. They not only provided me with meaningful suggestions for my honors thesis but also supported my undergraduate study at Emory University. Last but not least, I would like to express my thanks to my parents and friends who have shown support during the pandemic.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

The growth of studies in human-computer interactions through Artificial Intelligence (AI) prompts applications and developments of chatbots in different areas. A large amount of research focuses on task-oriented chatbots for customer service and question answering [23],[46]. Other studies have made progress towards open-domain bots such as using empathetic dialogue systems in order to develop more human-like chatbots, as emotion plays a crucial role in conversations and social media. Researchers have been tackling problems such as providing mental support for people through these agents [40].

As many current open-domain chatbots leverage deep learning methods, another important aspect that has been investigated by researchers increasing the interpretability of these models. Work such as Graph Reasoning for Inference Driven Dialogue (GRIDD) framework [10] has been proposed to increase the controllability and interpretability by enhancing the inferential ability of social chatbots, which emphasized discussions of personal thoughts and experiences in conversations. The framework enables better understanding of input semantics, more flexible initiative taking, and more novel responses that are coherent with the contextual information in dialogues. However, without an inference engine for analyzing users' emotions,

the framework is not capable of responding with empathy at this point. Responses generated in the framework are largely based on inferences and leave out reactions to speakers' emotions. In addition, the content about universities involved in this framework is not oriented towards a purpose of serving higher education institutions.

As part of the important steps towards developing such empathetic chatbots, emotion detection is often involved in the process. Current emotion classifiers have been using rule-based, machine learning based, or hybrid approaches [9] to tackle this problem. However, due to the limited number of available datasets and emotion labels, these models often produce basic emotion predictions. It is also hard to conclude the cause of predicted emotions from these models.

Social media analysis has been incorporating sentiment analysis and emotion detection as the most significant processes in understanding user behaviors and reactions online. For example, mental health problems are detected through analyzing social media data with sentiment analysis [27]. Studies of political science has also been using social media platforms to predict relevant events such as election results [3]. Emotions from social media data are helpful in summarizing social events and attitudes. With rich information on social media, these analyses produce insights that are hard to get through other type of textual data.

In the field of higher education, there have been attempts to develop chatbots, but many focused on information-based virtual agents to retrieve useful resources upon queries [14], [28]. These AI agents help facilitate online navigation systems of college websites. While these chatbots provide useful guidance for online users, they often lack human-like features such as showing acknowledgements and making appropriate inferences to speakers' inputs. Few studies have made contributions to design an open-domain and empathetic chatbot in college settings.

Given that there are few applications of conversational agents in higher education, and emotion provides meaningful insights in the process of developing these agents,

I was inspired to develop a well-performed emotion classifier targeting college-related social media contents in this thesis. Specifically, the thesis aims to approach the emotion detection problem with unsupervised models that produce more refined and elaborate emotions than existing models. This approach would help overcome common limitations in many of the current datasets and approaches that use these dataset, such as the lack of emotion labels and the lack of interpretability in commonly-used machine learning approaches. This would also provide a meaningful tool for future chatbot developments in higher education as an innovative solution for institutions. Such a university chatbot could provide students with services including academic advising, mental consulting, and career development sessions. Developing such a chatbot would also save much time and increase efficiencies for students who have personal questions they are not willing to discuss with people or prefer anonymous ways of talking due to private concerns.

In this paper, Chapter 2 summarizes the background of this thesis by summarizing related work on research that inspired this project and a comprehensive literature review on emotion detection, emotion-cause extraction, and empathetic dialogue systems. It also introduces relevant Natural Language Processing models used in this thesis. Chapter 3 introduces the data collection process and all the datasets used for model development in Chapter 4. In Chapter 4, I describe both baseline models and the unsupervised approach of detecting emotions in our Reddit dataset. Chapter 5 concludes the paper by summarizing findings and discusses potential future directions of this work for further improvements.

1.1 Thesis Statement

By developing a context-aware unsupervised approach of one-label and two-label emotion detection models, we expect these models to achieve three goals on the college-

related Reddit data: 1) the two-label emotion detection model will perform better than the one-label emotion detection model and baseline models; 2) the unsupervised approach will be able to refine neutrality as an additional emotion label that does not exist when using baseline models; 3) the unsupervised models will be applicable to future analysis of college-related contents and even other fields of studies.

Chapter 2

Background

This chapter introduces all the related work to my thesis. In the beginning stage of planning the thesis, I thought about two potential directions: 1) generating empathetic responses for college-related contents so that it would be helpful in developing university chatbots; 2) extracting emotion-cause pairs from textual data and generating reasoning structures because it would help understand the reasoning behind emotion predictions. After a thorough literature review and experiments, I ended up focusing my thesis on developing context-aware unsupervised approaches of detecting emotions.

2.1 Alexa Prize Socialbot Grand Challenge IV

The motivation of this thesis project comes from my participation in the Alexa Prize team to compete for the prestigious Amazon Social Chatbot Grand Challenge IV. Our novel approach of Graph Reasoning for Inference Driven Dialogue (GRIDD) framework [10] allowed the chatbot Emora to make appropriate inferences based on user's inputs and generate controllable human-like responses. To enable Emora's language capabilities, I built the knowledge base and ontology as concept graphs for common-sense reasoning, constructed implication rules for inferences, and coded

natural language templates for response generation. I also fine-tuned a deep learning model (RoBERTa) that allowed the matching of user inputs with the most relevant wise saying, providing an innovative solution for Emora to produce more engaging responses. These contributions helped our team to be selected as the finalist team in the competition.

Deeply captivated by Emora's mission of ultimately providing mental support for those in need through Natural Language Processing, I focused my honors thesis on the extraction of emotional information from discussion forums that I will elaborate in subsequent chapters. This project would make meaningful contributions for Emora's ongoing development of emotional reasoning, making it a more empathetic chatbot that helps people with mental health issues.

2.2 Collaboration Work

The work of this thesis is in the process of collaboration with another project of two members from the Emory NLP Lab, Mack and Daniil, who have been focusing on the automatic generation of multi-turn dialogues from Reddit data that will be introduced in Chapter 3. Inspired by the lack of high quality multi-turn dialogue data, they took the advantage of BlenderBot, BERT Next Sentence Prediction, and Reddit's structure to identify a strong-performing model for multi-turn dialogue assembly. Using comment threads, they enhanced their conversations by utilizing the conversational nature of threads on Reddit. Finally, they used a variety of metrics to help filter our resulting conversations for better results, especially focusing on conversation coherency.

This thesis intends to help improve the performance of their models by adding the emotion component, such as providing appropriate emotion prediction to generate responses or the next sentence. For example, with the emotion classifier I developed, it

could serve as one of the metrics used in their project to determine the best response. This would make the response generation more emotionally reasonable and cohesive, making the dialogue look more realistic.

2.3 Emotion Analysis

Given that data from social media and discussion forums contains rich information and my motivation with adding emotional capabilities to Emora, the original idea that I wanted to pursue was empathetic response generation with inference for university chatbots. To achieve this, I intended to integrate an emotion classifier into the GRIDD system as an inference engine by adding predicted emotions of inputs as predicates so that response templates could be added to generate empathetic responses. This would allow both emotion-aware and inferential responses taking the advantage of the controllability in the GRIDD system.

To achieve this, I completed a thorough literature review on relevant work, including emotion theories, emotion detection, emotion-cause extraction, and empathetic response generation. The literature review on these areas led me to focus on the final thesis direction of emotion classification and emotional information extraction in Reddit posts.

2.3.1 Theories of Emotion

Theories of emotion have been a fundamental element of understanding emotions. Many current work on emotion classification, introduced in Section 2.1.2, relies on major emotion theories in the Psychology field as basic ways of defining emotion categories. This section introduces two major types of models of emotion: discrete and dimensional models.

Discrete emotion models divide emotions into distinct categories. The Ekman's

basic emotion model [7] and Parrott's six basic emotions with extension [16] (see Figure 2.1) fall in this type of emotion models. Ekman's model considers emotions of Western cultures, including ANGER, DISGUST, FEAR, HAPPINESS, SADNESS, and SURPRISE. In comparison, Parrott's model starts with basic emotions of FEAR, SADNESS, SURPRISE, ANGER, LOVE, as well as JOY, and extends these emotions into a tree that encompasses 100 separate emotions. The OCC model (proposed by Ortony, Clore, and Collins) [29], another discrete emotion model, views emotion as a result of individual perceptions of events and considers emotional intensity to define 22 emotion categories. These discrete emotion models are significant in providing conceptual clarity to the study of emotions. They represent emotions with easy-to-understand labels. However, the limitation of these models is that emotional categories may not represent different emotional states even if the set of emotion categories is defined.

Dimensional emotion models suggest that each emotion is characterized by multiple dimensions. Circumplex's [18] and Whissell's [43] emotion models considers two dimensions. Circumplex model defines emotions according to their intensity in the vertical axis of arousal and the horizontal access of valence, while Whissell's model considers levels of positivity and activeness. Another multi-dimensional emotion model widely applied is Plutchik's Emotions Wheel [34], offering a more comprehensive and hybrid structure of organizing emotions into concentric circles with inner being more basic and intense and the outer more complicated and less intense emotions (see Figure 2.2). For example, fear is a more complicated than but less intense emotion than terror, so fear is placed in an outer position than terror in the Emotions Wheel. They provide measures of comparison between emotion categories, as adjacent classes in the space are very similar to each other.

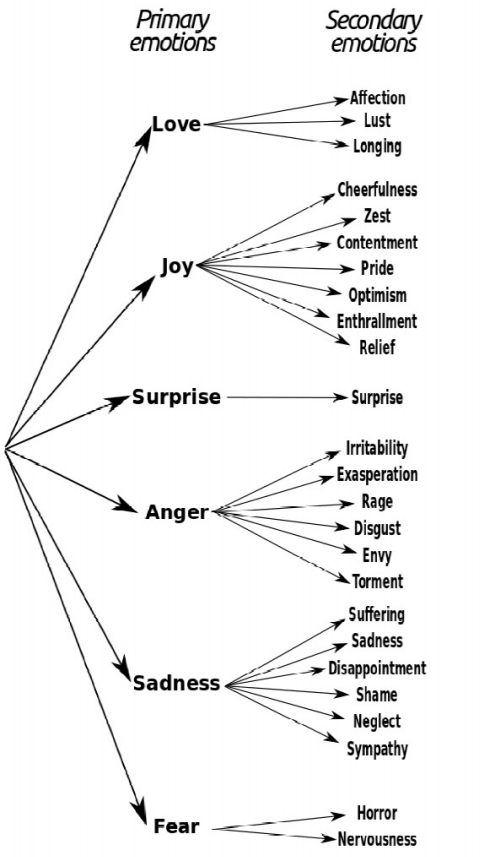

Figure 2.1: Parrott's six basic emtions with extension [16]

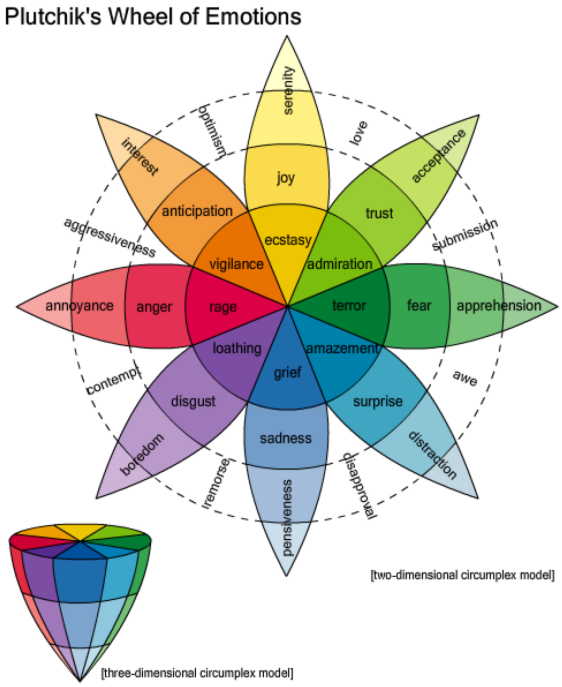

Figure 2.2: Plutchik's Emotions Wheel [34]

2.3.2 Textual Emotion Detection

Emotion detection is a substantial process in further applications such as building empathetic dialogue systems. Because my thesis is interested in leveraging textual data, I conducted a literature review mainly on textual emotion detection. Approaches of recognizing emotions mainly include rule-based, machine learning (ML), and hybrid methodologies.

