Utrecht University

Text-based empathy detection on social media

Master Artificial Intelligence

Nikolaos Bentis
6662889
n.bentis@students.uu.nl

First supervisor: dr. Dong Nguyen
Daily supervisor: Anna Wegmann MSc
Second supervisor: prof. dr. Albert Salah

July 12, 2021

Acknowledgement

I would like to thank dr. Dong Nguyen and Anna Wegmann for their continuous support, insightful suggestions and feedback. I would also like to acknowledge my fellow students and friends, Giannos and Bjorn, for motivating each other and for their collaboration. In addition, I would like to thank my friend Fotis, who inspired me to begin my studies at Utrecht University and supported me. Finally, I would like to thank my family, my partner and friends who sympathized with me and helped me complete my studies.

Abstract

The advancement of text-based empathy detection would be beneficial for the progress of affective computing. Affective computing is concerned with creating systems with emotional understanding and empathy is a key aspect of emotional intelligence. Also, empathy detection as a tool has many applications. However, only recently researchers started to focus on this topic, with most studies focusing on counseling data or on social media centered around psychological support. This study, first, takes a computational approach on the "Reactions to news stories" dataset, created by Buechel et al. (2018), with the usage of Transformer models. The pre-trained Transformer models of BERT and RoBERTa were fine-tuned on the data after a thorough hyper-parameter selection phase. In addition, the thesis explored data augmentation methods, but they did not improve performance on this task. During the model creation phase, the Transformer models improved approximately 10% on top of the baselines (CNN, FNN, Ridge regression), without using data augmentation methods. I conclude that Transformers are capable of predicting the EC and PD scores, even though the data had increased difficulty due to the sample size and because the scores were self-evaluated by the commenters. Additionally, this thesis investigates the differences between Reddit and Twitter on empathetic concern (EC) and personal distress (PD), using the selected BERT balanced model. The selected model was applied to user comments from Reddit and Twitter on the same news articles. This data were gathered during this dissertation. The results showcase significantly higher scores of EC and PD on tweets compared to Reddit comments on the same news articles. Further researcher should be made to investigate the reasons that lead to users having different behavior.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Text-based empathy detection

Why do we need systems that automatically detect empathy? This question falls under the scope of affective computing, the field of study that focuses on creating systems that understand human emotion. According to Rosalind Picard, who coined the term affective computing, there are multiple areas, such as education, healthcare and human-computer interaction, that will benefit greatly from having systems with emotional intelligence (Picard, 1997). An example would be applications that detect users' emotions and dynamically alter their behaviour to better facilitate the needs of each user better, in computer-assisted learning. The importance of empathy in affective computing is linked to psychology, as the key aspects of emotional intelligence are considered to be the abilities of understanding and expressing emotions and empathy (Mayer et al., 2008). Therefore, for machines to have emotional intelligence they would have to detect and simulate empathy. In contrast with its importance, the early research around affective computing in the field of natural language processing (NLP) did not prioritize empathy or other complex tasks. This trend has shifted with the study of more complex psychological constructs, such as humor (Taylor, 2009), irony (Reyes et al., 2012), sarcasm (Joshi et al., 2017; Mukherjee and Bala, 2017) and empathy. One of the reason for this shift are the recent advances of NLP, like Recurrent neural networks, word embeddings and the Transformers, that can produce more powerful and efficient models able to capture these complex concepts.

Social media and affective computing. When affective computing emerged in the year of 1997, society was functioning differently compared to today. Social media platforms did not concern the majority of people. Today, platforms, like Facebook, Twitter and Reddit, play a central role in our lives. Close to 4 billion people are members of at least one social media platform and that number grows daily [1]. People use social media for many tasks, from communicating with friends to getting their daily news coverage. Two-thirds of Americans report that they get at least some of their news updates on social media (Shearer and Gottfried, 2017). Even though this can be beneficial, social media have been linked to have various negative effects on human psychology and society, ranging from depression (Scherr and Brunet, 2017), to social anxiety (Primack et al., 2017), to targeted manipulation for influencing

[1] https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media

election results (Timberg, 2017). Moreover, studies have shown that different social media platforms have different effects on people. Users tend to adjust their behavior depending on the platforms they use, because they have a different audience on each platform (Davidson and Joinson, 2021) or because they have a different expectations (Velasquez and Rojas, 2017). So it is evident that these platforms, which are visited by a big percentage of the global population, affect us and in many cases without our knowledge. Thus, there is a need for studying and monitoring the effects that social media have on our psychology, but also on our society as a whole, and affective computing could facilitate it, as its focus is the creation of systems that understand human emotion. This makes empathy detection important for two reasons. First for the progression of affective computing and second for monitoring the effects of social media. Cases where users experience negative effects from social media platforms, should be investigated and resolved.

The situation of text-based empathy detection. In my thesis I am interested in empathy but specifically text-based empathy, as people mostly interact using texts on social media platforms. What is the situation on text-based empathy detection? Text-based empathy detection emerged 9 years ago. In the first half of this period the production of papers around text-based empathy detection was minimal; it revolved around counseling data and mainly originated by the research group of Xiao (Xiao et al., 2012, 2015). This situation started to change the last five years with an increase on the research output around the topic (Sharma et al., 2020; Khanpour et al., 2017; Zhou and Jurgens, 2020; Xiao et al., 2012, 2015; Gibson et al., 2015; Perez-Rosas et al., 2017). Although there is an increase on literature produced, the published work showcases a problematic situation, which is the ambiguity around the term of empathy, as studies have been used multiple conceptualizations. In addition, although the focus of most of the newer studies revolves around social media, it mainly concerns mental health social media.

Ambiguity of empathy. Cuff et al. (2014) reviewed literature from psychology and neuroscience about empathy and found 43 different conceptualizations of empathy. The research on text-based empathy detection inherits this ambiguity, as there have been used numerous different definitions, annotation methods and frameworks. "Showing an active interest by the therapist to understand what the client is feeling" is the definition used in most of the works with counseling data (Xiao et al., 2012, 2015; Gibson et al., 2015; Perez-Rosas et al., 2017). Empathetic concern (EC) (Batson, 1987), which includes concepts of sympathy, warmth and compassion, is a definition used mostly on social media data (Litvak et al., 2016; Abdul-Mageed et al., 2017; Buechel et al., 2018; Sedoc et al., 2019). Feeling the same emotion as the person you observe or communicate is empathy for others (Khanpour et al., 2017; Zhou and Jurgens, 2020). Finally, other researchers implemented a novel definition/framework of empathy called EPITOME, specifically in the context of empathy on text (Sharma et al., 2020). In my thesis, I follow the empathetic concern definition (EC), as it is the most adopted definition in studies on social media.

Sources of data. The first papers on empathy detection on text studied empathy on the transcripts of conversations between a therapist and their patient (Xiao et al., 2012, 2015). In the last five years, the focus has transitioned on social media data, as

their effects on our lives have drawn attention and because this type of data are easier to collect. The data used are either user comments and posts or conversations, both synchronous and asynchronous. Some produced research on analyzing user posts on social media platforms, but they included additional features, such as age, gender, likes, and not only textual (Litvak et al., 2016; Abdul-Mageed et al., 2017). Others focused on asynchronous text conversations (post and response) on mental health social media platforms, such as Reddit and Talklife (Khanpour et al., 2017; Sharma et al., 2020; Zhou and Jurgens, 2020; Hosseini and Caragea, 2021). Buechel et al. (2018) worked on predicting the empathetic concern (EC), which was annotated by social media users that first read a news article, then expressed how they felt and then commented on what they have read. This is the only dataset for empathy detection where users annotated themselves, in contrast to the rest of the datasets that used third parties to annotate the empathy level of the examined texts.

1.2 Research questions

Research question 1. The dataset created by Buechel et al. (2018), captures the self-felt levels of empathetic concerns and personal distress of users of social media. The way it is created makes this dataset ideal for usage on social media studies. This work, though, was published close to the birth of the Transformer architecture (Vaswani et al., 2017) and the state of the art models of today have not been applied yet on the dataset. The design of Transformers with the attention mechanism is ideal for capturing the context of longer texts, where other architectures struggle. These reasons lead me to my first research question.

I take a computational approach, using the Transformer architecture, on the empathetic reactions dataset published by Buechel et al. (2018). Therefore, I propose the following research question:
What is the performance of the Transformer architecture on predicting the empathetic concern (EC) and personal distress (PD) scores of reaction comments on news articles?
Due to the small number of samples in the dataset (1860) I also investigate the performance on the dataset with the usage of data augmentation methods.

Research question 2. As already discussed, one reason of why people use social media is to get their news updates. But social media lead to people changing their behavior depending on the platform and there are evidence that they affect our behavior in many ways. To the best of my knowledge, there are no studies trying to detect the different effect of social media on people who read news stories on these platforms. For this reason, I investigate the scores of empathetic concern (EC) and personal distress (PD) on the users of Twitter and Reddit, who use these platforms to get their news, using the empathy detection system produced on the first research question. Because Twitter and Reddit differ along many characteristics, I hypothesize that we will observe different EC and PD levels. Reddit includes sub-groups, called subreddits, that people select to belong in and follow the content posted on them. These groups have their own rules and their own moderators to monitor them. On the other hand, on Twitter each person chooses the people or organizations they

follow and they do not follow content through groups. Additionally, Twitter has a limit on the length of the text, in contrast to Reddit. According to Priya (2018), these and other characteristics reduce bias and extreme views in the comments of Twitter compared to Reddit.

I hypothesize that Twitter users show less personal distress and more empathetic concern when reacting to news stories compared to Reddit users from popular news subreddits. I propose this additional research question:
Do the comments reacting to news stories on Twitter express less personal distress and more empathetic concern compared to the comments on the same news articles on popular news subreddits?

1.3 Results

Main results In this work, I show that two pre-trained models of the Transformer architecture, BERT (Liu et al., 2019) and RoBERTa (Liu et al., 2019), have an approximately 10% improvement on predicting the EC and PD scores, compared to the baselines produced by Buechel et al. (2018). The data augmentation methods did not improve the performance of these models further. Continuing, to answer the second research question I created a dataset that includes Reddit comments and tweets on the same news articles. For this dataset I give access upon request and details of it are seen on the appendix. Only half of my hypothesis held true after the experiments on the second question. The tweets have significantly higher scores of EC and PD compared to Reddit comments on the same news stories, so the part of my hypothesis for less personal distress on Twitter does not hold.

Implications The context of this thesis does create some implications. First, it showcases that the "reactions to news stories" dataset, by Buechel et al. (2018), should be expanded as the number of utterances is low. Next, models of the transformers architecture show that they are better suited on this task compared to different architectures that are widely used (Convolution neural network, feedforward neural network). Moreover, the experimental results show that data augmentation methods have limitations. In addition, the results of the second research question shine a light on the different behaviors of users on different social media. This should draw further research on monitoring the effects of social media and on exploring which aspects are responsible for some behaviors. Finally, the created dataset could be used with the intention of comparing these two platforms or the included subreddits (news, worldnews, politics).

Relevance to Artificial Intelligence This work is relevant to the field of Artificial Intelligence for a number of reasons. First, the research topic of empathy detection has gained traction the last five years and my work will be a part of this. In addition, I use the state of the art models in natural language processing (Transformers). Moreover, I created a new dataset that could be used for additional studies.

Ethics considerations In my thesis, I am concerned with affective computing applied to social media data. These areas are sensitive as they are many ethical issues. A discussion of ethical considerations is required before continuing on the next chapters. Affective computing systems can be ethical and unethical, depending on the use case. I consider unethical the usage of affective computing systems for user manipulation without any explicit consent or without notifying users about the algorithms that work on the background. An example of an unethical system is an algorithm that studies the emotional status of a user to promote content that will keep the user engaged on the platform. In my case, I use affective computing to investigate the effects of social media platforms with the intention of showcasing that social media have different effects and should be monitored. Thus, I consider this usage ethical. In addition, during the second research question I created a dataset with data from Reddit and Twitter. For both I followed the terms of usage for using their APIs to acquire the data. Moreover, the data, which I will give access after receiving a informative request, will not carry the usernames or comment IDs. People, though, will still able to still track users by using the whole comment as a query. For the topic of empathy I think transforming the comments to avoid this will harm the quality of the dataset, thus I will only give access to the data to researchers that I agree with the scope of their work. At a final note, I understand that there are use cases with these type of data that can be on the unethical side, for example identifying patterns that keep users engaged on the platforms even if these affect them negatively. Applications that aim for human manipulation for profit should be regulated by independent third parties.

Following chapters. In the remaining chapters of this thesis, I first review literature around the areas of affective computing and empathy detection, before I focus on answering the two research questions. Specifically, in chapter 2, I present the field of affective computing and discuss the situation around the definition of empathy. In addition, I describe the work on text-based empathy detection and the NLP methods, which I will use in my experiments, such as Transformers and data augmentation. Chapter 3 and chapter 4 are dedicated on answering the first and second research question. Finally, in chapter 5 I present the discussion and conclusion of my thesis, along with opportunities for future research.