By constructing grammatical and logical rules to follow in order to recognize emotions in texts, rule-based approaches rely on emotion dictionaries or lexicons. The notable WordNet-Affect dictionary [39] is often used to perform keyword recognition, as the dictionary provides search words to assign emotion labels. Other notable dictionaries include EmoSenticNet [33], SentiWord Net [8], and National Research Council of Canada (NRC) lexicon [26]. Although this is a simple and straightforward way of classifying emotions, it faces challenges such as the complexity of generating reasonable dictionaries and the lack of semantic meanings. Researchers have also

developed lexical affinity methods to improve the keyword recognition method, which assigns emotion words with an additional probabilistic affinity. For example, the word "awesome" may be assigned with a probabilistic affinity of "positive". However, this method does not consider the context of the text and may lead to inaccuracies in emotion classifications. For example, the sentence "The rain ruined my awesome day" may be classified as "positive" because of the word "awesome", while it actually expresses a negative emotion.

The ML methodology approaches the emotion detection problem by supervised or unsupervised ways of classifying texts into different emotion categories. Supervised approaches mainly rely on datasets. Existing popular emotion datasets include Daily Dialogue [20], CARER [36], Empathetic Dialogues [35], MELD [31], GoodNewsEveryone [32], and GoEmotions [5] (see Table 2.1 for a comparison between different datasets). This thesis chose to use Empathetic Dialogues after comparison with other available datasets. These datasets Models such as ESTeR combines word co-occurrences and word associations from lexicons for unsupervised emotion detection, proposing a novel similarity function based on random walks on graphs [13].

Dataset	Year	Size	Number of Emotions
Daily Dialogue	2017	13,118	6 (Basic Emotions)
CARER	2018	20,000	6 (Basic Emotions)
Empathetic Dialogues	2018	22,908	32
MELD	2019	1400	7 (Basic Emotions + Neutral)
GoodNewsEveryone	2020	5,000	6 (Basic Emotions)
GoEmotions	2021	5,800	28

Table 2.1: 32 emotion labels of the *Empathetic Dialogues* dataset

Hybrid approaches combine both the rule-based and the ML methodologies into a unified model [9] to improve the performance of the emotion classification tasks.

However, existing classifiers often contain limited emotion labels.

2.3.3 Emotion-Cause Pair Extraction

As the ability of making inferences is essential to make chatbots more reasonable, there have been approaches to increase the interpretability of emotion detection and the capability of commensense inferences in recent years. This gives information about the reasoning behind classified emotions and makes machine learning models more understandable and applicable for downstream tasks. I discovered these approaches when conducting the literature review for emotion detection, giving me some new ideas of possible thesis directions at the beginning of my thesis planning stage.

Current work has used attention-based models to extract emotion-cause span with an emotion polarity classifier [21]. In the attention model they developed, they included both emotion-aware attention and context-aware attention to reinforce not only emotions but also contextual information that was important to the emotion. This work inspired my final decision on developing a context-aware approach for emotion detection.

Other approaches used concepts from psychology to aid the model developments. Inspired by the Cognitive Theory of Emotion, DialogueCRN [17] was designed to address multi-turn reasoning modules to extract and integrate emotional clues. This approach mimics the normal cognitive thinking process. Another work introduced CIDER [12] to perform dialogue-level natural language inference, span extraction of emotional commensense, and multi-choice span selection for implicit and explicit inferences, which extracted rich explanations from conversations that were conducive to improving downstream tasks.

2.3.4 Empathetic Dialogue Systems

Some current approaches to empathetic dialogue systems significantly rely on learning from large scale conversation data to generate responses. For example, Lin et al. (2020) [22] introduce a system that adapts Generative Pretrained Transformer (GPT) fined-tuned by PersonaChat dataset [45] and the EmpatheticDialogue dataset [35] for generating empathetic responses via transfer learning. Their model involved multi-task objectives, including response modeling, prediction, and detection of dialogue emotions. Xie et al. (2019) propose a Multi-turn Emotionally Engaging Dialog model (MEED) [44], which utilizes a hierarchical attention mechanism to track historical conversations and a dedicated embedding layer for emotion encoding in the field of open-domain dialogue systems. Although these approaches perform well in different aspects, responses generated by this type of data-driven method are often short and vague; they may also be inconsiderate and redundant since the outputs are more or less unpredictable.

Tracking emotion states of the speaker and the listener becomes one of the focuses in developing empathetic chatbots. In one paper, the attempt was to develop a sentiment look-ahead reward function to model the future user emotional state using reinforcement learning [37]. The model provided a higher reward when the generated utterance improved the user's sentiment and thus helped generate more empathetic responses. In addition, Emotional Chatting Machine (ECM) [15] generated appropriate responses not only in content (relevant and grammatical) but also in emotion (emotionally consistent) by embedding emotion categories, capturing changes in emotion states, and using external emotion vocabulary. This approach was based on the encoder-decoder framework of the general sequence-to-sequence model.

Recent advances have also attempted Generative Adversarial Net (GAN) to generate emotional responses because it has shown promising results in text generation. However, the texts generated by GAN usually suffer from poor-quality contents and

lack of diversity. Novel frameworks such as SentiGan [42] was thus proposed to address these problems by incorporating multiple generators and one multi-class discriminator to generate diversified examples of each sentiment.

2.4 Sentiment Analysis in Social Media

After confirming that the direction of this thesis would proceed with emotional information extraction from social media and discussion forums, I reviews work related to sentiment analysis in these platforms as well.

Social media analysis is important for many different fields. For the business sector, vendors use social media platforms such as Twitter to advertise product features and collect feedback from their clients [1]. These feedbacks from customers are valuable for companies to analyze customer behaviors and help them improve product or services. Sentiment analysis in this type of analysis assists marketers in understanding perspectives from customers so that they know customer attitudes towards their current products and how to change and improve the services they provide [19], [2]. The rise of social media and associated sentiment analysis thus reformat the way business runs.

For healthcare sectors, online social media platforms such as Twitter contain rich and essential information provided by professionals and citizens. In facing the outbreak of Covid-19, for instance, people have been using Twitter as a major resource hub and posting platform to share thoughts and opinions on the pandemic [11]. With the increased mental health issues stimulated by such situation, health professionals have conducted sentiment and emotion analysis by utilizing data from these platforms to detect psychological disorders such as depression [38].

The education sector also uses sentiment analysis to evaluate academic performances of students, teachers, and institutions as many schools are using platforms

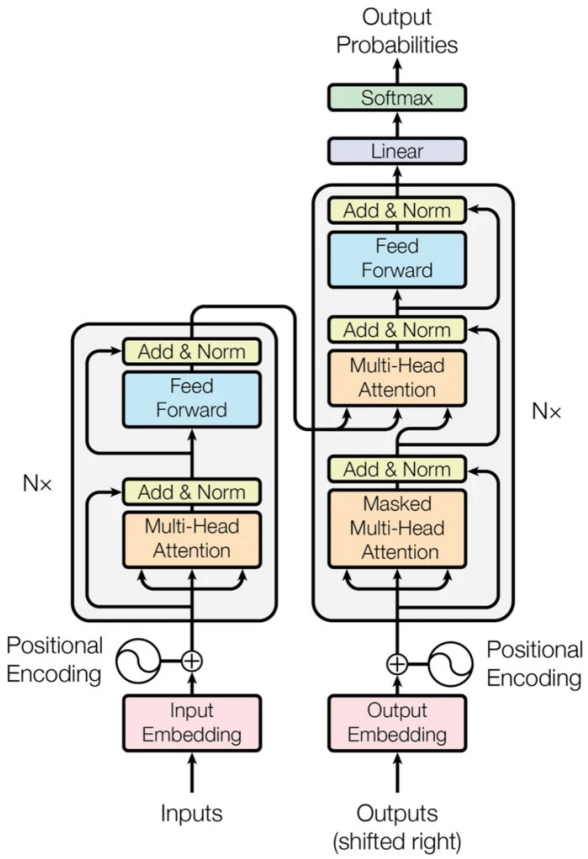

Figure 2.3: Transformer model architecture [16]

such as Facebook to collect feedbacks. Sentiment analysis performed with such data has thus become important in improving educational practices [25].

2.5 Transformer Models

Transformer models [41] have been providing substantive solution to sequential text problems in Natural Language Processing research, which help produce many state-of-the-art results in applications, including machine translation, language modeling, text classification, and document summarization. As my approaches were based on these models, this section introduces the general structure and variants of Transformer models. Figure 2.3 shows the model architecture of a Transformer model.

The model contains two major blocks: the encoder and the decoder. A softmax

activation function is added to normalize output probabilities. A sequence of data is the input to the model, and then the input words are passed through positional encoders, which assign vectors to words based on their positions in the sentence and extract contextual meaning of the input. The multi-head attention and a feed-forword network in the encoder blocks capture the relationship between words in the sentence by computing attention vectors, which are then passed to the decoder block. The decoder blocks also have another encoder-decoder attention layer which allows the decoder to focus on and pay attention to appropriate parts of the input.

Overall, by utilizing the attention mechanism, the Transformer models are able to store the hidden information of inputs by deciding the important part of the sequence.

BERT (Bidirectional Encoder Representations from Transformers) [6] is one of the variants of the Tranformer model introduced by Google. There are two models of BERT, including the BERT-base and BERT-large models. The BERT-base model is made up of 12-layered transformer encoder blocks with each block containing 12-head self-attention layers and 768 hidden layers, producing about 110 milliion parameters. The BERT-large model is made up of 24-layered transformer encoder blocks with each block containing 16-head self-attention layers and 1024 hidden layers, producing about 340 million parameters.

The Robustly Optimized BERT pre-training Approach (RoBERTa)[24] is a BERT variant that seeks to ultimately optimize BERT by tweaking various methodological parameters in the initial version of BERT released by Facebook. It is shown to have a better performance than BERT. RoBERTa-base is a model that uses BERT-base architecture, and RoBERTa-large is a model that uses BERT-large architecture.

Chapter 3

Dataset

3.1 Empathetic Dialogues (ED-32)

The *Empathetic Dialogues* (ED-32) dataset [35] is a novel large-scale dataset of 22,908 open-domain conversations grounded in emotional situations, released by Facebook. The dataset is divided into three sets: 17,623 for training, 2,747 for validation, and 2,538 for testing. Each dialogue consists of 1 of the 32 emotion labels (see Table 3.1), a description of the conversational situation, and a multi-turn dialogues between a speaker and a listener based on the situation. This dataset is used in models described in later sections for training and comparing purposes. Figure 3.1 shows an example of a dialogue in the ED dataset.

Label: Afraid
Situation: Speaker felt this when...
"I've been hearing noises around the house at night"
Conversation:
Speaker: I've been hearing some strange noises around the house at night.
Listener: oh no! That's scary! What do you think it is?
Speaker: I don't know, that's what's making me anxious.
Listener: I'm sorry to hear that. I wish I could help you figure it out

Figure 3.1: Example dialogue in the *Empathetic Dialogues* dataset

0	afraid	8	confident	16	furious	24	nostalgic
1	angry	9	content	17	grateful	25	prepared
2	annoyed	10	devastated	18	guilty	26	proud
3	anticipating	11	disappointed	19	hopeful	27	sad
4	anxious	12	disgusted	20	impressed	28	sentimental
5	apprehensive	13	embarrassed	21	jealous	29	surprised
6	ashamed	14	excited	22	joyful	30	terrified
7	caring	15	faithful	23	lonely	31	trusting

Table 3.1: 32 emotion labels of the *Empathetic Dialogues* dataset

This dataset is ideal for the purpose of this thesis because: 1) The dataset contains a total of 32 emotion labels, which provides a larger variety in emotion categories compared with many other classical datasets that only provide POSITIVE, NEGATIVE, and NEUTRAL labels, or basic six emotions of SADNESS, HAPPINESS, FEAR, ANGER, SURPRISE and DISGUST; 2) The distribution of the 32 emotions is relatively balanced in this dataset, which is significant for classification models because it helps generate higher accuracy models and higher balanced accuracy; 3) The situations in the dataset are relevant to daily life and more similar than other existing datasets to the college-related Reddit posts, which is important because this thesis was interested in the emotion analysis of data from this type of text.

However, the dataset has two limitations, which were tackled in this thesis: 1) This dataset does not contain the NEUTRAL emotion label, so this thesis focuses on adding neutrality into consideration when classifying emotions, as described in later sections; 2) Another limitation tackled in this thesis is that situations that convey similar emotions are labeled with different categories in the dataset, which is demonstrated by an example in Table 3.2 In the example, both situations could be alternatively labeled as either JOYFUL or SURPRISED.

3.2 College-Related Reddit Posts

Reddit College is a dataset created by Dr. Jinho Choi at the Emory NLP Lab [4],

Situation	Label
I went home, and my wife surpised me with a picture of our future baby.	joyful
I just found out that my sister and her husband are pregnant.	surprised

Table 3.2: Example of situations with similar emotions from the *Empathetic Dialogues* dataset

containing rich information on Reddit posts and corresponding comments in subreddits related to college. Up to December 14, 2021, there was a total of 36,044 posts extracted from Reddit (see Table 3.3).

Subreddit	Post Count
ApplyingToCollege	15,815
AskAcademia	2,616
College	8,753
CollegeAdvice	818
CollegeMajors	1,274
CollegeRant	2,480
Emory	1,591
GradSchool	3,515
Total	36,044

Table 3.3: Statistics of the *Reddit College* dataset

This dataset was ideal for this thesis because Reddit was one of the central hubs for students and educational professionals to discuss college-related concerns. I also selected to use Reddit data after comparisons with Twitter and Quora. While Twitter is a good platform for people to express their opinions, it contains messier language and many other elements such as images and videos under hashtags. Quora is a huge website for answering questions from people, but it produces more formal and professional responses than Reddit. Therefore, I chose Reddit as the discussions around college raised from people who are actually in institutions would be especially

helpful for developing university chatbots in the future.

In this thesis, only the textual data from the post part was considered because the interest of this thesis lies in concerns of members from college communities rather than the responses they get from other people. Specifically, values that were relevant and used included sid (the subreddit ID) and text (the content of the post part). Sentence tokenization was also performed to the dataset to split a post into its sentences. All special characters and URLs in the post text parts were also removed, so only the plain text data was left for further classification and analysis.

3.3 Data Annotation

A selective set of the *Reddit College* dataset was annotated to provide evaluation and development benchmarks for models in Chapter 4. This enhanced the specificity of the approaches on college contents. To further enhance the comparability among different Reddit posts, lengths of posts from the dataset were controlled to posts that consisted of 10 sentences. Expenses of the data annotation process were also considered to determine the number of posts. As a result, a total of 100 posts, 1000 sentences (utterances) were selected for annotation.

3.3.1 Merged Empathetic Dialogues (ED-8) and Neutrality

The original idea was to use the 32 emotion labels from ED-32 to classify emotions, but the accuracy of such models was low when tested on our Reddit dataset (see Chapter 4). It also did not provide NEUTRAL emotion label. Therefore, labels from the *Empathetic Dialogues* dataset were merged into 8 categories based on the theory of Emotions Wheel [30] to address the limitation of producing different emotion labels for similar sentences as specified in Section 3.1 (see Table 3.4). The merged *Empathetic Dialogues* dataset is denoted as ED-8 in the following sections.

Also, the category of NEUTRAL emotion was added to the annotation process, in addition to the exiting 8 emotion labels, to overcome the limitation of missing neutrality in the *Empathetic Dialogues* dataset. In the end, there was a total of 9 emotion labels used in data annotation.

Merged Label	Original Label(s) in the *Empathetic Dialogues* dataset
Joy	joyful, excited, content, proud, grateful
Trust	caring, trusting
Fear	afraid, terrified, embarrassed, ashamed, guilty, apprehensive, anxious
Surprise	surprised, impressed
Sadness	sad, devastated, sentimental, nostalgic, lonely, disappointed
Disgust	disgusted
Anger	furious, angry, annoyed, jealous
Anticipation	anticipating, hopeful, confident, prepared, faithful
Neutral	N/A

Table 3.4: 8 merged and 1 neutral emotion labels for annotation

3.3.2 MTurk Tasks

To allow models to achieve their best performance on the *Reddit College* dataset, data was crowd-sourced through Amazon Mechanical Turk (MTurk). MTurk is an online marketplace developed by Amazon that outsources jobs published by individuals and businesses to a distributed workforce who performs the tasks virtually. It enables the acceleration of data collection and analysis, streamline business processes, and machine learning development. MTurk is thus considered as an ideal platform for my thesis to annotate data because it broke down this time-consuming project into smaller, more manageable tasks to be completed by distributed workers.

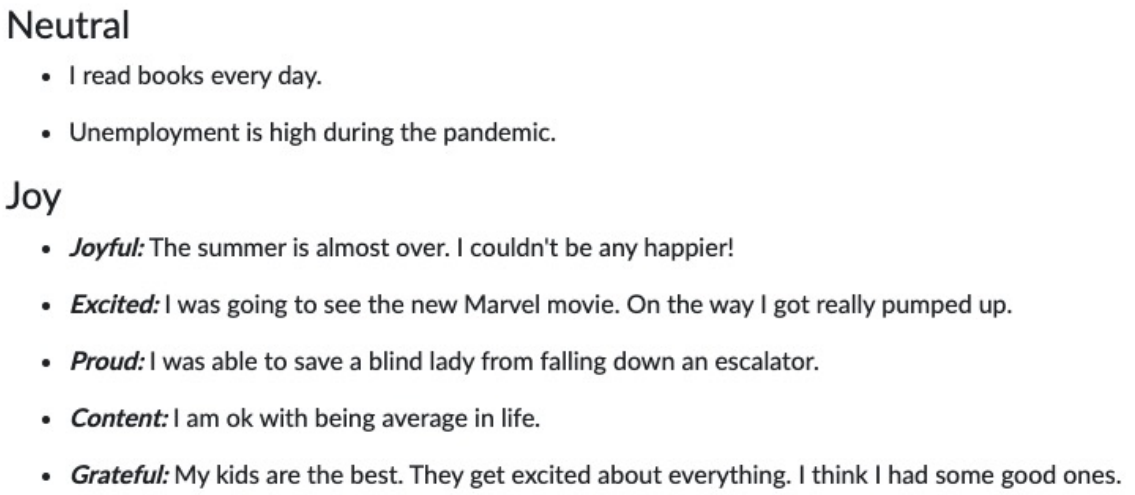

Figure 3.2: Good examples provided to MTurk workers in Task 1

Task 1: Emotion Classification

In this task, each MTurk worker was presented with a body paragraph from a Reddit post and its sentences to be classified. The worker was expected to select the primary sentiment expressed in each sentence out of the 9 sentiments specified in 3.3.1, given the whole paragraph as the context. If there were multiple emotions expressed, the worker was expected to use their best judgement and choose the strongest sentiment. Good examples of emotion classifications were provided to guide the workers (see Figure 3.2). Each set of paragraph and sentences was distributed to 2 workers for annotation. The interface on MTurk was designed to highlight instructions, the paragraph, sentences with contrasted colors, and drop-down selections for classification (see Figure 3.3).

Task 2: Emotion Judication

This task is a subsequent task following Task 1. The worker who received this task was shown a post, its sentences, and 2 annotated labels from Task 1 for each sentence. The worker in this task was expected to justify which sentiment annotation(s) match the emotion of the sentence. Workers may choose 1 of the 2 annotations, Both,

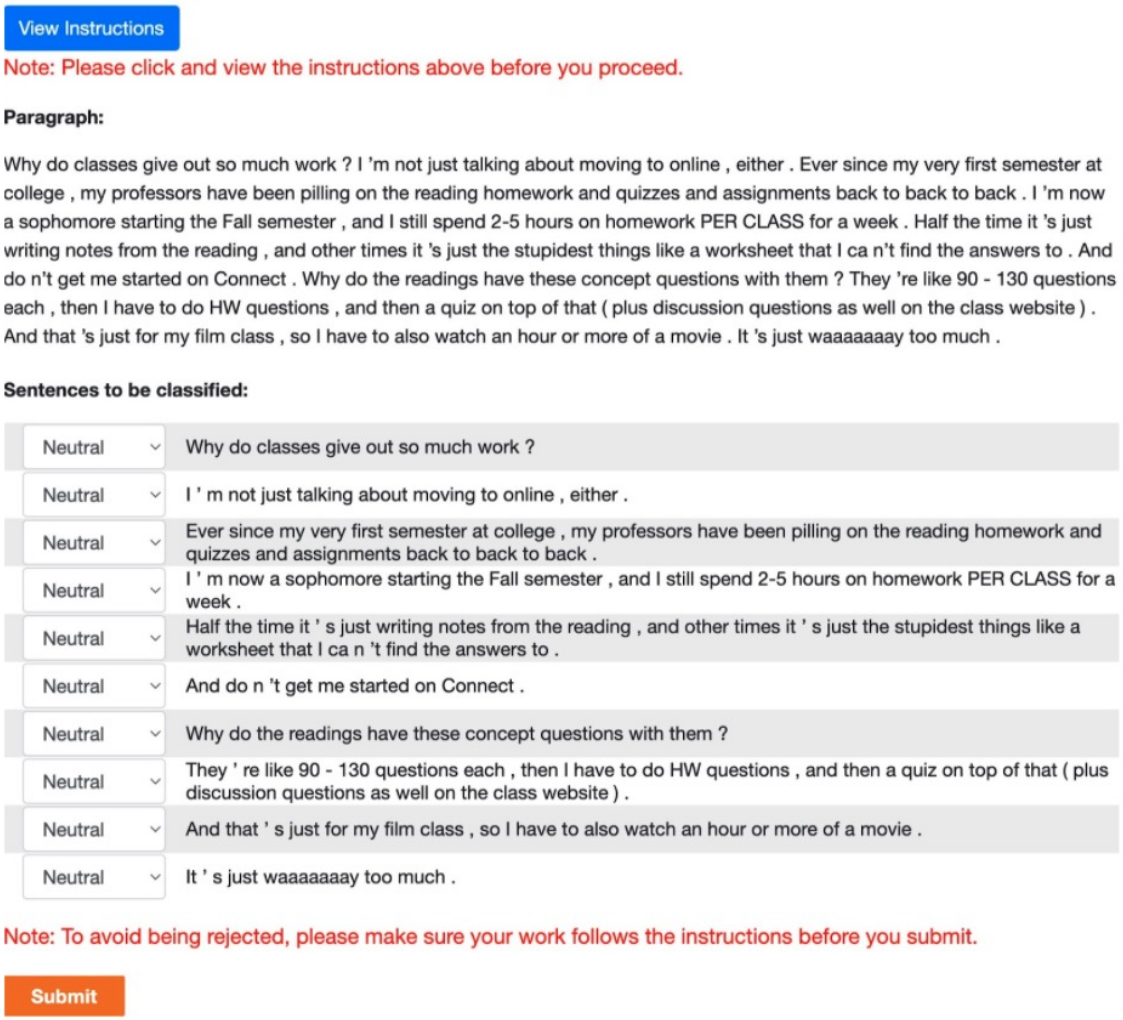

Figure 3.3: Example MTurk task interface of Task 1

or Neither, as the judication answer. Each set of paragraph and sentences with annotations was distributed to 2 workers for judication. The MTurk interface was designed to highlight instructions, sentences and their annotations with contrasted colors, and drop-down selections for judications (see Figure 3.4). In the end, one batch (20 posts, 200 sentences) of this task was published.