Chapter 2

Related work

This section discusses the literature around empathy detection. First, I introduce affective computing and how it led to empathy detection in section 2.1.. Second, I examine the different theories, coming both from psychology and neuroscience, about what empathy is and the aspects the researchers disagree on in section 2.2.. Then, the work about text based-empathy detection is explained thoroughly and the published machine learning systems and datasets are mentioned on section 2.3.. Finally, I present the Transformer architecture and data augmentation methods for NLP on section 2.4..

2.1 Affective computing

Introduction to the field. Affective computing is the field interested in the study and development of computer systems that are able to detect, interpret, influence and simulate human emotions (Daily et al., 2017). It is a multidisciplinary field that includes psychology, sociology, physiology, computer science and linguistics (Tao and Tan, 2005). The first mention of this topic of research was made by Rosalind Picard in 1995, who coined the term in her published paper with the same title (Picard, 1997) and led the way for other researchers to follow. The ideas of affective computing were mentioned before but she was the first to organize them and give a purpose to the field, which was to create systems with emotional intelligence. key aspects of emotional intelligence is the ability of expressing and understanding emotions and empathy (Mayer et al., 2008). This area of research, as it is relative new, has mainly revolved around the basic human emotions — happiness, sadness, anger, disgust, fear and surprise — or the expressed sentiment of an situation or phrase —positive, neutral, negative —, but more complex concepts have started to emerge lately.

Different aspects of affective computing. Affective computing can be separated into two different tasks, the affective understanding/sensing and the affective generation. The second builds on top of the first though, as one system needs to understand and differentiate what actions express which emotion to simulate human-like behavior (Strauss et al., 2005). Affective understanding is a complex task, as humans express their emotions in many different ways depending on the way they communicate at that moment, for example with their facial expressions, their physical stance, the way they speak, the way they write and many more. This situation

has sprung research in different areas and for different purposes. Some notably areas of research are facial recognition and speech recognition. In facial recognition, camera sensors capture our facial expression, as the way we our face reacts gives a lot of information about our mental state. In speech recognition, the way we speak —speed, tone, pitch — and how we construct our sentences indicate how we feel (Daily et al., 2017). In this thesis, I will focus on natural language processing, the automatic processing and analysis of human language, which increased its importance as people interact more and more online between each other or with chatbots the field.

Situation in the field. The most popular tasks of affective computing in NLP are sentiment analysis and emotion analysis, which are usually being treated as classification tasks. Although emotion analysis is more complex as a task due to having more classes and the separation between them in some cases is complex, both tasks progressed at the same time and with the same techniques, from keyword matching to rule-based classification to machine learning algorithms to deep learning systems or even hybrid ones (Alswaidan and Menai, 2020). Georgiou et al. (2011) published one of the first machine learning system that used only textual features to encode the behavioral activity of a person, but not their empathy. An utterance could be neutral or show acceptance, blame, positive, negative, humour or sadness. The positive results of the model showed that models using only textual information can distinguish and classify complex behavioral and psychological phenomena. Multiple applications of similar systems exist on healthcare support, on advertisement, on social media platforms, on finance, on education and in many more areas. Some of the applications are the usage of chatbots that understand the emotion of the person so communicate properly with users or systems that detect behaviours on a population level and extract mass opinions (Kratzwald et al., 2018). Until recently, these were the most complex tasks that people working on affective computing focused on, but with the creation of attention models and Transformers this changed (Vaswani et al., 2017). More details of how Transformers made NLP advance on section 2.4..

Ethical considerations. A final thing to consider about affective computing is its ethical aspect. Picard in her first publication about affective computing raised a warning about the "tragic consequences" to follow if a computer system is able to express itself emotionally and has certain capabilities (Picard, 1997). Affective computing systems can be both ethical and unethical depending on the situation that they are placed in and there should be restrictions made on their usage. For example, when an affective computing system (facial recognition or a speech recognition) is used to understand someone's emotions at that moment it may be a huge privacy invasion and the human party should know that they are under analysis. Additionally, by understanding how humans react and respond to every situation affective computing systems could be used to manipulate the person's emotions. This could be beneficial in a healthcare support system designed to help people, but in numerous situations it shouldn't be allowed. These are some of the ethical implications of having powerful systems able to understand humans and these need to be acknowledged, so the field could move on ethical ways (Daily et al., 2017).

2.2 Empathy definitions

The situation of the concept of empathy. What is empathy? If that question is asked most people would have an idea of what it is, but the answers would probably not align. This situation is observed also in the literature coming from different fields of study, such as developmental psychology, social psychology, cognitive neuroscience and clinical neuropsychology (Decety and Jackson, 2004), where researchers have different ideas about what empathy is. Cuff et al. (2014) provide an overview of different definitions of empathy written in English, by examining key papers and their references. In that way, they produced 43 different definitions, which they group into 8 empathy-concepts. For a term that has a significant role in social work and counseling, it is problematic to find so many definitions and disagreements between researchers for a number of reasons. First, someone that studies the literature of empathy must be very careful of the definition that is used in each one of the readings, to know exactly what that reading is measuring or discussing, which is time consuming and could lead to false assumptions (Gerdes et al., 2010). Second, while many researchers have worked on empathy, the reality is that people study different concepts and the quality of research is affected by the disagreement. An example to that is that the ambiguity of the term is responsible for inconsistent and non wide accepted measures of empathy (Wispe, 1986) (Eisenberg and Strayer, 1987). Supporters of one definition do not think that others measure empathy and vise versa, so the findings and conclusions of one is not accepted by the rest, which stalls the progression of the field.

Empathy as a greater emotional category. There are some key aspects responsible for the ambiguity around empathy. The main factor though, is how one relates the term with other complex psychological concepts, such as sympathy, compassion and others. Batson (1987) and Preston and de Waal (2002) claim that empathy is a category of emotional responses that includes other psychological concepts, like sympathy, compassion, tenderness and others. Batson names this category of emotional responses *empathetic concern* (EC) and indicates that there is an opposite category named *personal distress* (PD), that differentiates from the EC based on the motivation of the expressed emotions. These two emotional categories do not exclude each other and when they appear together they combine to produce a stronger emotional arousal. EC is rooted in altruistic motives, the person expressing empathy seeks to easy the suffering of the observed. PD has egoistic motives, the observer suffers from the feelings produced by what they experienced and acts with the intention of minimizing them. The author is one of the most cited researchers on the topic, but this is a controversial definition of empathy. The controversy originates from the fact that others do not see empathy as a category of emotional responses that include sympathy, but as an emotional response on its own that relates to the other concepts.

How do these concepts relate? Ickes (2003) by gaining inspiration from Becker (1931), tried to explain why and how the complex psychological terms, like empathy and sympathy, are connected in our cognition, thus making their distinction a difficult task that produces conflicts in research. He claims that empathy and other similar concepts, such as sympathy, unipathy, transpathy, mimpathy and compathy,

could be represented in a three dimensional space in our cognition. The first dimension is the degree of understanding the other's emotion, the second is the degree of sharing that emotion and the third is the degree that the observer can differentiate himself from the person they observe. He continues that the reason that empathy is hard to distinguish and some researchers project it as a greater category, that includes other concepts, is its place on the three dimensions. Depending on the dimension someone is focusing their research, empathy is very similar to other concepts, making it hard to disambiguate the terms with a strict definition.

"Feeling as the other" and "feeling for the other". The distinction between *"feeling as the other"* and *"feeling for the other"*, mentioned above as the degree of sharing the same emotion, is what separates empathy and sympathy for Hein and Singer (2008). For him and other authors (Cohen and Cohen, 1992) (Decety and Lamm, 2006) sharing the same feeling/emotion with the other person is what constitutes empathy, while in sympathy different emotions are produced while observing the other person. Following this distinction, the empathetic ability of someone can be assessed by how much their emotion differentiates from the observed person, which is called empathetic accuracy(Ickes, 1997). A different opinion in the literature is that empathy is not restricted and the observer could be empathetic with a different emotion (Preston, 2007) (Eisenberg et al., 2006), which could be similar or not. Various different definitions, that do not perceive empathy as a category, disagree on this aspect.

"Self-other merging" and "self-other distinction". The third dimension of the complex psychological concepts is the degree that the observer can differentiate himself from the person they observe. When someone can not differentiate themselves is called *"self-other merging"* and its opposite is called *"self-other distinction"* (Batson, 1987). Wispe (1986) instigated that the self-awareness, meaning that the observer understands that these feelings originate by other's experience and acts differently as if the same emotions occurred to them by their experience, is an important factor to distinguish empathy. Batson (1987) with the findings of their research supported the idea, that people can differentiate the emotions produced by observing others suffering and that is how he distinguished EC and PD as the two big categories. Technological advantages gave researchers the ability to investigate further by peeking on the activity of the brain. Evidence produced by experiments using fMRI data of the brain (Singer and Lamm, 2009) (Jackson et al., 2006) suggested that the distinction between empathy and *"self-other merging"* is not as clear as it was accepted in the past. The *"self-other merging"* exist in our cognition and has an important role in empathy as it is trivial in the process of understanding the emotion that others feel.

Cognitive or affective human process. Moreover, an important factor of disagreement between authors is if empathy is a cognitive or an affective human process (Gerdes et al., 2010). The ability to process and understand the feelings/emotions of the observed is called cognitive empathy (Ickes et al., 1990). The emotions produced by observing the experience of someone else is the affective empathy (Cuff et al., 2014) (Mehrabian and Epstein, 1972). For some, these are two different concepts of empathy, making some focus on one of them, excluding the other on their definition

of empathy. The brain activity evoked by them is associated with different areas in the brain as neurological evidence indicate (Shamay-Tsoori et al., 2009), but these concepts seem to interact so much and affect each other's activation that their separation is not suggested and it is now accepted that together they constitute empathy (Barker, 2008).

2.3 Empathy detection

In text-based empathy detection systems, textual features are used to predict if a text is empathetic or how empathetic it is on a scale. In this section, for each of the text-based empathy detection models I will discuss the methodology used, the way that the researchers define empathy, the process of gathering the data and how they judged if an utterance is empathetic or not. In Table 2.1 the papers showcased in this section are grouped according to the empathy definition they followed. Further, in subsection 2.3.1 I will provide a brief discussion, examples and statistics of the available datasets for empathy detection on text.

Patient — therapist conversations

First work on text-based empathy detection. The early studies on empathy detection was centered around therapy sessions. The following study is the first work that tried to identify textual features that signal empathy. Suchman et al. (1997) performed a study on emotion and empathy detection based on discussions between patients and physicians. Each member of the research team examined a a transcription of a conversation between the patient and the physician, identified the emotion of the patient based on their utterance and examined if the doctor's utterance was empathetic or not. This analysis led to the creation of a conversational guide that therapists could follow and detect when the patient seeks an empathetic response, with the goal of having more efficient dialogues. This work did not have as a goal to produce a machine learning system, but they still detected empathy using signals from text, a certain phrase or the way of expression for example. Thus, indicating that only the text of a conversation could provide the necessary information to detect empathy in therapy sessions. [h!]

Empathy definition	Papers
Showing an active interest by the therapist to understand what the client is feeling	(Xiao et al., 2012), (Xiao et al., 2015), (Gibson et al., 2015), (Perez-Rosas et al., 2017)
Batson's empathy scales: Empathetic Concern – Personal Distress	(Litvak et al., 2016) (Abdul-Mageed et al., 2017) (Buechel et al., 2018) (Sedoc et al., 2019)
Feeling the same emotion as the person observed	(Khanpour et al., 2017) (Zhou and Jurgens, 2020)
The emotional state occurring by observing someone else going through an emotional reaction and feeling the same emotion or similar one	(Alam et al., 2018)
EPITOME: three communication mechanisms of empathy – Emotional Reactions, Interpretations, and Explorations	(Sharma et al., 2020)

The empathy definitions in the literature and the papers they follow them

First machine learning systems for empathy on text. The following papers all aimed at producing a machine learning system that identifies empathy. In the first paper (Xiao et al., 2012) the researchers worked on two binary classification tasks (empathetic or not), one on utterance level and one on the whole conversation of clinical trials studies on substance use by college students, thus two datasets were created. They defined empathy as "showing an active interest by the therapist to understand what the client is feeling", differentiating it from warmth, sympathy and other concepts. For the first task they used the manual for the Motivational Interviewing Skill Code (MISC) (Miller and Moyers, 2008) to annotate the utterances. For the second task they annotated a dataset using the Motivational Interviewing Treatment Integrity coding manual (MITI) (Moyers et al., 2014). For the machine learning model of the second task, they used language features that were extracted from the bigram model which was the most succesful on the first task. Continuing the research the same team created the first automatic system for classification of a psychological session (Xiao et al., 2015). In this paper, they created a pipeline system, in which an Automatic Speech Recognition (ASR) module transcripts the language and a binary empathetic classifier rates the therapist as empathetic or not. For this purpose, they created a new dataset using MITI and they used n-gram features with the support vector machine algorithm for their model. At the end they constructed a system that is a concrete application of empathy detection.

Examination of different features. The next two papers investigated which type of features contribute more in classifying empathy. Both studies use counseling data and were annotated with the usage of MITI, as they use the same empathy theory. First, Gibson (Gibson et al., 2015) examined the different types of linguistic features. These are different types of n-gram models (unigram, bigram and trigram), the psychological dimensions of the Linguistic Inquiry and Word Count tool - LIWC features (Pennebaker et al., 1999), which include constructed count features based on different emotion or syntactic aspects on the whole text, and similar features to LIWC which were constructed by them. After performing a correlation study be-

tween the created models they found that the results of the n-gram features did not correlate with the results of the other constructs, hinting that the different types of features capture different information. In the second work, they constructed abstract behavioral features using acoustic and linguistic features and investigated which offers more information (Perez-Rosas et al., 2017). They concluded that therapists who show engagement (consistent interactions, allowing patients to speak), coordination (match the communication style of patient) and matching content (reflective language) are considered more empathetic. These two studies showcase that the accurate detection of complex concepts, such as empathy, require information that these features can not provide on their own.

Social media

Facebook activity. Litvak et al. (2016) and Abdul-Mageed et al. (2017) focus on user's Facebook data, including their posts and information about them. For both studies, user's gave their permission to have their data crawled and these were annotated using Davis' IRI (Davis, 1980a). Davis' IRI relates closely with the empathy definition of Batson (1987) and characterizes empathy as a class that includes empathetic concern, fantasy sub-scale, perspective taking and personal distress (Davis, 1980b). Litvak focused on creating a system that uses features, which originate from the user's activity on the last thirty months, to detect the empathy scores of the user. These features include the LIWC on their text history and constructed Facebook activity features. While their results indicated that there is correlation between the different writing style and empathy, the lack of a big sample size did not allow them to make robust conclusions. n the second paper, the emphasis was on detecting the personal distress score, the pathogenic empathy as they call it. For their regression task, of predicting the pathogenic empathy score using a user's post and their information, the models created used different mixture of features, from a pool of n-gram, gender and race information, topics of posts and word embeddings (word2vec). Their most succesful model was a mixed model with unigrams and gender information of the user.

Reaction to news stories. Buechel et al. (2018) took different approaches compared to other researchers on the text-based empathy detection. Initially, they emphasized on following the theory of Batson (1987), after criticizing the work so far due to the usage of shallow definitions of empathy not backed by psychology. For their dataset, they gathered the reaction comment of people after reading a news article and their EC and PD scores at that moment, which were extracted after they completed Davis' IRI questionnaire. They claim that this dataset is the first "gold-standard" dataset on text-based empathy detection because it captures the emotions of the people experiencing and not what others perceive. They formulated the detection of EC and PD as regression tasks and their most succesful model was a Convolution Network on top of pre-trained FastText embeddings. In addition, the same group continued their work on empathy by producing the first ever lexica with words judged on their EC and PD levels (Sedoc et al., 2019). For this purpose, they used the dataset of their previous work and their outcome was the first empathetic Lexica with 9356 word types (lower-cased, non-lemmatized, including named enti-

ties and spelling errors), ready to be used for prediction tasks or for enrichment of word embeddings. They concluded though that with different and more complex methods the Lexica could have a better quality. This group was the first to publish their dataset and the Lexica that they produced.

Asynchronous conversations focused on mental health. The following studies use responses/comments from asynchronous conversations on forums, discussion boards and social media platforms dedicated to mental health. Asynchronous are the conversations that have a time delay. This work was the first one that took a computational focus on empathy (Khanpour et al., 2017). They presented a machine learning model, for their binary classification task, that identified empathetic messages in discussion boards in online health communities. After collecting the data, they annotated them based on the empathy theory presented by Decety and Jackson (2004), on utterance level. In their proposed model, the input gets transformed into word2vector word embeddings, then it passes through a convolution layer and then through an LSTM layer before the softmax layer outputs the predicted class. Using the same definition for empathy detection, Hosseini and Caragea (2021) took a different approach. They gathered conversations from the discussion board of an online cancer network and annotated the sentences between three classes, none, "seeking" empathy and "offering" empathy to construct their dataset, named IEmpathize. It is the first work that also annotates when someone is seeking empathy. In their computational approach the used the pre-trained BERT model and fine-tuned it for this task. In both studies, the models were used to identify empathetic messages in conversations and observe the emotional change on the users that received such messages, showcasing the power of identifying empathy.

Sharma et al. (2020) took a computational approach into the text-based empathy detection on asynchronous dialogues on mental health support. The group criticized other researches that used questionnaires and empathy scales that were designed for spoken conversations (Davis' IRI and MITI). Instead, they defined a text-based framework for empathy called EPITOME, which separates empathy into three types:1) the emotional reaction - similar to empathetic concern that we have seen before (Batson, 1987),2) the interpretation - an utterance that shows an understanding to the problem and 3) the exploration - an utterance that shows a will to learn more details about the situation. Having defined empathy, they proceeded to create a dataset with asynchronous dialogues around mental health, which include a post from a user describing their emotions and a response from another user. The annotators had to identify for each aspect of EPITOME three levels, from none to strong, on the whole comment and also annotate the sentence of the response that includes that empathy type, the rationales. It is the first work on empathy that not only tries to detect different empathy types but also showcase the part of the text responsible for that classification result. The three multitask models that they created, one for each aspect of EPITOME, use a bi-encoder architecture with attention, based on RoBERTa (Liu et al., 2019). Overall, the group created a new framework for empathy detection on text, a dataset based on that framework with rationales highlighted, a powerful model that performed well on both tasks for every aspect and made them public.

Empathy in condolence messages. In the following paper, empathy was not their main focus, but a part of a bigger study from Zhou and Jurgens (2020) centered around condolence and distress in online communities, mainly on Reddit. For empathy, their goal was to detect which condolence messages contained empathy. Their empathy definition is strict and limits the concept to the cases when the observer has the same emotion as the observed person. Their dataset of contained distress-condolence pairs, which were annotated on a five-scale system to gather the regression score of the condolence reply. Two models were trained on the data, a random forest with n-gram features and a RoBERTa one (Liu et al., 2019), with the second one outperforming the first. Next, they used the model of Buechel (Buechel et al., 2018) to predict empathy scores on their dataset to observe how the different empathy theories relate. The Pearson's correlation result was 0.343 showing a positive correlation but it is obvious that the two models study different concepts, which is a logical result as the definition of this work is strict and the empathy studied by Buechel incorporates the terms of sympathy, compassion and more. Additionally, the two studies use different methods of annotating. This result highlights the importance of the definition and how careful one must be to compare different works.

Synchronous conversation on social media. Rashkin et al. (2020) focused their work on creating a system that produces empathetic responses while conversing, to promote empathetic messages on conversational agents. The dataset created includes 25 thousand dialogues between two people, where the initiator states their emotion and the other person tries to communicate empathetically to help the initiator. In this work they did not focus on detecting empathy, so they did not assess how empathetic the messages are and were not clear on what definition of empathy they used. Recently, Welivita and Pu (2020) having as a goal to advance the empathetic response generation systems used Rashkin's dataset and added level of details on the utterances. Welivita classified the empathetic utterances on 8 intention classes, but they did not use a definition of empathy or a scale-system on how empathetic is the message. By taking a sample of the original dataset, they annotated the utterances of the person responding from a list of 8 intentions or as neutral and then created a classifier using the BERT architecture to group the response utterances of the whole dataset. Finally, Welivita used a RoBERTa network and fined tuned it to classify an conversational utterance between 41 classes (33 emotion, 8 intentions).

Spoken synchronous conversations

Call sender. Alam et al. (2018) focused on the binary classification task of empathy using both acoustic and linguistic features. The data used in this study, were spoken conversations in Italian between a health call center operator and the people who called. For them empathy is defied as "the emotional state occurring by observing someone else going through an emotional reaction and feeling the same emotion or similar one", a definition related to Batson's empathetic concern (Batson, 1987). They used this definition to annotate utterances of the spoken conversations only on the operator level and they did not examine the utterances of the caller. Their goal was to create a whole system that runs during the duration of the call and

annotates the empathy of the operator in real time, with the intention of increasing the quality of the call. For its implementation they used three classifies, one trained on audio signals, one trained on n-gram features on text acquired by a ASR and one trained on LIWC features on the same text, and then through majority ruling they system decides if the utterance is empathetic or not.

2.3.1 Datasets

Reaction to news stories Buechel et al. (2018) were the first to make a resource about empathy detection on text public. In this work, they used the definition of empathy of Batson (1987) with the two scores of Empathetic Concern (EC) and Personal Distress (PD), to evaluate the comments of the users. The first category covers the other-oriented emotions evoked (sympathy, compassion, etc) and the second the self-oriented negative emotional responses. During the creation process, people first read a news story, then evaluated their EC and PD scores, using two multi-scale questionnaires and then commented. The creators decided to gather the scores using multi-scale questionnaires to counter the different perception of the question that each individual could have, as it is a self-evaluation. For this reason and because they did not use a third-party to evaluate people's comments, they consider their dataset as a "gold-standard" one. The set of articles (418 in total) that people read were selected by two researchers (psychology undergraduates) that were seeking articles that could evoke emotional reactions to the readers. After the post processing phase, the total amount of responses was 1860 with length varying between 300 to 800 characters and the median number of tokens per response was 84. By performing Pearson correlation on the two variables they observed a moderate correlation 0.451, supporting the research of Batson that claims that these two concepts co-exist. In Table 2.2, some examples of responses with their EC and PD scores are presented. Based on this work, Sedoc (Sedoc et al., 2019) constructed the first public lexica for empathy with the development of a machine learning system named Mixed-Level Feed-Forward Network (MLFFN), as self-annotation for such complex concepts has not proven succesful. The lexica contains 9356 lower-cased words with rating for EC and PD.

EC	PD	response
4	5.5	Here's an article about crazed person who murdered two unfortunate women overseas. Life is crazy. I can't imagine what the families are going through. Having to go to or being forced into sex work is bad enough, but for it to end like this is just sad. It feels like there's no place safe in this world to be a woman sometimes.
1	1.3	I just read an article about some chowder-head who used a hammer and a pick ax to destroy Donald Trump's star on the Hollywood walk offame. Wow, what a great protest. You sure showed him. Good job. Lol, can you believe this garbage? Who has such a hollow and pathetic life that they don't have anything better to do with their time than commit petty vandalism because they dislike some politician? What a dingus.

Table 2.1: Examples of Reaction to news stories dataset

Epitome dataset Sharma et al. (2020) after the creation of their Epitome framework, used it to annotate a new dataset. The data sources for the asynchronous conversations were the network TalkLife, which allows conversation threads and communities of Reddit (subreddits) focused on mental health. They gathered 10,143 pairs of user post and peer response. Each sentence was annotated for each of the three aspects of the Epitome framework. The classes for each aspect were none or weak or strong. The process of annotation was succesful with an inter-annotator agreement was 0.6865. Their intention with the inclusion of the rationales in the dataset was the advancement of our understanding on how the system classified the utterance, which could provide insides to the people using the system to improve their empathetic skills. The Table 2.3 includes two responses from the EPITOME dataset.

ER	I	E	Response
Strong	Strong	None	**If that happened to me, I would feel really isolated.** Let me know if you want to talk. **I really hope things would improve.**
None	None	Strong	**wonder if this makes you feel isolated.** Let me know if you want to talk.

Table 2.2: Examples of responses on EPITOME dataset

A Taxonomy of Empathetic Response Intents in Human Social Conversations The two datasets we have seen so far have restricted topics, as the first one captures the reactions on news articles and the second one captures conversation from mental health discussion groups. Rashkin et al. (2020) and Welivita and Pu (2020), who annotated the dataset released by Rashkin, published an open-domain conversation dataset. Rashkin gathered 24,856 empathetic dialogues grounded on the emotion of the person initiating the dialogue, with an average length of 4 turns. There are 33 different emotions of ground truth in the conversations. Welivita decided to annotate the responders' texts between 8 empathy intention classes (Questioning, Acknowledging, Consoling, Agreeing, Encouraging, Sympathising, Suggesting, Wishing) from 500 conversations. Then they used a BERT Transformer-based classifier to annotate the rest of the utterances in the dataset with either an emotion class out of the 33 or an intention class out of the 8. The group was inspired to annotate on the dataset, as a generated response could be classified with a different intention if the initiator is happy or if he is sad, hoping to advance empathy text generation. The annotation took place on the whole response and not on sentence level, but the responses on this dataset tend to be short as they include conversations. An example of a conversation is presented in Table 2.4.

IEmpathize Hosseini and Caragea (2021) constructed the dataset, named IEmpathize by gathering asynchronous conversations from the discussion board of an online cancer network. In more details, they annotated the sentences between three classes. The classes are "offering" empathy, "seeking" empathy and none. For "offering" empathy the definition of Decety and Jackson (2004) was used, stating that empathy is "the psychological recognition and understanding of the others' feelings,

Class	Utterance
Disgusted	S: Bleh, I just had the worst food ever.
Questioning	L: What did you eat?
Disgusted	S: I was at Mcdonalds and was given a rotten cheese burger. I almost puked after I ate it.
Disgusted	L: Oh gross, makes me never want McDonalds again.

Table 2.3: A conversation from the empathetic taxonomy

thoughts, or attitudes". "Seeking" empathy is described as "asking to be truly understood, which is why people seek each other out in hard times". In total 5,007 sentences were annotated, 1,046 are "Seeking" empathy, 966 are "offering" and the rest are none. The inter-annotator agreement on the dataset was 84% measured by Cohen's kappa coefficient. In Table 2.5, some sentences with their class are presented.