Experiments and Evaluation

For Task 1, annotation agreement was measured by calculating the number of utterances labeled with the same emotions from the 2 workers, divided by the total number of utterances. A total of 3 batches were distributed to online workers with different number of posts and different attempts to adjust the task interface each. The first batch was published with detailed instructions and examples for guiding the annotation, which resulted in a 24.5% agreement score for the 20 posts. In the second batch, reminders of reading instructions were added in red and bold font before workers started to select emotion labels and before they submitted their answers. However, the agreement was low for another 20 posts and dropped by 5% as compared with the first batch. In the final batch, automatic tests were created to reject invalid workers. That is, a list of sentences with expected labels was used as the true labels. If the workers selected wrong labels for those sentences, their submission would be rejected, and the website would automatically re-publish the task for other workers to complete. This attempt did not improve the agreement score, as it yielded a 18% agreement, lower than previous two batches. For agreed annotations, the accuracy of them was 71.4% overall and 90.1% for the SADNESS emotion label.

One possible reason of the low agreement score from Task 1 was that 2 annotations may both be correct. Therefore, the remaining disagreed annotations were assessed manually to see the distribution of situations where both annotations were correct. Results showed that annotations were both correct for 19% of the time, and neither

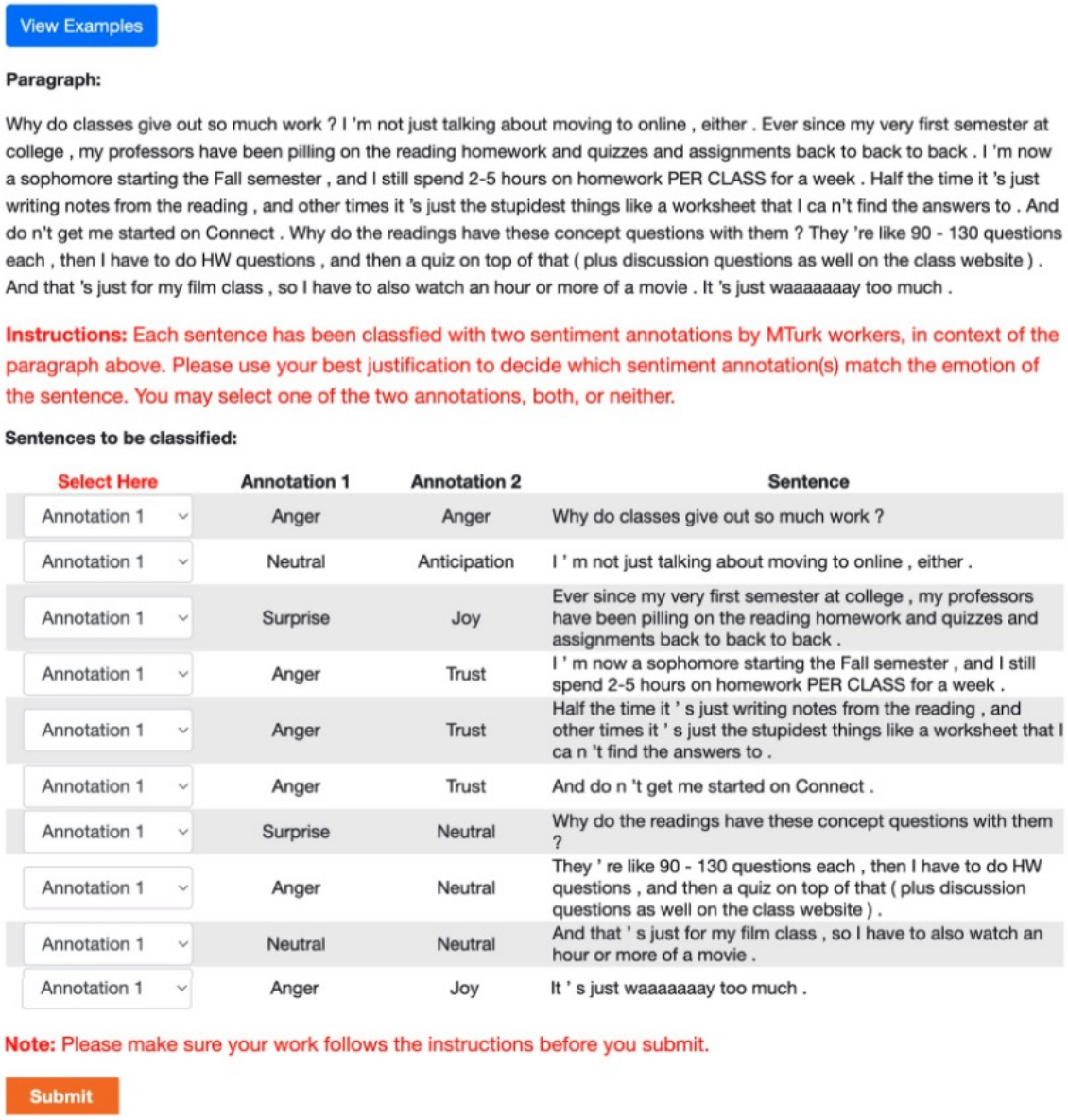

Figure 3.4: Example MTurk task interface of Task 2

Batch	Number of Posts	Attempt	Annotation Agreement
1	20	Original interface	24.5%
2	20	Reminders in bold and red font	19.5%
3	60	Tests to reject invalid workers	18%

Table 3.5: Batches of tasks published for Task 1 (Overall Agreement: 19.6%)

of the 2 annotations was correct for 17.2% of the time. More specifically, there was a total of 25.1% of the cases where one of the annotations was NEUTRAL. After checking the true sentiment distribution, I found that NEUTRAL was true for 45.5% of the time. However, because of the small size of such samples, we may not conclude that the annotation of NEUTRAL was trustable. These statistics indicated weak reliability of the annotations from MTurk.

In addition, Task 2 was published to justify the results of Task 1 via MTurk. In Task 2, judication agreement was measured by calculating the number of utterances labeled with the same judication selections from the 2 workers, divided by the total number of utterances. A total of 1 batch of judication tasks was published to workers, which corresponded to data from Batch 1 (20 posts, 200 sentences) in Task 1. The overall judication agreement was 37.5% for this batch, and the overall accuracy of these judications was 56%. Out of these judications, the accuracy of BOTH as the judication was the highest, which was 94.7% (see Table 3.6).

To assess results from annotations and judications deeply, I analyzed situations for agreed NEITHER judication. An example of agreed NEITHER judication and their corresponding annotations was provided in Table 3.7, with bolded NEITHER indicating correct judications and others indicating ajudications. We could see that only 2 out of 6 (33.3%) NEITHER judications were true, as Annotation 2 (ANTICIPATION) for Sentence 0 was actually true and both annotations (DISGUST) for sentence 2 were true.

Agreed Judication	Accuracy
Both	94.7%
Neither	33.3%
1 of the 2 Annotations	44.0%

Table 3.6: Accuracy of agreed judications (Overall Accuracy: 56%)

For ajudications, annotations by workers from Task 1 were also revisited to de-

Index	Sentence	A1	A2	J1	J2
0	Unfortunately , many of us are increasingly desperate for opportunities and lots of scumbags (especially MLMs) are going to try to take advantage of us .	Neutral	Anticipation	**Neither**	**Neither**
1	I ' ve had enough of it , it ' s unbelievably annoying .	Anger	Anger	Neither	Neither
2	Cool , I did n 't know this internship was LinkedIn , but please carry on .	Disgust	Disgust	**Neither**	**Neither**
3	I woke up at 7 AM today and threw up right as I got out of bed .	Joy	Joy	Neither	Neither
4	I do n 't really know if this post fits here but I really wanted to get this out .	Anger	Neutral	Neither	Neither
5	I managed to somehow score a C on the first exam , did n 't take the second because I was too anxious , and now my third exam is coming up and I do n 't even know where to begin ???	Fear	Fear	Neither	Neither

Table 3.7: Example of agreed NEITHER judication and their annotations (A: Annotation; J: Judication. Bolded text indicates true judication.)

Index	Sentence	A1	A2	J1	J2
0	I have been avoiding studying for organic chemistry (a big no no) since the beginning of the semester pretty much .	Disgust	Fear	Neither	**A2**
1	This kid in my internship constantly has to brag about about where he attends school and how he ' s the only legal intern .	Disgust	Disgust	Neither	**Both**
2	I did my fair share with research , and during the day it was due , we discussed .	Neutral	Joy	**A1**	Neither
3	But I feel like I am wasting my time on something I will never use again and sorta is the focus of the degree .	Neutral	Joy	Both	**Neither**

Table 3.8: Example of one-NEITHER judications and their annotations (A: Annotation; J: Judication. Bolded text indicates true judication.)

termine if the judications were trustworthy and if they could provide schemes for potential automatic annotation on the rest of the data. The analysis started with assessing one-NEITHER judications. Overall, the accuracy of NEITHER was 10.7%, and the accuracy of the non-NEITHER judication was 84%. Table 3.8 shows an example of comparisons between one-NEITHER judications and one-NEITHER ajudications. We could see from the example that when NEITHER was an ajudication, the non-NEITHER judication was trustable. However, because of the relatively small sample (25 out of 200) of such cases, this rule could not be ultimately used to automatically generate correct annotations for the remaining data.

Then, for one-BOTH situations, the BOTH judication was correct for 29.4% of the time, and 66.7% of the non-BOTH judications were true. Specifically, I analyzed one-BOTH judications in disagreed annotations to see if we could find patterns when

Index	Sentence	A1	A2	J1	J2
0	A legitimate company will provide you with any materials necessary to do the job .	Trust	Surprise	Both	**A1**
1	I ' m glad I stuck with it and having a valuable degree is worth the extra work .	Joy	Surprise	**Both**	A2
2	One guy said " I ' m almost done with it already , I ' ll send it to you guys to review and turn it in at 7:00 pm " he never sent me the video to review , so i texted the other guy who is also waiting for it to review , because he promised to make a groupchat but did n 't so he is the only one I can text directly about the group , and he never responded .	Anger	Joy	A2	Both

Table 3.9: Example of one-BOTH judications for disagreed annotations (A: Annotation; J: Judication. Bolded text indicates true judication.)

one-BOTh judications were presented. Table 3.9 shows an example of such situation for disagreed annotations, including when BOTH was true, non-BOTH judication was true, and two judications were incorrect. After examining these results, although over half of the non-BOTH judications were true, the sample of such data was still too small (24 out of 200) to allow automatic annotations.