Class	Sentence
Seek	I cannot imagine living the rest of my life this way, I am sick to my stomach every day.
Offer	I know you feel really down about this, but look at me I'm still here and have a reasonably good quality of life.
None	I used Aquafor skin lotion/gel (over the counter) for radiation side effects on my skin.

Table 2.4: Examples of the IEmpathize dataset

Condolence and empathy in online health communities. The final resource by Zhou and Jurgens (2020) is the first one to have a narrow definition of empathy. Empathy here is defined as the situation where the observer has the same feeling as the person that they observe. In this dataset they gathered 1000 distress posts with the condolence reply of another user from Reddit. The annotators ranked the condolence replies on a five-point Likert scale to extract the empathy score per text. The dimensions of the scale are: 1) pleasantness, 2) anticipated effort in dealing with the situation, 3) situational control, 4) how much oneself or another person was responsible for the situation and 5) attentional activity and certainty about what was happening in the situation. Following all the previous works, the annotation has happened on the whole response and not on sentence level. The Pearson correlation between the two annotators was at 0.58 and was judged as moderate by the researchers.

2.4 Natural language processing

In this section, I will discuss the NLP methods that I will use in my thesis. First, I will present the Transformer architecture and some of the Transformer-based pre-trained models. Then, I will explore different techniques of data augmentation,

which is used to augment the size of the training sample.

Transformer architecture

Attention is all you need In 2017 the Transformer architecture was introduced (Vaswani et al., 2017) and replaced as the state of the art the recurrent neural networks architectures, such as long short-term memory (LSTM), in the NLP tasks where sequence matters. Why is this the case? The Transformer architecture is composed of two parts, the encoder and the decoder. The encoder and the decoder can be seen at figure 2.1, with the left block being the encoder and the right being the decoder. The role of the encoder is to encode the input sequence in a state and this state is then passed to the decoder. In both of these parts only the mechanisms of multi-head attention and forward pass are being used. In more details, the input tokens first are converted to a sequence of word-embedding vectors and then positional information is added on these vectors. After that, multi-head attention is applied on the whole sequence, which performs matrix multiplication, and then the output passes through the feed forward step. In contrast with the recurrent architectures, this architecture does not require the sequential processing of the data, allowing parallelization, which reduces the training time. Additionally, the Transformer architecture achieved better results to the various NLP benchmarks, like machine translation, and does not suffer from the other disadvantages of the recurrent architectures, such as the vanishing gradient problem.

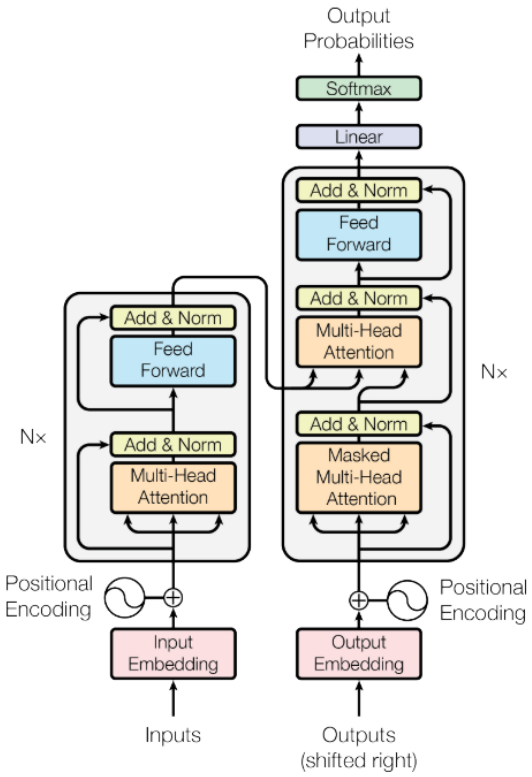

Figure 2.1: The Transformer architecture (Vaswani et al., 2017)

The ability of parallelization enabled researchers to train models using unsupervised tasks on enormous datasets, such as the Wikipedia's corpus. With the use of transfer learning, these pre-trained models could be used by other researchers, after a fine-tuning process on their specific task. This characteristic of the Transformers allows the usage of very powerful models to researchers who did not have this option before. Examples of pre-trained Transformers are BERT and RoBERTa.

BERT. BERT (Devlin et al., 2019), whose initials stand for Bidirectional Encoder Representation from Transformers, is a pre-trained Transformer model designed by Google. In the pre-training phase, a vast amount of unlabeled data were used, 16 GB from Wikipedia and BooksCorpus, on the unsupervised tasks of Masked Language Model (MLM) and Next sentence Prediction (NSP). The MLM task is the reason that BERT learns bidirectional, which was not possible before. In this task, a percentage of the tokens of a sentence is masked and the learning objective is to identify them. In the task of NSP, BERT tries to predict if two sentences are pairs (the second follows the first) or the one does not follow the other, making the model able to learn sentence relationships. After the pre-training phase BERT can be fine-tuned in a supervised task using a smaller amount of data than it was required before, due to the pre-training phase. When BERT was published in 2018, it created the state of the art models, with the use of fine-tuning, on 11 NLP tasks, which belong to text classification, question-answering and textual entailment, with large margins. On the tasks SQUAD (Rajpurkar et al., 2016) and SWAG (Zellers et al., 2018), BERT was able to surpass human-level performance.

RoBERTa. The initials RoBERTa stand for robustly optimized BERT pre-training approach and they indicate that the developers had an alternate approach on pre-training BERT model (Liu et al., 2019). The researchers from Facebook, by studying BERT, found that different aspects of it could be changed to improve the final model. Initially, they adapted the unsupervised tasks for the pre-training process, they decided to use Dynamic masking instead of the MLM task and not use the NSP task at all. During the training phase of BERT, masking was performed only once which means same input masks are fed to the model on every epoch. In Dynamic masking, for each epoch there are different versions of the sentences with masks in different tokens. Moreover, they increased the training data size from 16 GB to 161 GB, with many more sources of data, and decided to train RoBERTa with longer sequences compared to BERT. Their final model managed to outperform BERT and other pre-trained models on GLUE (Wang et al., 2018), but it requires more computational power to be fine-tuned and predict compared to BERT.

Data augmentation

In any possible task in natural language processing, the performance of it is impacted by the amount and quality of data available for training. If the amount of data is problematic, the solution is to either gather more data or construct them, which is called data augmentation. Data augmentation includes techniques that can be used to either use a training sample by modifying it slightly to acquire a new instance or by processing that sample and synthesize a new one from it. Below

follows the review of three techniques: back-translation, easy data augmentation — EDA and contextual augmentation — CLARE.

Back-translation. The idea of back-translation is simple to grasp. Paraphrase the available samples for training by translating the original text back and forth (Sennrich et al., 2016). The benefit that it brings compare to other methods is that the returned sentences have more syntactic diversity, as their structure could be different than the original. Yu et al. (2018) successfully introduced and used this technique, on their question-answering task, by translating their samples from English to French and then to English again.

Easy data augmentation. EDA includes four traditional and simple methods to create slightly different samples than the originals (Wei and Zou, 2020b). These techniques are: synonym replacement, random insertion, random swap and random deletion. All these techniques do not consider the stop words of the text. In synonym replacement n number of words are randomly selected from the sample and replaced with their synonym to produce a new sample. In random insertion a synonym of n words from the sentence are placed in random places in the sentence to produce the new one. In random swap words are selected randomly and exchange position. In random deletion each word has a probability to be deleted from the sentence to produce the new one. These techniques might be simple but they have showed that they improve the performance and generalization of the models.

Contextual Augmentation. Kobayashi (2018b) introduced the contextualized augmentation. In this method words from a sentence are selected randomly for replacement, similarly to EDA, but their replacement mechanism is more complicated compared to synonym replacement. They have created a bi-directional LSTM-RNN language model that is able to find the most replacement word based on the context. Their experiments showed a constant improvement on top of synonym replacement methods, but this improvement was marginal in some tasks.

Chapter 3

Automatic empathy detection

This chapter is dedicated on the first research question. With this research question I ask "What is the performance of the Transformer architecture on predicting the empathetic concern and personal distress scores of reaction comments on news articles?". To answer this question I used the "Reaction to news stories" dataset created by Buechel et al. (2018), which I presented in Subsection 2.3.1. During their work, they also used this dataset to produce models, which I treated as baselines.

In section 3.1, I performed analysis on the dataset to acquire insights about the data. Section 3.2 presents the methodology of the experiments in parallel with the results. It is divided into Subsections between the experiments with the original dataset and the experiments using data augmentation methods. In Section 3.3, I discuss the findings of the chapter and answer the research question.

3.1 Data analysis

Dataset Before starting with the analysis on the dataset, I will summarize some of its details. The data include people's comments and their empathetic concern (EC) and personal distress (PD) scores. People first read a news article, then self-evaluated their EC and PD and then commented on the article. The participants were instructed to write a comment as they would do on a social media website and to use between 300 and 800 characters. In Table 3.1 two instances of the dataset are presented. These examples showcase that the language in the comments can differ a lot. By looking specifically on the second example, constructs such as sarcasm and acronyms exist in this dataset. The existence of sarcasm raises the quality of the data, as it resembles real-life comments, but at the same time it raises the difficulty of the NLP task.

Number of instances Starting with analysis of the dataset I will present some details of it and then examine the distributions of the two scores. In total the corpus includes 1860 comments with their EC and PD scores. Before releasing the dataset its creators manually inspected every comment, removing comments that did not follow the described task.

Number of tokens The participants of this dataset were instructed to have a length between 300 and 800 characters on their comments, leading to sentences

EC	PD	response
4	5.5	Here's an article about crazed person who murdered two unfortunate women overseas. Life is crazy. I can't imagine what the families are going through. Having to go to or being forced into sex work is bad enough, but for it to end like this is just sad. It feels like there's no place safe in this world to be a woman sometimes.
1	1.3	I just read an article about some chowder-head who used a hammer and a pick ax to destroy Donald Trump's star on the Hollywood walk of fame. Wow, what a great protest. You sure showed him. Good job. Lol, can you believe this garbage? Who has such a hollow and pathetic life that they don't have anything better to do with their time than commit petty vandalism because they dislike some politician? What a dingus.

Table 3.1: Example sentences of the "reaction to news stories" dataset. Parts of second sentence is colored, for the identification of sarcasm (red) and of acronyms (green).

with varying length. The median number of tokens (words) per sentence is 84, with the shortest sentence having 52 tokens and the longest having 198. In total the number of tokens are 173686.

EC and PD distribution Next, I explored the distribution of the EC and PD scores, as these two concepts should be correlated according to the empathy theory of Batson (1987). The two histograms in Figure 3.1 showcase these two distributions of the scores. Both scores follow a similar, but not identical, distribution. In more details, both have population peaks around the scores of"1" and "4", but these peaks are slightly more populated on empathy. The difference between the two distributions are that for PD the population is more spread across all scores. Additionally, the two distributions score a r=0.451 on Pearson's R correlation (Table 3.3) meaning that these two distributions have a moderate correlation. This result support the claim of Batson, that these two phenomena are distinct but relate closely and could co-exist in most cases (Batson, 1987). Below follows the Table 3.2 with the mean, standard deviation and some additional statistics about each distribution. These statistics validate how similar these two distributions are, as their means, standard deviations and the percentiles have similar values.

Sentences with a difference in their EC and PD scores Since this correlation between EC and PD is substantial, it would be interesting to see examples where the two scores differ a lot. In this dataset, out of the 1860 instances there are 44 with high distress (equal and over 6) and low empathy (equal and lower than 2), and 39 with the opposite. In Table 3.4 I present two example sentences for this phenomena. Looking at the first example, the participant doesn't show any emotion related to empathy, as the language used does not have sign of warmth or sympathy. However, they seem very distressed about the article they have read. By looking at the second

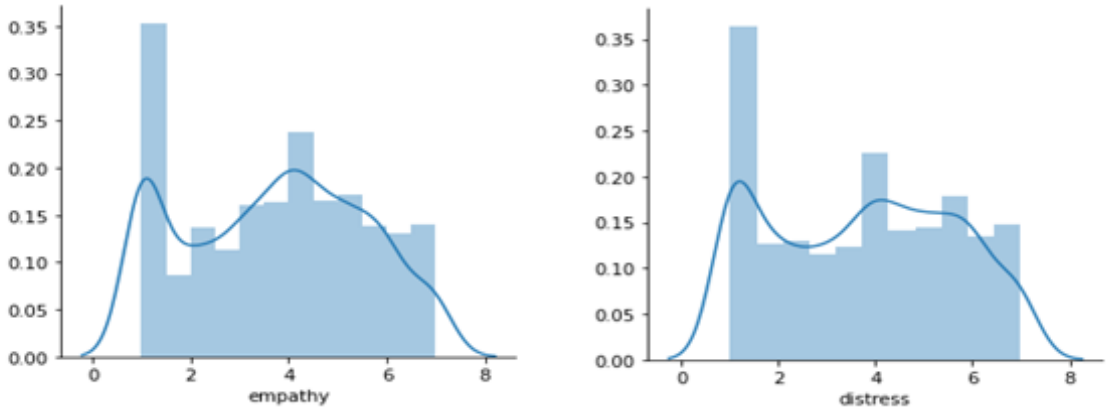

Figure 3.1: Histogram of EC and PD

measure	EC	PD
count	1860.0	1860.0
mean	3.7	3.73
std	1.82	1.9
min	1.0	1.0
25%	2.0	2.0
50%	3.83	3.88
75%	5.17	5.38
max	7.0	7.0

Table 3.2: Statistics about EC and PD

example, the language does not seem to favour either score as the participant doesn't really express their feelings, but the self-evaluation scores are different from what we might expect. This is an example that showcases the difficulty of the task. People evaluated themselves and then commented, but as each person has a different way of expressing their feelings, there can be sentences that do not align with the common perception of empathy and distress, with that sentence being a good example.