The final assessment considered judications for disagreed annotations. I found that when judications agreed for disagreed annotations, 79.6% of these judications were true. Table 3.10 shows an example of true and false cases for agreed judications. Overall, the agreed judications achieved much higher accuracy for disagreed annotations than disagreed judications.

Given results from Task 1 and Task 2, the agreement scores were low for both annotation and judication processes. The judication process also stopped after finishing the first batch due to its low reliability. The accuracies of these tasks were also not stable. Therefore, data annotated via MTurk was not reliable enough for pro-

Index	Sentence	A1	A2	J1	J2
0	I ' m now a sophomore starting the Fall semester , and I still spend 2-5 hours on homework PER CLASS for a week .	Anger	Trust	**A1**	**A1**
1	Why do the readings have these concept questions with them ?	Surprise	Neutral	A2	A2

Table 3.10: Example of agreed judications for disagreed annotations (A: Annotation; J: Judication. Bolded text indicates true judication.)

ceeding with model development described in Chapter 4. Automatic data annotation according to the MTurk results was also hard to achieve.

3.3.3 Self-Annotated Reddit Posts

Annotations from Section 3.3.2 did not improve much after the quality control process by attempts of adjusting the task interface, and judications were not helpful enough for developing automatic annotations. Therefore, results from the previous section were not used in this thesis. A self-annotated Reddit post dataset was created during the assessment, so models in Chapter 4 used this dataset. For the selected 100 posts (1,000 sentences), each sentence was labeled with up to 2 emotions. A sentence was classified with 2 emotions only if they were equally reasonable for representing emotions conveyed by the sentence.

Table 3.11 shows an example of comparisons between annotations created by MTurk workers and my annotations. My annotation refined emotions and removed incorrect annotations from MTurk (e.g., sentence 0) as well as revised some emotion annotations that did not exist in MTurk annotations (e.g., sentence 1 and sentence 2). I also completed a 1-round validation of annotations to confirm that each sentence was labelled with the most accurate emotion(s).

This dataset was further divided into two sets, the development set and the test set, with similar sentiment distributions for model development (see Figure). Each set

Index	Sentence	Annotation(s) from MTurk	My Annotation(s)
0	Last semester I was super close with my roommates and we just had a great time .	Joy, Neutral	Joy
1	Any advice on how to deal with this / how to tell them would be greatly appreciated :)	Trust, Anticipation	Joy, Anticipation
2	However , stuff happened and now the three of them have just stopped including me / totally ignoring me .	Sadness, Disgust	Anger, Surprise

Table 3.11: Example of comparisions between annotation(s) from MTurk workers and my annotation(s)

contains 50 posts, 500 sentences. Detailed statistics of this dataset will be introduced in Chapter 4.

Chapter 4

Emotion Classification

This chapter introduces Transformer-based models as baselines as well as unsupervised approaches of developing emotion classifiers.

4.1 Emotion Distributions in Datasets

The experimental data that I used for developing models in this section was the *Empathetic Dialogues* (ED-32 and ED-8) and the self-annotated *Reddit* data. The distribution of emotions was almost uniformed in ED-32 (see Figure 4.1).

After merging the 32 emotions into 8 emotions, the distribution became less uniformed for ED-8 (see Figure 4.2). However, the distribution in ED-8 looked similar to that in the self-annotated *Reddit* data (see Figure 4.3). The emotion distributions in the development set and the test set were also similar for the *Reddit* data so that parameters that later described to be tuned in Section 4.3 on the development set would be applicable to the test set.

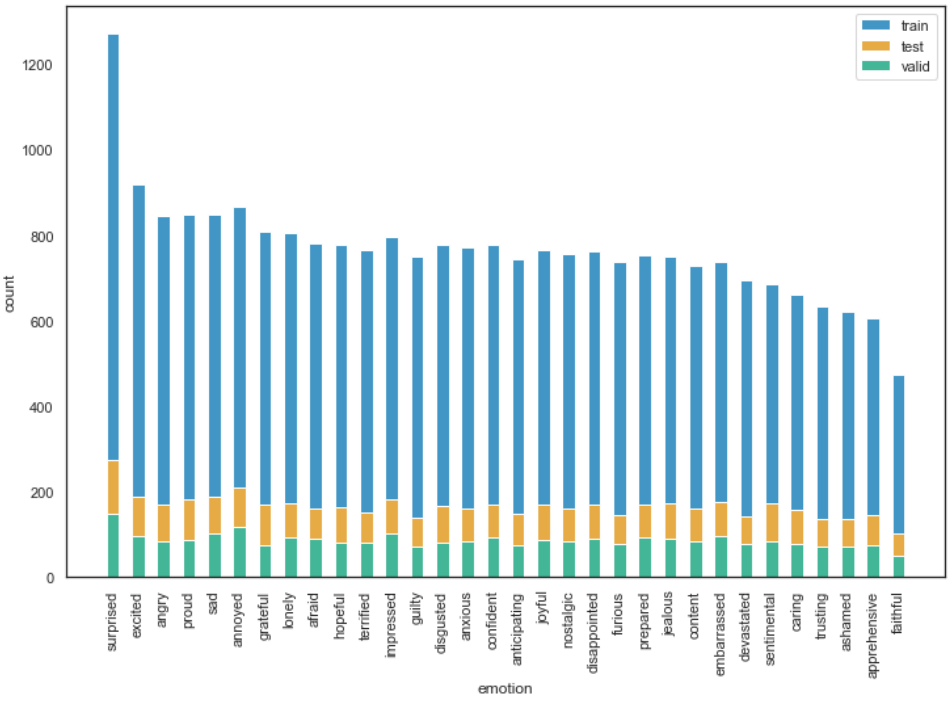

Figure 4.1: Distribution of emotions in the ED-32 dataset

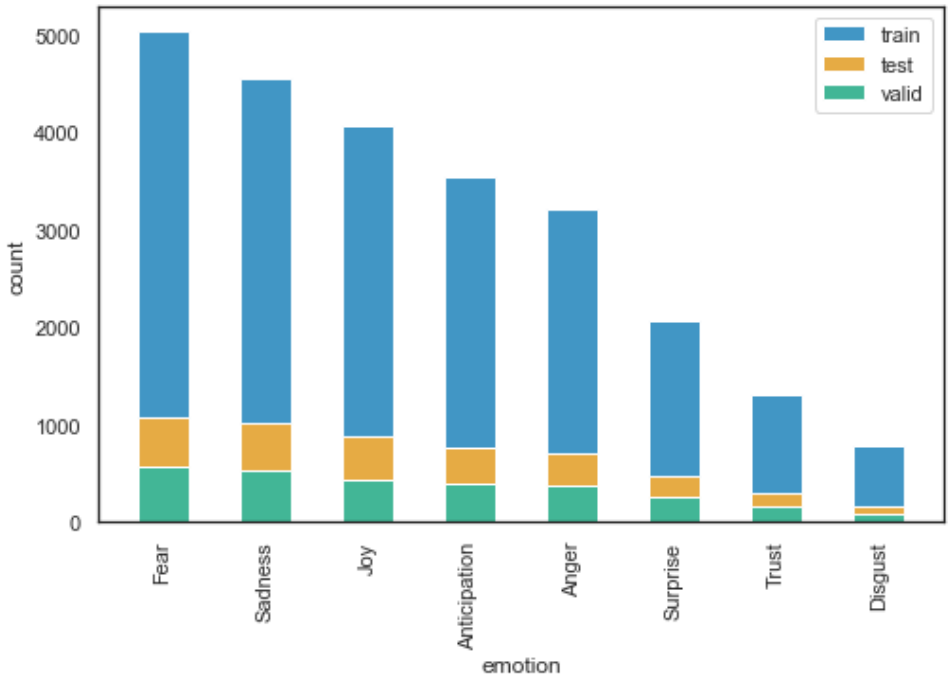

Figure 4.2: Distribution of emotions in the ED-8 dataset

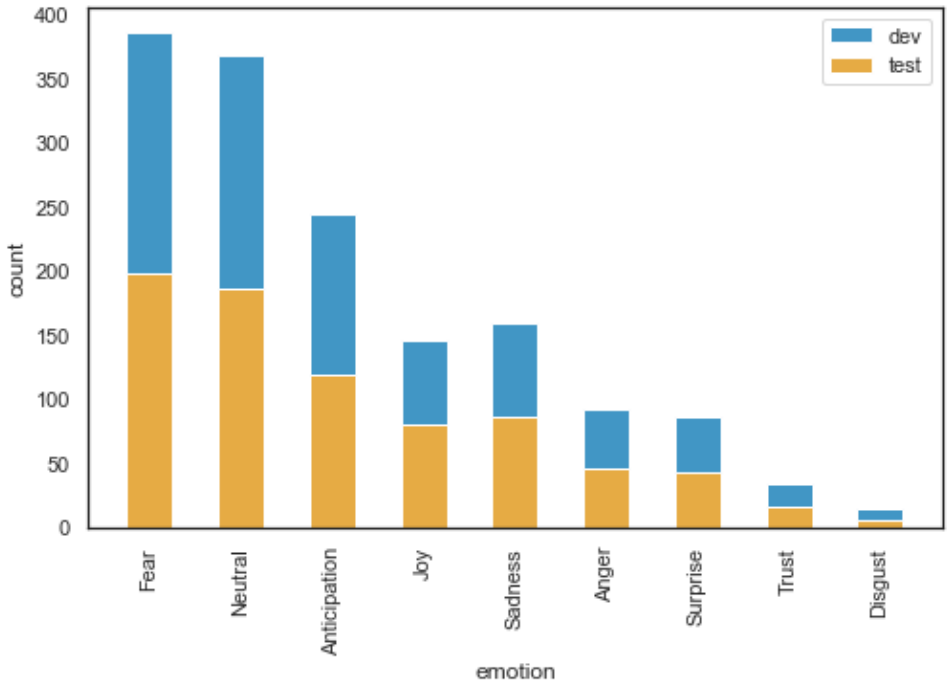

Figure 4.3: Distribution of emotions in the self-annotated Reddit dataset

4.2 Transformer Baseline Models

To determine the best-performed emotion classifier for college-related Reddit posts, baseline models were developed first to be compared with unsupervised models introduced in later sections. Transformer models (introduced in Section 2.5) that were used as baselines included BERT and RoBERTa (base and large). To control the experiments so that I could make comparisons between models, the configurations of all models were set to be the same: for example, they used a batch size of 32 with 3 epochs of training, dropout probability of 0.1, and 12 number of hidden layers.

The dataset used to train the model was the training set from the *Empathetic Dialogues* (ED-32 and ED-8) with situations and their corresponding emotion labels. The goal of this training process is to allow the emotion classifier to detect emotions given a situational sentence that is similar to the language from the *Reddit College* data. These models were also compared with unsupervised approaches in later sections.

4.2.1 32To8 Single-Label Approach

In this approach, Transformer models were first trained with the *Empathetic Dialogues* situations and their corresponding 32 emotion labels (ED-32). After testing the model with ED-32, the model was tested with ED-8 and the self-annotated *Reddit College* dataset that contained 8 emotion labels as introduced in Section 3.3.1. This helped assess the performance of the models in their original on the task-specific Reddit post data.

4.2.2 Merged-8 Single-Label Approach

In this approach, Transformer models were first trained with the *Empathetic Dialogues* situations and their merged 8 emotion labels (ED-8) following merging rules introduced in Section 3.3.1. The models were then tested with ED-8 and the self-annotated *Reddit College* dataset that contained 8 emotion labels. The approach directly classified input utterances into 1 of the 8 emotions.

4.2.3 Experiments and Evaluation

Both the 32To8 single-label approach and the Merged-8 single-label approach were experimented with BERT, RoBERTa-base, and RoBERTa-large models and tested with ED-32, ED-8, and the the self-annotated *Reddit College* datasets.

The evaluation metrics used here varied according to the test datasets. When the test data was from ED-32 and ED-8, model accuracies were calculated as the number of true predictions divided by the total number of predictions since the dataset only contained one true label for each utterance, which means that the number of predictions equal to the number of true labels. When the test data was from the self-annotated *Reddit College* dataset, three metrics were calculated: 1) Precision, the number of true predictions divided by the total number of predictions; 2) Recall, the

number of true predictions divided by the total number of true labels; 3) F1 Score, the harmonic mean of precision and recall (2 times the product of precision and recall divided by the sum of precision and recall).

For each of the approaches (32To8 and Merged-8 single-label approaches described above), model accuracies with BERT, RoBERTa-base, and RoBERTa-large for detecting emotions with the corresponding test set and number of emotions were measured and compared. Among 32To8 single-label classifiers, accuracies increased for all Transformer models after the 32 emotions were merged into 8 labels (see Table 4.1). This means that the process of merging was effective in detecting emotion more accurately for the *Empathetic Dialogues* dataset. The accuraccies in 32To8 models for ED-8 were also slightly higher than the ones produced by Merged-8 models, meaning that classifying utterances with 32 emotions and then merging them into 8 emotions was more effective than directly classifying utterances with 8 emotions. Overall, the model with the highest accuracy was the 32To8 single-label approach with RoBERTa-base, which had an accuracy of 0.819.

Model	Approach	Dataset	Accuracy
BERT	32To8	ED-32	0.575
		ED-8	0.770
	Merged-8	ED-8	0.762
RoBERTa-base	32To8	ED-32	0.604
		ED-8	0.808
	Merged-8	ED-8	0.801
RoBERTa-large	32To8	ED-32	0.627
		ED-8	**0.819**
	Merged-8	ED-8	0.805

Table 4.1: Accuracy of single-label baseline models on *Empathetic Dialogues*

These models were then tested on the self-annotated *Reddit College* test set de-

scribed in Section 3.3.3. For BERT-based models, the performance of the 32To8 approach was worse than that of the Merged-8 approach. However, for RoBERTa-base and RoBERTa-large models, the 32To8 approach achieved better results than the Merged-8 approach. Overall, the model that achieved the highest precision, recall, and F1 score was the 32To8 single-label approach with RoBERTa-base. This model reached an F1 score of 0.540, 0.011 higher than the second largest F1 score produced by 32To8 model with RoBERTa-large.

Model	Approach	Precision	Recall	F1 Score
BERT	32To8	0.524	0.345	0.416
	Merged-8	0.586	0.386	0.465
RoBERTa-base	**32To8**	**0.680**	**0.448**	**0.540**
	Merged-8	0.664	0.437	0.527
RoBERTa-large	32To8	0.666	0.439	0.529
	Merged-8	0.646	0.426	0.513

Table 4.2: Evaluation of single-label baseline models on the self-annotated *Reddit College* test set

4.2.4 Results and Analysis

Although 32To8 RoBERTa-large model outperformed all other models on the ED-8 dataset, it did not achieve the highest precision, recall, or F1 scores on the self-annotated *Reddit College* test set. The 32To8 RoBERTa-base model was the best-performed model on the dataset.

By taking a specific post as an example, where bolded font indicated correct classification outputs and normal font indicated wrong classification outputs, we could see that the RoBERTa-base models had the highest number of correct classifications (see Table 4.4). The basic label assigned by the 32To8 BERT model deviated the most regarding the emotion expressed by the utterance (see Table 4.3), as it classified

the second sentence as JOY while other models classified it as emotion labels with opposite meanings. The Merged-8 RoBERTa-large model also produced the most wrong predictions in this specific example (see Table 4.5).

4.3 Unsupervised Models

Because baseline models did not provide the emotion label of NEUTRAL, I attempted the unsupervised approach to assign NEUTRAL labels considering results produced by the baseline models as well as the context of the post. This approach was novel in the sense that most existing models on emotion detection relied on supervised learning and only predicted a limited number of emotion labels. Algorithm 1 demonstrates the logical processes of merging labels based on consecutive sentences to produce at most two emotion predictions.

Hyper-parameter tuning was performed on the development set from the self-annotated *Reddit College* to first find the best parameters for the models. These parameters were then used to predict results on the test set. The algorithm demonstrated the following steps to make emotional predictions after finding the optimal number of these paraemeters, including standard deviations N_σ, consecutive distance threshold A_t, and percentage change threshold p_t for development set:

1. Transformer models trained on the *Empathetic Dialogues* dataset from Section 4.1 (32To8 and Merged-8 for both RoBERTa-base and RoBETa-large) were used as the base models to classify emotions of the input utterances to get: a) top 2 emotion labels and their probability scores produced by the model, from which their probability difference (percentage change) was calculated; b) embeddings from the last layer to be used as the sentence representation.

2. Centroids of emotion labels were calculated with last layer embeddings of the training set from the *Empathetic Dialogues* dataset. Then, distances of last layer

Index	Ordered Sentences	True Label(s)	32To8	Merged-8
0	Why do classes give out so much work ?	Anger, Surprise	Sadness	**Anger**
1	I'm not just talking about moving to online , either .	Neutral	Joy	Sadness
2	Ever since my very first semester at college , my professors have been pilling on the reading homework and quizzes and assignments back to back to back .	Anger, Surprise	**Anger**	**Anger**
3	I'm now a sophomore starting the Fall semester , and I still spend 2-5 hours on homework PER CLASS for a week .	Fear, Anger	**Anger**	**Anger**
4	Half the time it's just writing notes from the reading , and other times it's just the stupidest things like a worksheet that I can't find the answers to .	Fear, Anger	**Anger**	**Anger**
5	And don' t get me started on Connect .	Neutral	Anticipation	Anticipation
6	Why do the readings have these concept questions with them ?	Anger, Surprise	Fear	Fear
7	It's just waaaaaaay too much .	Anger, Surprise	**Anger**	Fear

Table 4.3: Example emotion classification with baseline **BERT** models on a particular Reddit post (bolded text indicated correct prediction)

Index	Ordered Sentences	True Label(s)	32To8	Merged-8
0	Why do classes give out so much work ?	Anger, Surprise	**Surprise**	**Anger**
1	I'm not just talking about moving to online , either .	Neutral	Fear	Fear
2	Ever since my very first semester at college , my professors have been pilling on the reading homework and quizzes and assignments back to back to back .	Anger, Surprise	Anticipation	**Anger**
3	I'm now a sophomore starting the Fall semester , and I still spend 2-5 hours on homework PER CLASS for a week .	Fear, Anger	**Fear**	**Anger**
4	Half the time it's just writing notes from the reading , and other times it's just the stupidest things like a worksheet that I can't find the answers to .	Fear, Anger	**Anger**	**Anger**
5	And don' t get me started on Connect .	Neutral	Fear	Anticipation
6	Why do the readings have these concept questions with them ?	Anger, Surprise	**Surprise**	**Surprise**
7	It's just waaaaaaay too much .	Anger, Surprise	**Anger**	Fear

Table 4.4: Example emotion classification with baseline **RoBERTa-base** models on a particular Reddit post (bolded text indicated correct prediction)

Index	Ordered Sentences	True Label(s)	32To8	Merged-8
0	Why do classes give out so much work ?	Anger, Surprise	Sadness	**Anger**
1	I'm not just talking about moving to online , either .	Neutral	Joy	Sadness
2	Ever since my very first semester at college , my professors have been pilling on the reading homework and quizzes and assignments back to back to back .	Anger, Surprise	**Anger**	**Anger**
3	I'm now a sophomore starting the Fall semester , and I still spend 2-5 hours on homework PER CLASS for a week .	Fear, Anger	**Anger**	**Anger**
4	Half the time it's just writing notes from the reading , and other times it's just the stupidest things like a worksheet that I can't find the answers to .	Fear, Anger	**Anger**	**Anger**
5	And don' t get me started on Connect .	Neutral	Anticipation	Anticipation
6	Why do the readings have these concept questions with them ?	Anger, Surprise	Fear	Fear
7	It's just waaaaaaay too much .	Anger, Surprise	**Anger**	Fear

Table 4.5: Example emotion classification with baseline **BERT** models on a particular Reddit post (bolded text indicated correct prediction)

Index	Ordered Sentences	True Label(s)	32To8	Merged-8
0	Why do classes give out so much work ?	Anger, Surprise	**Anger**	**Anger**
1	I'm not just talking about moving to online , either .	Neutral	Fear	Fear
2	Ever since my very first semester at college , my professors have been pilling on the reading homework and quizzes and assignments back to back to back .	Anger, Surprise	**Anger**	Anticipation
3	I'm now a sophomore starting the Fall semester , and I still spend 2-5 hours on homework PER CLASS for a week .	Fear, Anger	**Anger**	**Fear**
4	Half the time it's just writing notes from the reading , and other times it's just the stupidest things like a worksheet that I can't find the answers to .	Fear, Anger	**Anger**	**Fear**
5	And don' t get me started on Connect .	Neutral	Anger	Fear
6	Why do the readings have these concept questions with them ?	Anger, Surprise	Fear	Sadness
7	It's just waaaaaaay too much .	Anger, Surprise	**Anger**	Fear

Table 4.6: Example emotion classification with baseline **RoBERTa-large** models on a particular Reddit post (bolded text indicated correct prediction)

Algorithm 1: Unsupervised Algorithm

Data: sentence list S from the same post, number of output labels N_{label}
Result: emotion label(s)
Parameters: Optimal number of standard deviations N_σ, consecutive distance threshold A_t, percentage change threshold p_t for development set ;
foreach *s in S* **do**
    Classify with Transformer baselines to get top 1 emotion e_1 and top 2 emotion e_2, percentage change p between probabilities of e_1 and e_2, and embedding v;
    Compute a distance list D between v and embeddings of all-emotion centroid list C;
    Get adjusted distance list $A = p * D$;
    Compute the distance d between current v and consecutive embedding v_c from previous emotion e_c;
    if *Every element of* $D \in (\mu_A - N_\sigma * \sigma_A, \mu_A + N_\sigma * \sigma_A)$ **then**
        if $d < A_t$ **then**
            Merged label e_m = previous or next emotion e_c
        else
            e_m = NEUTRAL
        end
    else
        $e_m = e_1$
    end
end
if $N_{label} == 1$ **then**
    return e_m;
end
if $N_{label} == 2$ **then**
    $y_1 = e_m$;
    if $e_m == e_c$ ***or*** $e_m ==$ *NEUTRAL* **then**
        y_2 = original label e_1 (Experiment 1) or NEUTRAL (Experiment 2)
    else
        $y_2 = e_2$ for $p <$ threshold p_t;
    end
    return y_1, y_2;
end

sentence embeddings from emotion centroids were calculated for each utterance.