	EC	PD	length
EC	1.0	0.451	0.136
PD	0.451	1.0	0.153
length	0.136	0.153	1.0

Table 3.3: Correlation table of EC, PD and length

Correlation with length Last, I want to explore the relation of each score with the length of the comments. Considering that the creators of the dataset allowed sentences with varying length, there might be a linkage between the levels of empathy or distress with the length of the comments. The correlations were studied with Pearson's R correlation. As we can observe in Table 3.3 there is a small positive correlation between the length with each score, r=0.136 with empathy and r=0.153

EC	PD	response
1	7	We need to destroy the system that allows these things to happen. Capitalism that reinforces these disgusting habits must be destroyed utterly and the Christian religion should be wiped out in its entirety. There is no excuse for this kind of repulsive behavior that drives people to abuse these children.
7	1	I recently read an article about how the band Eagles of Death Metal visited the Paris attacks memorial. A year before they were playing at a concert hall when the attacks occured a year prior. Sting was the first person to play at he concert hall since the attacks. The front man of the Eagles of Death Metal was denied access to the event. He had told Fox news that the muslim security collaborated with the attackers at the concert hall.

Table 3.4: Sentences that had a big difference between their EC and PD scores

with distress. These slight correlations show a linkage between longer texts and increased levels of self-reported empathy and distress.

3.2 Experimental phase and results

The "reaction to news stories" dataset includes the empathetic concern (EC) and personal distress (PD) scores of the people who first read and then commented on the articles, after evaluating themselves. This means that the research question includes two regression tasks for predicting separately the EC and PD scores. For these regression tasks, I examined the performance of two types of pre-trained Transformer models, BERT and RoBERTa base models, against the baseline models created by Buechel. For this process, I fine-tuned the pre-trained models on the dataset. Before comparing these models with the baselines, I first performed a hyper-parameter selection on each one to find the optimal settings. After experimenting with this data, I explored data augmentation methods on the original sentences and then used the Transformer models, with the selected hyper-parameters. This choice was based on the small number of utterances in the data. During this section, I present the methodology steps along with their results. Before beginning with the experimental phase, I first present the baseline models, a summary of the methodology that I followed and some training details.

Baseline models The the models that I treat as baselines for these tasks are the ones created by Buechel et al. (2018) on the data. The baseline models vary in complexity. These are a Ridge regression model, a Feed-forward neural net (FNN) model and a convolution neural net (CNN) model. The FNN model has two hidden layers of 256 and 128 units respectively. The CNN model includes a single convolution layer which is followed by a single dense layer with 128 units. For all baselines, they used the Fast-Text word-embeddings as inputs on the models. The performance of these models are presented on the following sections.

Methodology The methodology that I follow to answer this research question can be divided on two sections, one with the original data (reaction to news stories) and one with the augmented data. On the original data section, I first performed 5-fold Cross-Validation (CV) using all data except a test set for hyper-parameter selection for BERT and RoBERTa. After I selected the best versions of these models, I tested them with two methods. Initially, I performed 10-fold CV on the whole dataset similar to what Buechel did, so I can compare these two models with the results of Buechel. The other method was to test them using the test set that I have separated beforehand, to get insight on the performance of these models.

On the data augmentation section, I did not perform hyper-parameter selection and I used the already selected hyper-parameters. Using the different methods for data augmentation, I performed 10-fold CV on both models for each method to compare them with the baselines and then at the end I used the original test set to evaluate the models further.

Training details For the training of the BERT-based and RoBERTa-based models, I used the NLP library from Hugging face (https://huggingface.co/), in which there are implementations of these Transformer architectures. I chose to use the base edition of both models, which include 110 million and 125 million parameters respectively. The optimal choice for the models are the large editions, but due to their parameter size and the length of the sentences the memory of the GPU I am using (NVIDIA QUADRO RTX 6000) was not enough, even with the batch-size of 2 for the training phase, an option that is not recommended.

3.2.1 Original dataset

Hyper-parameter selection methodology The first step of these experiments was the separation of a test set for validating the results at the end. 10% out of the 1860 comments (186) were placed on the test. The rest of the data were used for 5-fold Cross-Validation (CV) to identify the optimal hyper-parameters for both Transformer models. After I select for Both BERT and RoBERTa the best performing hyper-parameters, I will compare them with the baseline models. Important to note, that the optimal hyper-parameters were selected by the average Pearson's R correlation score of the two models, because the final models will have identical settings. The hyper-parameters that were examined, during the 5-fold CV phase, are presented in Table 3.5. The batch size is limited to 8 and 16 for memory purposes, with 16 being suggested by the creators of the models (Devlin et al., 2019). Important to note, that initially the distilled versions (Sanh et al., 2019) of both pre-trained models were used to narrow the hyper-parameters to the ones presented in Table 3.5, with the assumption that the base versions will perform best in that space. The distilled models are smaller in size (almost half the number of parameters) but have a small drop of performance (5%) compared to the base models.

Hyper-parameter selection on BERT First, I present the best performing models of BERT in both regression of empathy and distress in Table 3.6. The criteria to select the best performing models out of the fifty that were tried, with all combinations of the hyper-parameters, is the average Pearson's R correlation

hyper-parameters	options
batch size	8, 16
learning rate	1e-05, 2e-05, 3e-05, 4e-05, 5e-05
weight decay	0, 0.1, 0.2, 0.3, 0.4

Table 3.5: Hyper-parameter space for both BERT and RoBERTa models.

score on the validation set of the 5-fold CV. As we can observe in the Table, for all hyper-parameter combinations the best performing model is the the one on the regression of distress. The majority of the selected models have a batch size of 16, which supports the suggestion of the creators of BERT that 16 or 32 are usually the best option for batch size. As it seems, the highest learning rates perform best and the weight decay hyper-parameter varies. The first model in Table 3.7, i.e. the one that achieves an average correlation of 0.44, is selected for the next phase. This model achieved the highest correlation on empathy and the third highest on distress. The hyper-parameters of it are a batch size of 16, learning rate of 0.00005 and weight decay of 0.2.

Batch size	learning rate	weight decay	EC corr.	PD corr.	average
16	5e-05	0.2	**0.426**	0.453	**0.44**
16	5e-05	0.1	0.42	**0.458**	0.439
16	5e-05	0	0.42	0.455	0.437
16	5e-05	0.3	0.417	0.45	0.434
16	3e-05	0.3	0.424	0.443	0.433

Table 3.6: Results of 5-fold CV BERT for hyper-parameter selection on empathy and distress using 90% of the dataset. The highlighted row is the one with the selected model.

Batch size	learning rate	weight decay	EC corr.	PD corr.	average
8	2e-05	0.3	0.431	**0.468**	**0.45**
8	2e-05	0.1	**0.438**	0.457	0.447
8	3e-05	0.1	0.432	0.461	0.447
16	4e-05	0.2	0.431	0.459	0.445
8	2e-05	0.2	0.427	0.46	0.443

Table 3.7: Results of 5-fold CV RoBERTa for hyper-parameter selection on empathy and distress using 90% of the dataset. The highlighted row is the one with the selected model.

Hyper-parameter selection on RoBERTa Next, I present the results of the RoBERTa models on the same task in Table 3.7. The results indicate a better performance of the RoBERTa models in comparison with the BERT models, as all the showcased models of RoBERTa achieve a higher correlation score compared to the most successful BERT model. In similar fashion with BERT models, the scores

are higher in the regression task of Distress. In this model though, the batch size of 8 is performing better and the learning rate of 5e-05 is not included in any model, which was the most successful learning rate previously. The architecture of these pre-trained models might be similar, but the optimal hyper-parameters seem to be different. These results will be verified during the prediction on the test set. For the next steps, the model with the hyper-parameters of batch size 8, learning rate 0.00002 and weight decay 0.3 was selected. This model achieved the highest average with a score of 0.45, while having the highest score of all models on distress and the fifth highest on empathy.

Model comparison with 10-fold CV After selecting the best performing models for both BERT and RoBERTa, I wanted to compare them between each other and with the baselines models, the ones created by Buechel. et. al. For this purpose, I followed the exact methodology performed by them. I performed a 10-fold CV on the whole dataset with the selected hyper-parameters. The performance of the models is measured with Pearson's R correlation, which is the same metric used by Buechel, between the predicted values and the true ones. By mimicking the experimental setup of Buechel, I can confidently compare the results of BERT and RoBERTa.

Results of 10-fold Cross Validation on original data In Table 3.8, there are presented the results of the baseline model and the BERT and RoBERTa models, with hyper-parameters that I selected with the 5-fold CV. The results highlight an improvement close to 10% from both BERT and RoBERTa models compared to the best performing model of Buechel et. al. The biggest improvement is observed on the regression task of empathy, in which all models seem to struggle so far. No correlation score surpassed the threshold of 0.5 to be considered strong and in both tasks for both models the correlation is moderate. Contrary to the previous results of the 5-fold CV, this time the BERT model has a higher Pearson's R correlation score than RoBERTa by a close margin on each task. BERT achieved 0.4549 and 0.4773 and RoBERTa achieved 0.4489 and 0.4732, in these two tasks.

model	empathy	distress	average
Ridge	0.385	0.41	0.397
FNN	0.379	0.401	0.39
CNN	0.404	0.444	0.424
BERT	**0.4549**	**0.4773**	**0.466**
RoBERTa	0.4489	0.4732	0.461

Table 3.8: The results of the 10-fold CV using the whole dataset. The first three models (orange) are the ones created by Buechel and the next two (pink) are the best performing BERT and RoBERTa models.

Testing using the separated test set The next step of the experimental phase is use the test set (not used for hyper-parameter selection) and to further evaluate these two models. First, I trained both BERT and RoBERTa models with the

selected parameters and then I predicted on the test set. With the results of the testing, I produced plots that offer insight on their predictive ability.

Results of testing The test results for the two models on both tasks, using Pearson's r correlation, are placed in Table 3.9. Similar to the results of the 5-fold CV but against the results of the 10-fold CV, the RoBERTa model outperformed the BERT one by a close margin, with the average scores of 0.4812 to 0.4731. While RoBERTa outperformed BERT on the average score, BERT achieved a higher r score on the PD task. In more details, the BERT model had PD score of 0.5419 and an EC score of 0.4043 and the RoBERTa model had a PD score of 0.5372 and an EC score of 0.4252. Both models showed a significant increase of performance at the regression task of PD, while suffering a drop of performance on predicting EC, in comparison to the results of both the 5-fold CV and the 10-fold CV experiments.

model	B.s.	l.r.	w.d.	Emp. corr.	Dis. corr.	average
BERT	16	5e-05	0.1	0.4043	0.5419	0.4731
RoBERTa	8	2e-05	0.2	0.4252	0.5372	0.4812

Table 3.9: Pearson's r correlation Test results of the selected Bert and RoBERTa models on both EC and PD tasks. B.s: batch size, l.r: learning rate and w.d: weight decay.

Significance of difference After performing the comparison of the two models on the test set, difference on the scores of EC and PD on both models indicates that this models have the same capabilities. But it is necessary to verify if the prediction differences of these models are significant or not. For this reason I chose to perform a two-tailed t-test for paired samples [1]. The t-test is ideal for comparing samples of distributions that are matched on "pairs" to signal the significance of their differences. Table 3.10 presents the results of the two t-tests, one for each task. The p-value of the test for the EC task is higher than the threshold of 0.05, which means that these distributions are similar and that both models could perform similarly. On the other hand, the p-value of the t-test on the PD task is lower than 0.05, leading to the conclusion that the distributions are different. Even if the BERT model has a lower average Pearson's r correlation compared to RoBERTa, there is a significant difference on the PD task, in which the BERT model achieved a better score than the RoBERTa model. These facts in combination with the results of the 10-fold CV, provides confidence for favouring the BERT model.

Scatter plots between predicted and true scores With the intention of acquiring a better understanding of how the models performed I plotted, for each model on each task, the true scores of the test set against the predicted scores. These plots are presented in Figures 3.2 and 3.3. The scatter plots for both models on the task of empathy showcase a problematic situation, as both models tend predict values between two and five and are not able to predict values close to their high margin, which is seven. The highest values that these models predict is slightly

[1] https://en.wikipedia.org/wiki/Student%27s_t-test

type of comparison	statistic	p-value
EC	-0.450	0.654
PD	-3.117	0.002

Table 3.10: T-test results between the BERT and RoBERTa models on both tasks of EC and PD. The p-value on the EC task is above 0.05 and is insignificant. The p-value on the PD task is lower than 0.05, thus the BERT model has a significant difference compared to the RoBERta model.

higher than five. A different situation is observed for both models on the distress task, as the correlation scores are far higher and are considered strong. On the distress task both model could predict values close to six, but still they do not predict a value close to seven. Possible explanations of this situation are the small number of instances in the training set (1674), which might not allow the models to fit correctly and the difficulty of the task, which I have mentioned previously on the 3.1 section. The difficulty of the task is increased because the scores for prediction are based on the self-evaluation of people, but the comments might be very different. People by answering the questionnaire provided information about what the news stories made them feel, but as each person expresses themselves differently, an exact prediction of their EC and PD scores might not be possible.