3. Distances were then adjusted by multiplying probability differences as weights. This is found to be the best equation of adjusting distances after experiments with fixed threshold for probability differences. The adjusted distances were checked for even distributions using the standard deviation method. That is, the distances were considered evenly distributed when they were within a certain number of standard deviation N_σ around the distance mean. If the distances were evenly distributed, the corresponding input utterance was considered as ambiguous. For ambiguous predictions, distances between consecutive pairs of last layer embeddings were calculated. The ambiguous prediction was then merged to consecutive emotion label(s) if the distance was below a certain distance threshold. If the distance was above the distance threshold, the input utterance would be labeled as NEUTRAL. For non-ambiguous distances, the corresponding emotion did not merge and was the same as the top 1 emotion predicted by the Transformer models.

4. Different sets of refined emotion label(s) were returned according to the number of labels needed. This will be described in details in the following sections.

4.3.1 Single-Label Approach

The single-label approach produced 1 of the 9 merged emotion labels as the output emotion label. After the steps 1-3 described above, this approach simply returns the newly merged emotion labels in step 4. Figure 4.4 displays the pipeline of all the procedures and steps involved in this approach.

4.3.2 Two-Label Approach

Two-label approach follows the general pipeline described in steps 1-3 above and contained an extra hyper-parameter of probability difference. The approach produced up to 2 emotion label as the output. For this approach, step 4 of refining final emotion

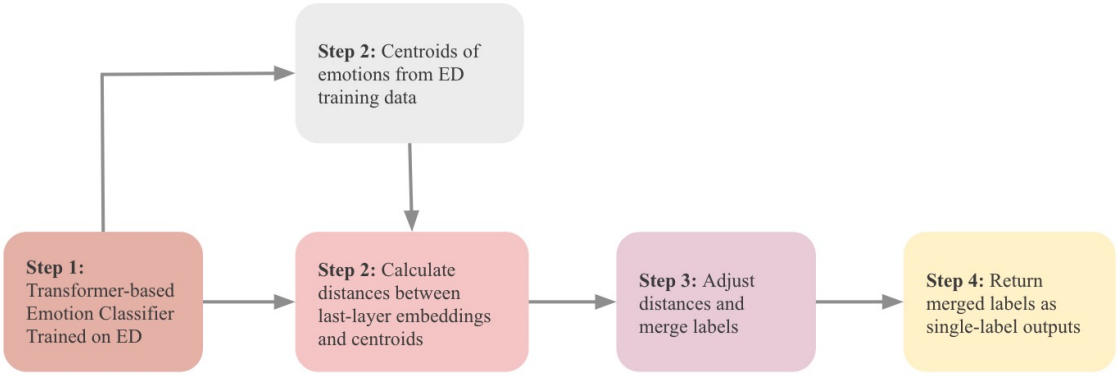

Figure 4.4: Pipeline of unsupervised single-label approach

predictions was different than that of the single-label approach (see Figure 4.5).

For input utterances identified as ambiguous in step 3, two experiments was conducted. One experiment considered the merged label from step 3 as the first emotion label and the original prediction by the Transformer baseline model as the second emotion label (Experiment 1). The other experiment considered the original prediction by the Transformer baseline model as the first emotion label and NEUTRAL as the second emotion label (Experiment 2).

For non-ambiguous input utterances, the prediction by the Transformer baseline model was considered as the first emotion label. The probability difference was used here to determine if the second possible emotion should be one of the output labels. If the probability difference was below a certain threshold, top 2 emotions would be the two output emotion labels.

4.3.3 Experiments and Evaluation

Unsupervised models described above were all evaluated with the self-annotated *Reddit College* test set. Evaluation metrics used for unsupervised approaches were consistent with the three metrics for this dataset as described in Section 4.2.3, including precision, recall, and F1 score.

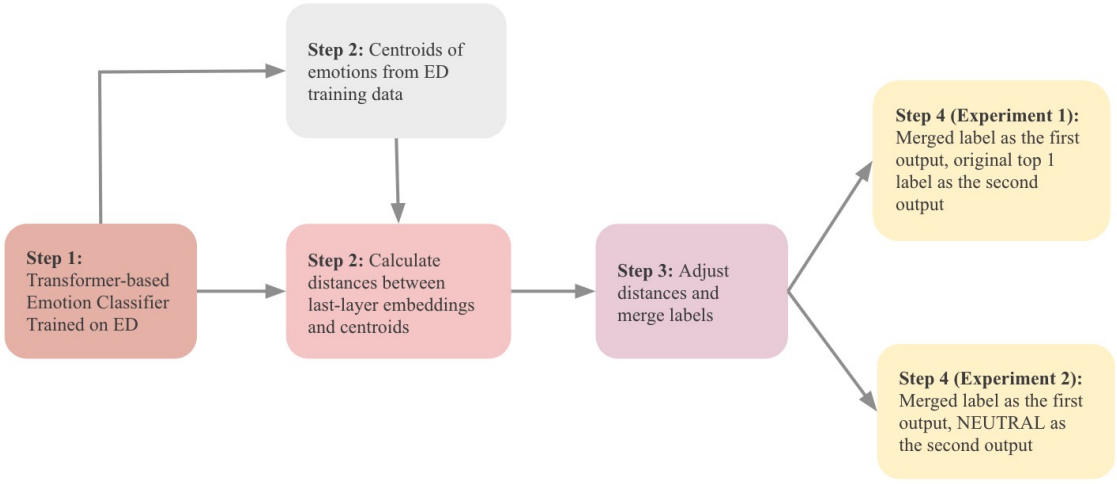

Figure 4.5: Pipeline of unsupervised two-label approach

According to the evaluation metrics for single-label approaches shown in Table 4.7, the model with the best performance was the 32To8 approach with RoBERTa-base, producing the highest precision, recall, and F1 score. Both the 32To8 and Merged-8 approaches with RoBERTa-large performed worse than those with RoBERTa-base.

Model	Approach	Precision	Recall	F1 Score
RoBERTa-base	**32To8**	**0.735**	**0.492**	**0.589**
	Merged-8	0.681	0.456	0.546
RoBERTa-large	32To8	0.708	0.469	0.565
	Merged-8	0.672	0.449	0.538

Table 4.7: Evaluation of single-label unsupervised models on the self-annotated *Reddit College* test set

For two-label approaches, two experiments were performed to compare and select the best model. Experiment 1, as described in Section 4.3.2, chose the merged label as the first output emotion label and the original top 1 prediction from the Transformer baseline models as the second output for ambiguous input utterances. According to results shown in Table 4.8, the 32To8 approach with RoBERTa-base outperformed

all other models with the highest precision, recall, and F1 scores. Overall, the 32To8 approach achieved higher F1 scores than the Merged-8 approach for both RoBERTa-base and RoBERTa-large models.

Model	Approach	Precision	Recall	F1 Score
RoBERTa-base	**32To8**	**0.643**	**0.602**	**0.622**
	Merged-8	0.568	0.534	0.550
RoBERTa-large	32To8	0.697	0.528	0.601
	Merged-8	0.564	0.528	0.545

Table 4.8: Evaluation of two-label Experiment 1 models

Experiment 2, as described in Section 4.3.2, chose the original top 1 prediction from the Transformer baseline models as the first output and the original top 2 prediction with a certain probability difference threshold as the second output. According to results shown in Table 4.9, the 32To8 approach with RoBERTa-base and RoBERTa-large performed roughly the same with the highest F1 scores. The difference lies in that the 32To8 RoBERTa-base had a higher precision score than the RoBERTa-large model. Overall, the 32To8 approach achieved higher precision, recall, and F1 scores than the Merged-8 approach for both RoBERTa-base and RoBERTa-large models in this experiment.

Model	Approach	Precision	Recall	F1 Score
RoBERTa-base	**32To8**	**0.667**	**0.544**	**0.599**
	Merged-8	0.573	0.528	0.549
RoBERTa-large	**32To8**	**0.636**	**0.566**	**0.599**
	Merged-8	0.558	0.534	0.546

Table 4.9: Evaluation of two-label Experiment 2 models

4.3.4 Results and Analysis

When two experiments were compared, the 32To8 RoBERTa-base model achieved the highest F1 Score among all models for unsupervised approaches overall. Also, the 32To8 RoBERTa-base model from Experiment 1 achieved a higher F1 score than the one from Experiment 2, which was therefore used when analyzing results in the following parts.

From results produced by single-label unsupervised models as shown in Table 4.10, the 32To8 RoBERTa-base model correctly classified 7 out of 8 sentences in the example post. The sentence corresponded with the only wrong prediction was also classified wrongly by the 32To8 RoBERTa-large model. The model with the most number of wrong predictions in this example was the 32To8 RoBERTa-large model (see Table 4.11), and it classified the first sentence with FEAR while all other models classified it as ANGER. Similarly, for other sentences such as sentences 3, 5, and 7, it predicted completely different results compared with other models. Merged-8 models performed similarily for both RoBERTa-base and RoBERTa-large models in this example, as they had the same number of correct predictions.

Results predicted by two-label unsupervised models were shown in Tables 4.12 and 4.13. With up to 2 possible predictions, the 32To8 RoBERTa-base model still had the highest number of correct predictions (see Table 4.12). The two wrong predictions produced by this model were also not completely different from the true labels for the last two sentences, as they were all negative emotions. There was one unreasonable prediction produced by the Merged-8 RoBERTa-base model, which was sentence 7, since the emotion of the sentence was not JOY at all. The performance of predictions made by RoBERTa-large models for the two-label classification here was similar to that for the single-label classification (see Table 4.13). While these models had some correct predictions, they also produced similar number of wrong predictions.

Index	Ordered Sentences	True Label(s)	32To8	Merged-8
0	Why do classes give out so much work ?	Anger, Surprise	**Anger**	**Anger**
1	I'm not just talking about moving to online , either .	Neutral	**Neutral**	Fear
2	Ever since my very first semester at college , my professors have been pilling on the reading homework and quizzes and assignments back to back to back .	Anger, Surprise	**Anger**	**Anger**
3	I'm now a sophomore starting the Fall semester , and I still spend 2-5 hours on homework PER CLASS for a week .	Fear, Anger	Neutral	**Anger**
4	Half the time it's just writing notes from the reading , and other times it's just the stupidest things like a worksheet that I can't find the answers to .	Fear, Anger	**Anger**	**Anger**
5	And don' t get me started on Connect .	Neutral	**Neutral**	Anticipation
6	Why do the readings have these concept questions with them ?	Anger, Surprise	**Anger**	**Surprise**
7	It's just waaaaaaay too much .	Anger, Surprise	**Anger**	Fear

Table 4.10: Example 1-label classification with baseline **RoBERTa-base** models on a particular Reddit post (bolded text indicated correct prediction)

Index	Ordered Sentences	True Label(s)	32To8	Merged-8
0	Why do classes give out so much work ?	Anger, Surprise	Fear	**Anger**
1	I'm not just talking about moving to online , either .	Neutral	Fear	Fear
2	Ever since my very first semester at college , my professors have been pilling on the reading homework and quizzes and assignments back to back to back .	Anger, Surprise	Anticipation	**Anger**
3	I'm now a sophomore starting the Fall semester , and I still spend 2-5 hours on homework PER CLASS for a week .	Fear, Anger	Anticipation	**Anger**
4	Half the time it's just writing notes from the reading , and other times it's just the stupidest things like a worksheet that I can't find the answers to .	Fear, Anger	Neutral	**Anger**
5	And don' t get me started on Connect .	Neutral	Fear	Anticipation
6	Why do the readings have these concept questions with them ?	Anger, Surprise	Fear	**Surprise**
7	It's just waaaaaaay too much .	Anger, Surprise	**Anger**	Fear

Table 4.11: Example 1-label classification with baseline **RoBERTa-large** models on a particular Reddit post (bolded text indicated correct prediction)

Index	Ordered Sentences	True Label(s)	32To8	Merged-8
0	Why do classes give out so much work ?	Anger, Surprise	**Anger**	**Anger**,Sadness
1	I'm not just talking about moving to online , either .	Neutral	**Neutral**	Fear, Anticipation
2	Ever since my very first semester at college , my professors have been pilling on the reading homework and quizzes and assignments back to back to back .	Anger, Surprise	**Anger**	**Anger**
3	I'm now a sophomore starting the Fall semester , and I still spend 2-5 hours on homework PER CLASS for a week .	Fear, Anger	**Fear, Anger**	**Fear, Anger**
4	Half the time it's just writing notes from the reading , and other times it's just the stupidest things like a worksheet that I can't find the answers to .	Fear, Anger	**Anger**	**Anger**
5	And don' t get me started on Connect .	Neutral	**Neutral**, Anticipation	Joy, Anticipation
6	Why do the readings have these concept questions with them ?	Anger, Surprise	**Anger**, Fear	Fear, **Surprise**
7	It's just waaaaaaay too much .	Anger, Surprise	**Anger**	Fear

Table 4.12: Example 2-label classification with baseline **RoBERTa-base** models on a particular Reddit post (bolded text indicated correct prediction)

Index	Ordered Sentences	True Label(s)	32To8	Merged-8
0	Why do classes give out so much work ?	Anger, Surprise	**Anger**, Fear	**Anger**
1	I'm not just talking about moving to online , either .	Neutral	Fear	Fear, Anticipation
2	Ever since my very first semester at college , my professors have been pilling on the reading homework and quizzes and assignments back to back to back .	Anger, Surprise	**Anger**, Anticipation	**Anger**, Fear
3	I'm now a sophomore starting the Fall semester , and I still spend 2-5 hours on homework PER CLASS for a week .	Fear, Anger	Anticipation	**Fear, Anger**
4	Half the time it's just writing notes from the reading , and other times it's just the stupidest things like a worksheet that I can't find the answers to .	Fear, Anger	Neutral, **Anger**	**Anger**
5	And don' t get me started on Connect .	Neutral	Fear	Anticipation
6	Why do the readings have these concept questions with them ?	Anger, Surprise	**Anger**, Fear	**Surprise**
7	It's just waaaaaaay too much .	Anger, Surprise	**Anger**	**Anger**, Fear

Table 4.13: Example 2-label classification with baseline **RoBERTa-large** models on a particular Reddit post (bolded text indicated correct prediction)

Chapter 5

Conclusion and Future Work

As stated in the results and analysis section in Chapter 4 for different approaches, the 32To8 RoBERTa-base model generally yields the best result for the Reddit dataset. Since the model had the highest performance scores, we would be able to use it for further emotion analysis on a larger Reddit data.

The 32To8 RoBERTa-base model successfully adds the neutral emotion label in addition to the original 8 emotion labels, which does better than many existing emotion classifiers that only contain six basic emotions or "negative, positive, neutral" only labels. It also allows us to detect emotions based on surrounding sentences, which provide a context-aware tool for classifying emotions for textual data. Further, the approach we developed was founded in the higher education sector. It therefore provided a well-performed tool for future analysis and applications in developing empathetic chatbot agents or services in higher education institutions.

Overall, the context-based unsupervised approach for emotion detection does satisfy our expectations at the beginning, as described in the thesis statement in Chapter 1. The two-label emotion detection model performed better than the one-label emotion detection models as well as baseline models. The unsupervised approach also refined neutrality as an additional emotion label that does not exist when using base-

line models. The model also promotes many potentials in future applications.

5.1 Future Work

There are many future directions this thesis could help with. This section discusses some of the future work that could be done to apply the emotion classifier developed in this thesis.

5.1.1 Emotion Analysis on Reddit

With the emotion classifier developed, we may conduct further emotion analysis on a larger Reddit corpus to perform text summarization and emotion-cause pair extraction. For example, we may construct storylines according to different topics of the college aspect to provide insights in higher education. Then we would be able to know structure such as what and how people would normally start with or end with when posting concerns about college. This type of information may help in developing chatbots in higher education. We may also use emotion classifier as a tool of automatically generating reasoning structures and content structures by associating concrete sentences with emotions. That is, we may use these emotional structures to summarize text by pairing sentences with their corresponding emotions, which is innovative when compared with existing methods of summarizing text based on topics.

We may also extend the emotion classifier to other areas of contents on Reddit using similar approaches, such as healthcare. In this way, a more comprehensive emotion classifer would be developed with 9 emotions, which is a larger number of emotion categories than many existing emotion classifiers.

5.1.2 Applications in Dialogue Systems

This emotion classification approach may be combined with dialogue system strategies to make more empathetic chatbots. For example, comments of Reddit posts may further be incorprated into the emotion analysis to generate reasonable situation-reaction pairs according to emotions. These emotional pairs would help in determining whether a generated response by the chatbot is meaningful from the emotion perspective, thus serving the evaluation and generation purposes. Another possibility is that we could use the emotion-cause pairs extracted from the data to train models that generate responses based on those pairs so that the responses are not only empathetic but also interpretable.

Bibliography

[1] Ijabadeniyi A Agbehadji IE. Approach to sentiment analysis and business communication on social media. in: Fong s, millham r (eds) bio-inspired algorithms for data streaming and visualization, big data management, and fog computing, springer tracts in nature-inspired computing. *Springer*, 2021.

[2] Mangal H-Putluri K Reid B Hanna G Sarkar M Al Ajrawi S, Agrawal A. Evaluating business yelp's star ratings using sentiment analysis. *Materials Today: Proceedings*, 2021.

[3] Sharma N. Chauhan, P. and G. Sikka. The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, 2021.

[4] Jinho Choi. Reddit college. `https://github.com/emorynlp/reddit-college`, 2021.

[5] Movshovitz-Attias D. Ko J.-Cowen A. Nemade G. Demszky, D. and S. Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

[6] Chang-M.W. Lee K. Devlin, J. and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 1999.

[8] Sebastiani F Esuli A. Sentiwordnet: a publicly available lexical resource for opinion mining. *Proceedings of the LREC*, 2006.

[9] El-Haggar N. Fathy, S. and M.H. Haggag. A hybrid model for emotion detection from text. *International Journal of Information Retrieval Research (IJIRR)*, 2017.

[10] Finch-J.D. Huryn D. Hutsell-W. Huang X. He H. Finch, S.E. and J.D. Choi. An approach to inference-driven dialogue management within a social chatbot. *arXiv preprint arXiv:2111.00570*, 2021.

[11] Berton L Garcia K. Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. *Appl Soft Comput*, 2021.

[12] Hong P. Shen S. Majumder-N. Mihalcea R. Ghosal, D. and S. Poria. Cider: Commonsense inference for dialogue explanation and reasoning. *arXiv preprint arXiv:2106.00510*, 2021.

[13] Rozenshtein P. Gollapalli, S.D. and S.K. Ng. Ester: Combining word co-occurrences and word associations for unsupervised emotion detection. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020.

[14] V. Putnala H. K. K., A. K. Palakurthi and A. Kumar K. Smart college chatbot using ml and python. *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2020.

[15] Tianyang Zhang Xiaoyan Zhu Hao Zhou, Minlie Huang and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[16] R. Harré and W.G. eds. Parrott. The emotions: Social, cultural and biological dimensions. *Sage*, 1996.

[17] Wei L. Hu, D. and X. Huai. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*, 2021.

[18] Russell JA. A circumplex model of affect. *J Pers Soc Psychol*, 1980.

[19] Lee Y Kwon O Jang HJ, Sim J. Deep sentiment analysis: mining the causality between personality-value-attitude for analyzing business ads in social media. *Expert Syst Appl 40(18):7492–7503*, 2013.

[20] Su H. Shen X. Li W. Cao Z. Li, Y. and S. Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

[21] Zhao H. Su H. Qian Y. Li, M. and P. Li. Emotion-cause span extraction: a new task to emotion cause identification in texts. *Applied Intelligence*, 2021.

[22] Xu P. Winata G. I. Siddique F. B. Liu Z. Shin J. Fung P. Lin, Z. Caire: An end-to-end empathetic chatbot. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[23] Jiang J. Xiong C. Yang Y. Liu, C. and J. Ye. Towards building an intelligent chatbot for customer service: Learning to respond at the appropriate time. *In Proceedings of the 26th ACM SIGKDD international conference on Knowledge Discovery Data Mining*, 2020.

[24] Ott M. Goyal N. Du J. Joshi M. Chen D. Levy O. Lewis M. Zettlemoyer L. Liu, Y. and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[25] Delgado-Vera C. Solís-Avilés E. Espinoza A.H. Ortiz-Zambrano J. Mite-Baidal, K. and E. Varela-Tapia. Sentiment analysis in education domain: A systematic

literature review. *In International conference on technologies and innovation*, 2018.

[26] Turney PD Mohammad SM. Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. *Association for Computational Linguistics*, 2010.

[27] Ashraf N.-Ahmed F.S. Ferzund-J. Shahzad B. Mustafa, R.U. and A. Gelbukh. A multiclass depression detection in social media based on sentiment analysis. *In 17th International Conference on Information Technology–New Generations (ITNG 2020)*, 2020.

[28] D. A. Patel N. P. Patel, D. R. Parikh and R. R. Patel. Ai and web-based human-like interactive university chatbot (unibot). *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2019.

[29] Collins A Ortony A, Clore GL. The cognitive structure of emotions. *Cambridge, MA: Cambridge University Press*, 1990.

[30] R. Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 2001.

[31] Hazarika D. Majumder N. Naik G. Cambria E. Poria, S. and R. Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[32] Hazarika D. Majumder N. Naik G. Cambria E. Poria, S. and R. Mihalcea. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020.

[33] Cambria E Hussain A Huang G-B Poria S, Gelbukh A. Emosenticspace: a novel framework for affective common-sense reasoning. *Knowl-Based Syst*, 2014.

[34] Plutchik R. A general psychoevolutionary theory of emotion. *Elsevier*, 1980.

[35] Smith E. M. Li M. Rashkin, H. and Y.-L. Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[36] Liu H.C.T. Huang Y.H. Wu J. Saravia, E. and Y.S. Chen. Carer: Contextualized affect representations for emotion recognition. *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018.

[37] Xu P. Madotto A. Shin, J. and P. Fung. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. *arXiv preprint arXiv:1906.08487*, 2019.

[38] Pandey S Singh M, Jakhar AK. Sentiment analysis on the impact of coronavirus in social life using the bert model. *Soc Netw Anal Min*, 2021.

[39] Valitutti A Strapparava C. Wordnet affect: an affective extension of wordnet. *Lrec*, 2004.

[40] Wisniewski H. Halamka J.D. Kashavan M.S. Vaidyam, A.N. and J.B. Torous. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 2019.

[41] Shazeer N. Parmar N. Uszkoreit J. Jones L.-Gomez A.N.-Kaiser Ł. Vaswani, A. and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017.

[42] Ke Wang and Xiaojun Wan. Sentigan: generating sentimental texts via mixture

adversarial networks. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018.

[43] C.M. Whissell. The dictionary of affect in language. *In The measurement of emotion*, 1989.

[44] Ekaterina Svikhnushina Yubo Xie and Pearl Pu. A multi-turn emotionally engaging dialog model. *arXiv*, 2019.

[45] Dinan E. Urbanek J. Szlam A. Kiela D. Zhang, S. and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[46] Zhang Z. Li J. Huang Y. Zhu, P. and H. Zhao. Lingke: A fine-grained multi-turn chatbot for customer service. *arXiv preprint arXiv:1808.03430*, 2018.