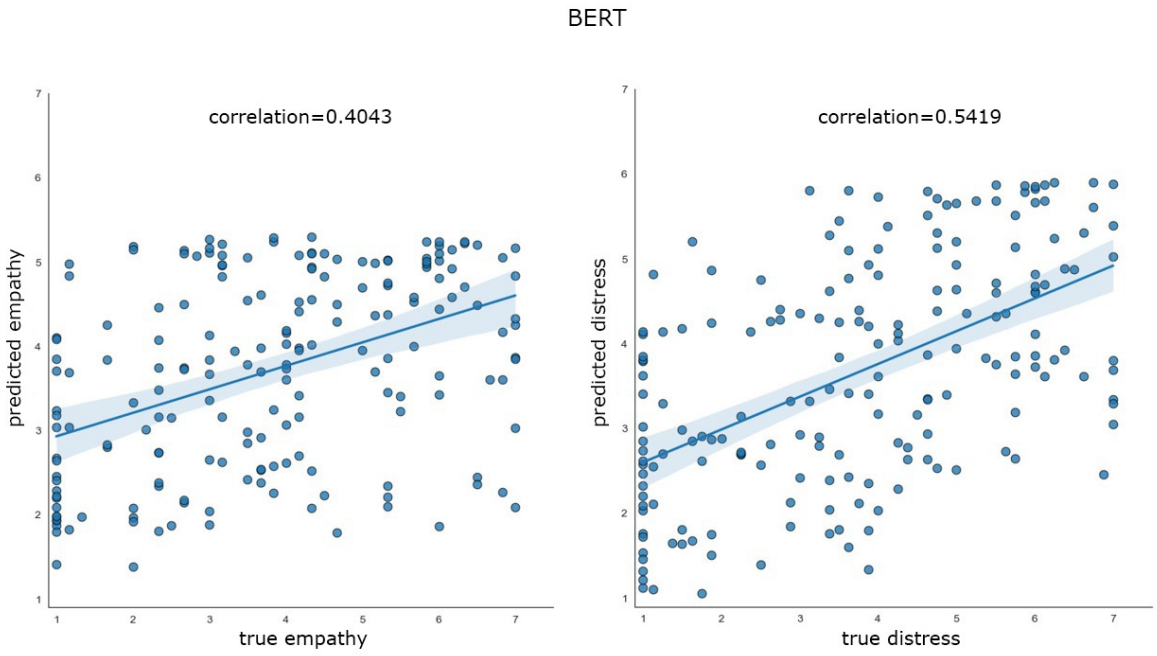

Figure 3.2: Scatter plot with correlation line of BERT predictions on the test set

Residual plots Continuing with the analysis of the test results, I have included the residual plots with the predicted values against the standardized residual values for each model on each task in Figure 3.4. Below, I include the formula on how each residual is calculated.

$$residual = observed - predicted$$

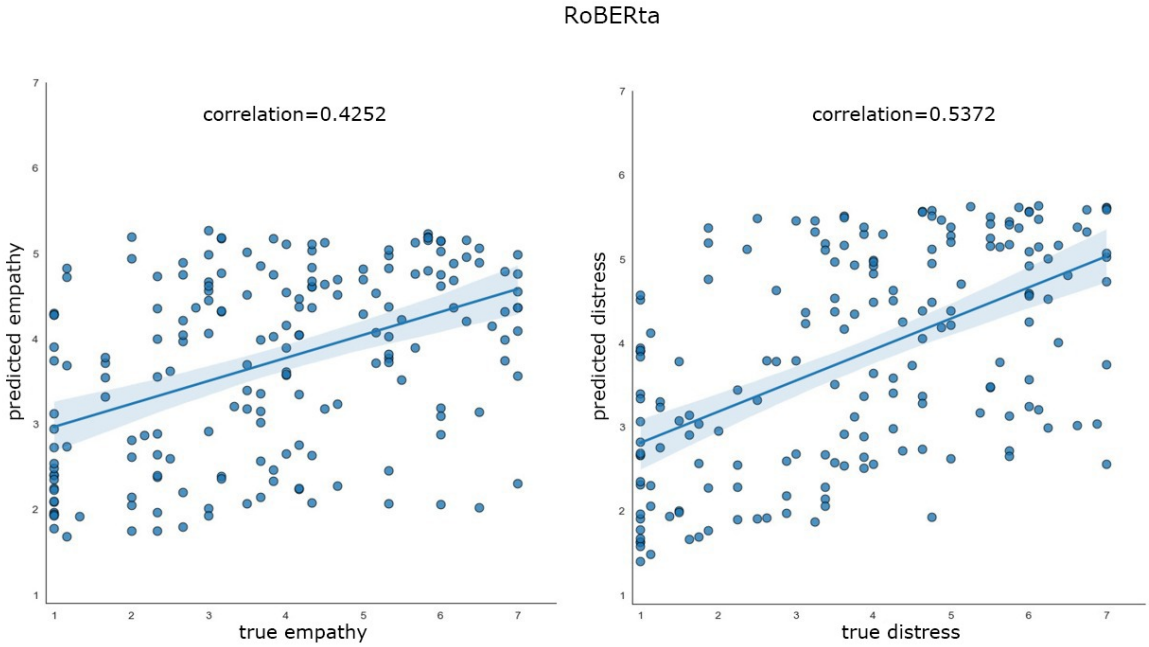

Figure 3.3: Scatter plot with correlation line of RoBERTa predictions on the test set

The residual plots for all cases do not showcase any clear prediction patterns on the residuals and are evenly distributed. Additionally, the Loess curve[2](green line on the plots) remains stable around the zero value, indicating that the models used (BERT and RoBERTa) are capable of fitting these regression tasks. Nevertheless, the plots indicate that the models tend to predict a lower value than the true scores, as the residuals (Y-axis) tend to be larger on the positive side rather than the negative one. In addition, for all residual plots the Loess curve (green line) favours the negative residual side at the predictions close to 1, which means that the predicted values were higher than the observed ones at that stage. Moreover, there are some cases with a high residual values - outliers, especially on the positive side of the Y-axis, which could be cases with a comment that did not align with the EC and PD scores, similar to what was seen during the data analysis at section 3.1. All the plots so far (scatter and residuals) have showed us, the inability of the models to predict values close to the prediction margins (1 and 7).

Cases of bad predictions As a final step in the analysis of the performance on the test set, I investigate some cases that the models did not predict correctly. Table 3.11 presents three cases that both models could not predict correctly on both regression tasks. More specifically, these three cases have one of the highest average (empathy and distress) absolute residual value on both BERT and RoBERTa. Table 3.12 includes the true EC and PD scores alongside predicted from both examined models.

First sentence The first sentence has a true value of seven on both tasks, but both model heavily under-predicted these scores. Possible explanations for this

[2]https://en.wikipedia.org/wiki/Local_regression

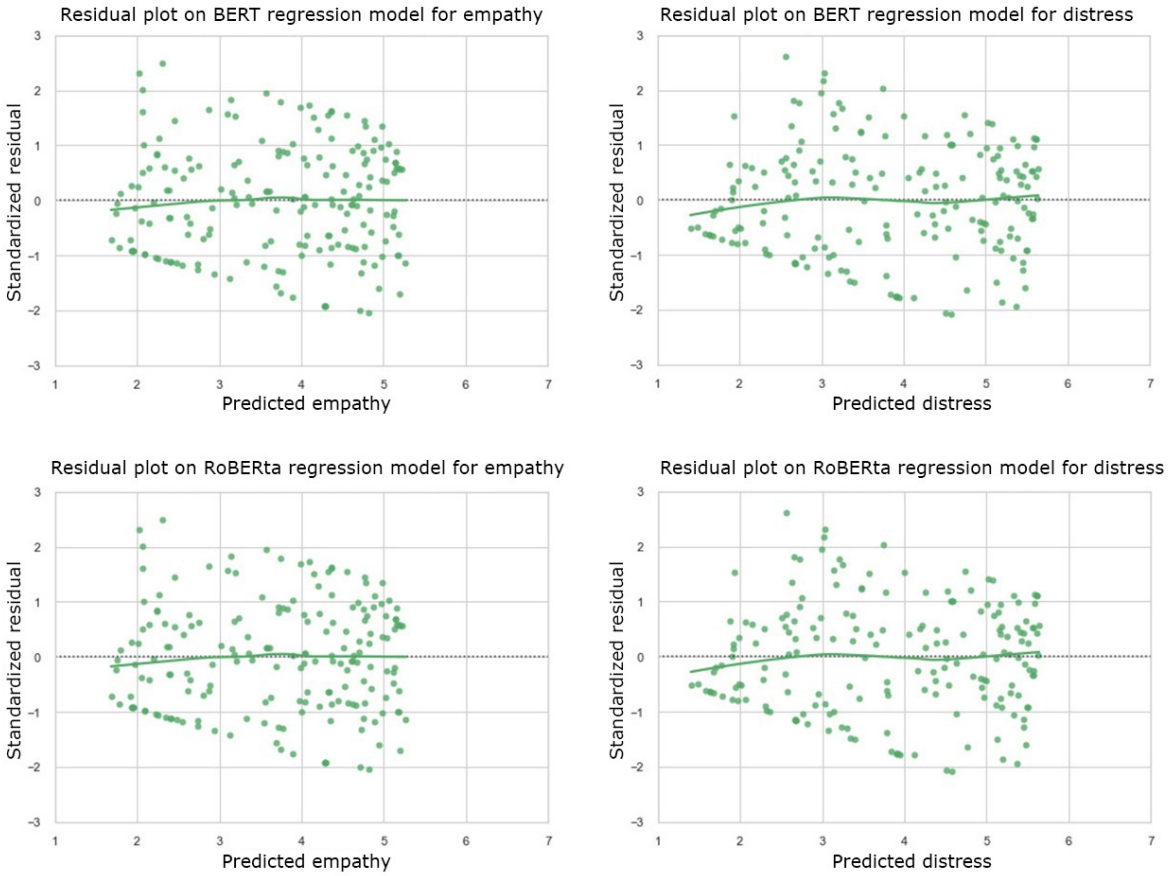

Figure 3.4: Residual plots for both BERT and RoBERTa models on both EC and PD tasks. Y-axis: residuals, X-axis: predicted scores. The green line in the plots are the Loess curve, indicating if there are patterns in the prediction of the models.

underestimation could be that the author used the words "the good news" in the beginning of the sentence, which might hinted to the model that the situation is not linked with high empathy and distress, or that the language used does not show a person who experiences high empathetic concern or empathetic distress.

Second sentence The second sentence had a true value close to 1 on both tasks, but both models over-predicted the scores. By reading the comment, the language used express emotions like sympathy and compassion related to empathetic concern and thus confusing the models. Especially the phrase "It's so sad for their families!" does not align with the person's self-evaluation. In regards to personal distress, the sentences do not have clear signs of PD, except the already mentioned sentence.

Third sentence In this sentence we observe the same situation with the previous one, as the true scores are 1 on both tasks and the model over-predict them. The author of the sentence describes what they read in the article in their comment but they are not empathetic or distressed. The situation that they describe though includes words, like "bombing", "casualities" and more, which usually relate to EC and PD, possibly explaining the higher predictions on the scores.

index	sentence
1	The good news is we live in a place that does not have much air pollution. The bad news is that people that have lung cancer that is exposed to air pollution may have a shorter survival time. Air pollution kills thousands of people each year. Researchers claim that the median survival for people diagnosed with early stages of lung cancer is expected to live 3 times shorter amount of time than those that do not. Of course, we all should stay away from cigarette smoke but it is unfortunate that we have air pollution. They can figure out a way to get to the moon but can not seem to figure out a way to clean our air.
2	It's bad that these 2 died, but I have to believe they had to have know the risks involved. I think so many of us, even knowing the risks, think that "it can't happen to me" and do things we shouldn't. It's so sad for their families! I'm not sure what is the right thing to do. Signs are posted, and divers know the risks. But would it be better to permanently close the area? I'm on the fence.
3	There was more bombing over in Syria in the news over the weekend. There were over 300 casualties including small children. A father and son were found dead in the rubble as well as two boys whose mother survived in an adjoining room. It seems like the violence in the Middle East is never ending.

Table 3.11: Cases of bad predictions. The colored words indicate what parts might have lead to the bad predictions.

index	true EC	true PD	BERT EC	BERT PD	RoB. EC	RoB. PD
1	7	7	3.847	3.054	3.566	3.75
2	1.167	1.125	4.98	4.818	4.822	4.125
3	1.0	1.0	4.085	4.147	4.282	4.517

Table 3.12: EC and PD scores of cases that produced bad predictions

These three cases enforce the observations so far. The first and second sentence are cases that even a human-rater would assign different values, as in the first case the text does not show signs of extreme distress or empathy and in the second case the text does not indicate the low scores. The third case could be explained by the small number of utterances in the dataset, as the commentor shows apathy, but the situation they describe usually creates stronger emotions.

3.2.2 Augmented dataset

After completing testing using the original dataset, I continued with investigating if the data augmentation methods of Easy Data Augmentation (EDA) Wei and Zou (2020a), back-translation Li et al. (2020) and contextualized word augmentation Kobayashi (2018a) would increase the performance of the selected models. The data augmentation methods were tried due to the size of the dataset, 1860 instances. For each method I augmented a single sentence per sample, so the total number of

samples becomes 3720.

EDA settings Let us first discuss the settings of each method separately. EDA provides four mechanisms to augment a sentence on word level. These mechanisms are synonym replacement (using Wordnet), random deletion, random insertion and random swap (Wei and Zou, 2020b). For our purposes I chose to use only synonym replacement, with a probability of 15% for each word, and random deletion, with a probability of 5%. These decisions relate closely to the default settings. My settings differ from the default at the synonym replacement, which I increased from 5%, and at the random deletion, which I decreased from 10%. These changes are based on my intuition that randomly deleting words has a high change of altering the context of the sentence, thus I wanted to minimize this mechanism and increase the synonym replacement. To implement this method on this data set I used the approach described here [3].

Back-translation settings For back-translation the only decision was to select to which language I wanted the original text to be translated before getting it re-translated, so it can be paraphrased. I choose to translate each text in French and back to English, as it is the default option (Sennrich et al., 2016). For this method I used the MarianMT model produced by Hugging face [4].

Contextualized word augmentation settings The final method, contextualized word augmentation, examines the words by looking the BERT based word-embeddings and decides which words should be replaced. The new word is decided by looking the word-embeddings of the surrounding words. For this method, I chose to replace 30% of the words in the text, as recommended (Kobayashi, 2018a). Table 3.13 presents an example of an augmented sentence for each method. More details about the methods are presented in section 2.4.

Methodology for data augmentation After creating the augmented data sets, I decided to follow the same methodology as before to be able to compare the produced models. In more details, for each data augmentation dataset I explored both BERT and RoBERTa models, with the selected hyper-parameters, on both empathy and distress. The 10-fold CV method was implemented on each model for comparability between these models with the non-augmented models and the baselines, but with an alteration from before. First, I used the non-augmented sentences to create the 10-folds and then I called the augmented instances that I need for training, with the intention of avoiding having augmented sentences in the validation set of the 10-fold CV. For the BERT models two epochs were used during training, instead of the three epochs that were used while training with the non-augmented data, as signs of over-fitting started to appear. On the other hand, for the training of RoBERTa models three epochs were used, similarly to the non-augmented methods.

Results of 10-fold CV using Data augmentation In Table 3.16, there are presented the results of the 10-fold CV for BERT and RoBERTa models per data

[3]https://github.com/jasonwei20/eda_nlp
[4]https://huggingface.co/

method	*sentence*
Original	no matter what your heritage, you should be able to serve your country. her thai heritage shouldn't preclude her and shouldn't have been an issue in this debate. tammy duckworth and her family should be congratulated on the services they have provided to this country. any type of racism should not be allowed in a debate
EDA	no matter what your heritage you should be able to answer your nation her thai heritage shouldnt preclude her and shouldnt have been an issue in this disputation tammy duckworth and her family should be congratulated on the table service they have bring home the bacon to this nation any type of racism should not be allowed in a disputation .
Back-translation	Whatever your legacy, you should be able to serve your country. its Thai heritage should not prevent it and should not have been a subject in this debate. Tammy Duckworth and his family should be commended for the services they have provided to that country. any kind of racism should not be allowed in a debate
Contextual	no matter what your heritage, you should be able to serve a country. her thai heritage shouldn't preclude her and shouldn't ever developed an issue in this debate. tammy duckworth and her family should be congratulated of the care they have provided to this country. any sort of racism should n be allowed in any debate.

Table 3.13: Example of augmented sentence

augmentation method. As we can observe, not one out of the six options surpassed the non-augmented BERT model, which scored an average Pearson's correlation of 0.466. The model that performed the closest was the RoBERTa model that used the augmented data of EDA, with a an average score of 0.4613. It is important to note that this RoBERta model is the only model that was able to surpass the performance of its non-augmented version by 0.0002, which is a very minimal increase.

model	**empathy**	**distress**	**average**
BERT EDA	0.4149	0.4856	0.4502
RoBERTa EDA	0.4387	0.4839	0.4613
BERT BT	0.4237	0.4637	0.4437
RoBERTa BT	0.4326	0.4798	0.4562
BERT CE	0.4403	0.4736	0.4569
RoBERTa CE	0.445	0.472	0.4585

Table 3.14: Results of 10-fold CV with the data augmentation methods of Easy Data Augmentation (EDA) , back-translation and contextualized word augmentation.

Analysis of the results of the 10-fold CV In general, these six Transformer models achieved average Pearson's R correlation scores close to each other. As we can observe, there is a drop of performance at the prediction task of EC compared to the non-augmented models and for all models but one, there is an increase of performance at the regression task of PD. By treating each task separately, the small performance increase for the PD task is in agreement with the expectation of an improved performance. More specifically, out of the three methods the worst performing was Back Translation, which could have been considered the method that changes a sentence the most. The least sophisticated method, the EDA was the method with the highest average results. In EDA words were replaced by their synonym or deleted randomly, which might have altered the meaning of the phrase less than the translated sentences, as the random chosen word might not offer much in the context. The translated sentences could have replaced words that contributed to EC and PD scores strongly. Finally, for all three methods the RoBERTa model out-performed the BERT model and in the case of EDA and BT by a considerable difference.

Predicting on the test set using the EDA RoBERTa model To get a better understanding how the data augmentation method of EDA altered the behaviour of the RoBERTa model, I used the training set and its augmented sentences to train the model and then predicted on the test set, in a similar way as the non-augmented methodology. In Table 3.17, we can see its results and the results of the non-augmented BERT and RoBERTa models, which were already presented in Table 3.9. The average person's r correlation score of this augmented model is lower than both models, validating that the EDA did not increased the performance of the dataset. In the case of EC though, the augmented RoBERTa model was able to surpass the BERT model by a small margin. Interesting enough, we can see the same behaviour in all three models, with a decrease on the EC score and an increase to the PD score in comparison to the 10-fold CV.

Scatter plots for the EDA RoBERTa model The scatter plots in Figure 3.5, indicate this same predictive behavior. The EDA RoBERTa model on the regression task of distress shows an ability to predict closer to the margins (1 and 7) in comparison to the empathy task, similarly to the non-augmented models. In case of the empathy task, the problematic predictive behavior of the models got worse in comparison to the non-augmented ones. The EDA RoBERTa model predicted values only between two and five, excluding from its predictions the scores from one to two and from five to seven.

model	B.s.	l.r.	w.d.	Emp. corr.	Dis. corr.	average
BERT	16	5e-05	0.1	0.4043	0.5419	0.4731
RoBERTa	8	2e-05	0.2	0.4252	0.5372	0.4812
RoBERTa EDA	8	2e-05	0.2	0.4102	0.533	0.4716

Table 3.15: Test results of the non-augmented models and the EDA RoBERTa model. B.s: batch size, l.r: learning rate and w.d: weight decay.

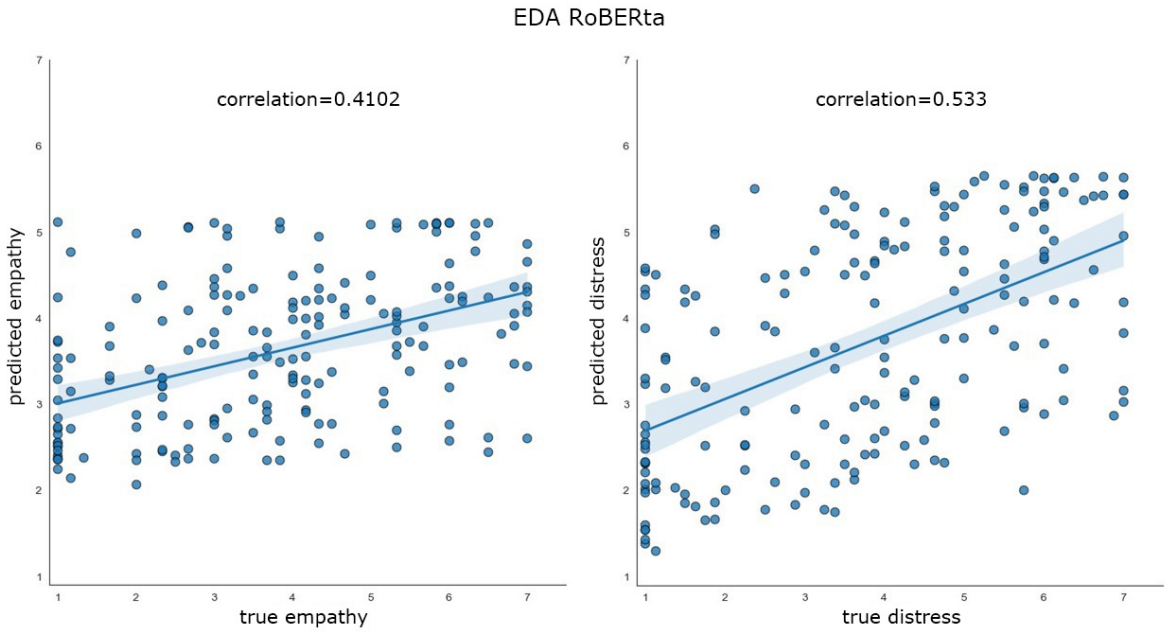

Figure 3.5: EDA RoBERTa test scatter plot with correlation line

Additional augmentation methods Since data augmentation methods on the whole dataset did not improve the results of the selected models and the problematic predictive behaviour of not predicting close to the margin remains, there is room for further exploration. I decided to attempt to tackle this favouritism towards the lower scores in the dataset that was observed during the residual study and might be the reason of not predicting values above five. More specifically, the dataset contains 568 instances with a score of 2.5 and lower and 380 with a score of 5.5 or higher. Towards that end two ideas were tested. The base of both ideas was to increase the samples that are under represented in the dataset. First, I decided to still use the data augmentation methods, but only on the instances that have a score above four. Second, I decided to balance the dataset by duplicating only once the instances above four, so augmenting the data with the same instances that already exist. The number four is very close to the mean, but this threshold was decided to avoid infusing to the dataset only extreme cases. Both ideas were first tested using the 10-fold CV and the augmentation happened on the instances of the training set per fold, so no information of the validation set would have been passed by the training set.

10-fold CV results on the additional data augmentation experiments In Table 3.16, there are placed the results of both ideas. In total, there are presented eight models, the first six for the data augmentation methods and the last two for the augmentation with duplication. The most succesful model was the BERT model without a data augmentation method, which scored the highest average correlation on the distress task and the fourth highest on the empathy task. Out of the data augmentation methods, with these settings the best performing method is back-translation. The BERT model of back-translation succeeded the second best empathy score at 0.4444 and its average is 0.0012 lower than the not augmented

BERT model. In a similar way as before, no method produced a model in the empathy task that surpassed the performance of the selected models with the original datasets.

model	empathy	distress	average
BERT BT	0.4444	0.473	0.4587
RoBERTa BT	0.4319	0.4798	0.4558
BERT EDA	0.4224	0.4604	0.4414
RoBERTa EDA	0.4479	0.4667	0.4573
BERT CE	0.4131	0.452	0.4325
RoBERTa CE	0.4303	0.4684	0.4493
BERT balanced	0.4366	0.4831	0.4598
RoBERTa balanced	0.4419	0.4732	0.4575

Table 3.16: Results of 10-fold CV with data augmentation methods applied only on utterances with scores higher than 4

Testing of BERT balanced model These new attempts might not have produced a more succesful model, but it would be beneficial to investigate the predictive ability of one on the test set to observe its predictive behavior. For this reason, I decided to use the BERT model that was augmented by the original dataset (BERT balanced) and predict on the test set. In table 3.17 there are presented the results of this BERT model, with the original BERT and RoBERTa models. The BERT balanced model during the testing phase scored a higher Pearson's R correlation on PD compared to the 10-fold CV and lower on EC. This trend was seen in all the combinations of models that were examined on both the 10-fold CV and on the testing set. In addition, the BERT balanced model performed better than the others in both empathy and distress and it was almost able to reach an average Pearson's R correlation of 0.5. Next I performed a paired two-tailed t-test for paired samples between the non-augmented BERT model and the BERT balanced model. The results of the test are placed in Table 3.18 and in both cases the difference between the models is significant.

model	B.s.	l.r.	w.d.	Emp. corr.	Dis. corr.	average
BERT	16	5e-05	0.1	0.4043	0.5419	0.4731
RoBERTa	8	2e-05	0.2	0.4252	0.5372	0.4812
BERT balanced	16	5e-05	0.1	0.4312	0.5678	**0.4995**

Table 3.17: Pearson's r correlation Test results test results of the non-augmented models and the BERT balanced model. B.s: batch size, l.r: learning rate and w.d: weight decay.

Scatter plot of balanced BERT model Figure 3.6 presents the scatter plots on both tasks from the BERT balanced model. For the first time, we see that one model on the task of empathy is able to predict higher than five, but the model still

type of comparison	statistic	p-value
EC	-3.528	0.0005
PD	-7.272	0.000000000009

Table 3.18: T-est results between the non-augmented BERT model and the BERT balanced on both tasks of EC and PD. The p-values on both tasks are lower than 0.05, so in both cases the difference between the models are significant.

has a limit at six. For the distress task, the model has succeeded the highest testing correlation value so far and it is observable on the scatter plot.

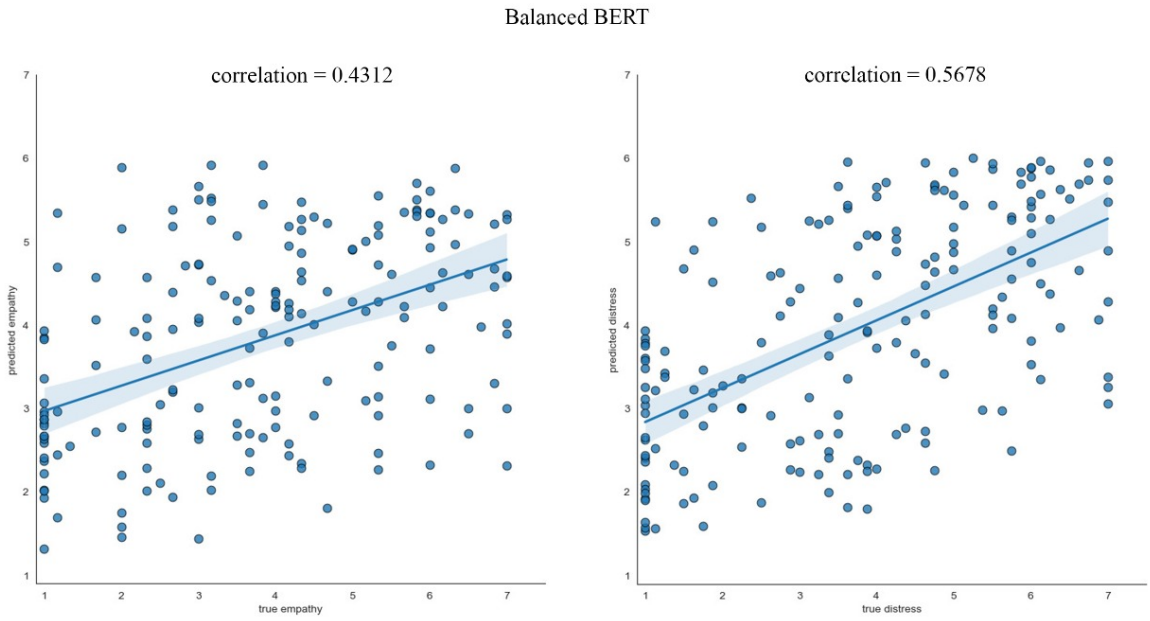

Figure 3.6: augmented BERT with original dataset test scatter plot with correlation line

3.2.3 Model selection

After the completion of all experiments and the investigation of the results, it is to select a model that will be used during the second research question. To compare all the models, I used both a 10-fold CV and these models were tested on the test set. During the 10-fold CV the non-augmented BERT model achieved the highest score (Table 3.8, score 0.466), but on the prediction of the test set the BERT balanced model performed best (Table 3.17, score 0.4995). The t-tests that followed the predictions on the test set, showed that the difference between the BERT balanced model and the rest is significant. All these tests lead me to choose the BERT balanced model for usage on the second research question. In addition, my decision was supported on the scatter plots of Figure 3.6, which indicated a better predictive ability from the model on both tasks.

BERT balanced The BERT balanced model, the hyper-parameters of which are seen in Table 3.17, is the chosen model. In summary, for this model I increased the samples of the dataset during the training phase by duplicating the instances that had a score above four on each task. This decision was based on the fact that the previous models tended to not under-predict scores.

3.3 Discussion

Before continuing to the next research question, I first summarize the results of my experiments and reflect on how they relate to the research question. Moreover, I discuss other findings that occurred during this chapter. With the first research question I asked "What is the performance of the transformers architecture on predicting the empathetic concern and personal distress scores of reaction comments on news articles ?"

Summary of first research question For the two regression tasks of predicting the EC and PD scores, I used two different pre-trained Transformer models, BERT and RoBERTa, which I fine-tuned on the dataset. Both BERT and RoBERTa models improved upon the baselines, which are the models inherited by Buechel et al. (2018). The BERT and RoBERTa models, that were selected after a thorough cross-validation phase, had an increase higher than 10% approximately on the average Pearson's R correlation score of the best model (CNN). I used the average score of the EC and PD regression tasks, similarly to Buechel et. al. to be able to compare their results. These results showcase that the Transformer architecture is more capable on predicting empathy and distress compared to other methods used on natural language processing (Ridge regression, FNN and CNN). Although there was a significant improvement, the average Pearson's R correlation of the selected model (BERT balanced) fell shortly below 0.5 (0.4995) and did not pass the threshold to be considered significant (0.5). The BERT balanced model was selected to be used on the second research question. This decision was based on the fact that it performed better on the test set compared to the non-augmented BERT and RoBERTa models in a significant way according to the t-test. Although the model did not surpass the 0.5 threshold, I argue that these results are strong and the Transformer architecture is more than capable of predicting the EC and PD scores, as the task at hand has an increased difficulty due to the small number of instances in the dataset.

Difficulty of task There are multiple reasons why these two tasks can be considered difficult. I believe that the main reasons of the difficulty is that people self-evaluated themselves before commenting and that the dataset is small. In my opinion, the self-evaluation of the comments is beneficial as not all people think alike, thus leading to a stronger dataset that examines different ways of thought that could lead to models that generalize better. But at the same time there is the risk of people not evaluating their emotions correctly. Unfortunately, the small number of instances does not allow the model to generalize well though. In multiple instances of the dataset that were examined, the context of the comments did not align with what the person's evaluation (Table 3.11 and 3.12). This situation highlight that not all people express what they feel in the same way. This limitation can

be surpassed by an increased amount of instances in the dataset. In addition, the language used in the dataset raises the difficulty of the task, as constructs such as sarcasm and acronyms are observed in the data. The existence of these constructs though make the data more legitimate.

Data augmentation During the process of answering this research question, I experimented with data augmentation methods to increase the instances in the training set. The data augmentation methods that I tried (back-translation, easy data augmentation and contextualized embedding) did not increase the performance of the models. The expectation before the experiments was a small increase in the performance of both regression tasks, as the number of utterances is small, but this was not the case. The results showed a negative effect using the data augmentation methods, supporting the work of Longpre et al. (2020), in which they examined pre-trained Transformer models with various data augmentation methods, including EDA and BT, on different tasks. Their conclusions were that on pre-trained Transformer models these methods do not improve the performance, even on very small datasets. Interestingly enough, the least sophisticated data augmentation method (EDA) was the one that performed better (Table 3.14) in our case. My belief is that the reasons behind the drop of performance are that by trying to augment the comments, the meaning is altered leading to a decrease in their performance due to distorted context. The EDA, which is the least sophisticated, as a method randomly deleted words or replaced them with their synonyms and in many times it might not altered the meaning of the sentences.

Correlation with length Finally, I also examined the correlation between the length of the text and the empathetic concern and personal distress scores on the samples of the dataset. The reason behind this is that on the second research question, I used the selected model to examine the EC and PD scores of Twitter and Reddit. From these two platforms, Twitter has a character limit for the user's post and Reddit does not. So, there was the question if there is a correlation between those two and then if this correlation holds. There was seen a small correlation between the self-reported EC and PD scores with the length of the text (Table 3.3).

Chapter 4

Empathy detection on Twitter and Reddit

In this chapter, I answer the second research question of my thesis. "Do the comments reacting to news stories on Twitter express less personal distress and more empathetic concern compared to the comments on the same news articles on popular news subreddits?". To answer this question I needed a dataset that includes news articles that have been posted on Reddit alongside the users' comments about each article and tweets discussing the same news articles. Unfortunately, a dataset fitting that description was not available, thus I created one with tweets and Reddit comments about the same news articles. The process of constructing this dataset is described in the first section of this chapter (section 4.1). In section 4.2, I present the analysis of the crawled data to get some insight before continuing with the predictions. After gathering, pre-processing and analyzing the data, I used the BERT balanced model from the first research question to predict the EC and PD scores, in section 4.3. Finally, in section 4.4 I discuss observations produced during the experiments and answer the research question. Before continuing with the rest of the chapter, I present briefly the platforms of Reddit and Twitter.

Reddit Reddit is a social platform with approximately 500 million active users per month [1]. The structure of Reddit resembles the structure of a forum with smaller groups. It is consisted of sub-communities, the subreddits, which can be created by any user, they can revolve around any subject and be regulated by their own members. Users decide on their own which subreddits they want to belong in and follow them, with the exception of some subreddits that everyone belongs to when they join the platform but they are free to unsubscribe. Users are able to create a post about anything on the proper subreddit and other users can up-vote it, down-vote it or comment on it. The post can either be an external link or a self-post, which doesn't link to an outside source and it usually contains only text. Important to note, there is no length limit for neither a post or a comment. When a post is created it is placed on the "new" list of the subreddit, but if it gets lots of upvotes it is added to other lists with higher importance and more people can see it. These are named "rising", "trending" and "top". Every user can either see posts from inside each subreddit or follow their "front-page" where posts are shown from all the subreddits they follow.

[1] https://www.businessofapps.com/data/reddit-statistics/

Twitter Twitter is a different social media platform compared to Reddit. It is a "micro-blogging" system that allows users to either send or receive posts, called tweets. Tweets were first limited to 140 characters but this was recently changed to 280 characters. They can contain text and external links. Twitter does not have a concrete structure as Reddit. Its members can "follow" other users and see their tweets in their "timeline". A person can follow any other user, except if they are blocked by them. A user can be a real person or an organization. Users have also the ability to re-tweet another user's tweet, meaning that they republish it for their followers to see.

4.1 Dataset creation

Since there is no public dataset that meets the criteria for this research question, the creation of one became the first step for answering the question. A necessary step for the creation of the dataset was to identify tweets and Reddit comments on the same news articles, with the intention of minimizing the factors that could be responsible for the different scores of empathetic concern and personal distress. Having this in mind, the pipeline for acquiring the data begun its process with Reddit, because on Reddit one can identify popular news stories by looking at the "hot" pages of News subreddits. So, the pipeline's initial step was the identification of news stories from the "hot" pages of three news related subreddits, then it crawled the comments of them that met the criteria, which are explained later. Next, for each of the selected articles the system searched for tweets discussing them. First, I describe in more details the part of the pipeline dedicated to Reddit and then describe the process for Twitter.

Reddit The subreddits that were used for this dataset are r/news, r/worldnews and r/politics. These are the most popular news subreddits, their language is English and they do not side with a particular political opinion. In addition, all focus on news coverage but have a different scope. R/news is dedicated to American news stories, r/worldnews has no geographical restriction and r/politics mainly focuses on American political news. As previously said, the pipeline crawled articles that were listed on the "hot" page of each subreddit. This process was executed daily between the dates of 07/04/2021 and 10/05/2021, except the days were the procedure failed. There were also additional criteria for a post to be selected. It was required to not be a self-post, as these couldn't be linked to Twitter, and to have 1000 or more comments. This threshold was placed arbitrary to limit the number of selected articles, because crawling from Twitter has a time limit and during trials without a threshold, the pipeline required more than a day to crawl tweets about just one day. So, my intention was to only crawl news stories that showed user interaction. Then from these posts I crawled the immediate replies of users (comments on the article) and the replies to the replies, but I avoided crawling further down the reply's tree. This decision was based on the fact that I couldn't be sure that the additional comments would still comment on the article or have a different topic. Except from the body of the comments, I also extracted the id of the article and its URL, the ids of the comments, the number of upvotes and the date of the comment. For all

these tasks, I used the free to use PRAW API [2] in python to connect with Reddit and get the necessary information.

Twitter After selecting the articles from Reddit, I was able to crawl tweets and tweet replies that discuss these articles. For the process of crawling from Twitter I used Tweepy [3] to connect with Twitter's API. To search tweets while interacting with Twitter's API one has to formulate queries that will return tweets that match the exact query. After experimenting with different queries, I decided on the following strategy: search for tweets that include the URL of the article or tweets that include the title of the article, excluding retweets. Moreover, from these crawled tweets I searched and crawled the replies from other users. The only requirement for the crawled tweets/replies was to be in English. There are other works that followed a different strategy for their queries (Priya, 2018)(Wei and Gao, 2017). They separated the title into bi-gram pairs of words and crawled tweets that matched. With that strategy the scope of the search grows and the query could return a lot of tweets that are relevant, but they might discuss the topic and not that specific news story. In my case, the methodology that I followed for the queries might limit the number of crawled tweets per article but, as I mentioned previously, making sure that the crawled tweets speak about the specific article was the priority. I understand though that aiming for high precision, will affect negatively the number of crawled tweets, as not all people tweet about a news story alongside the title or the url of the story.

Preprocessing After acquiring the comments from these platforms, these needed to be preprocessed. For all comments and tweets I followed the same preprocessing steps. First, I lower-cased the texts, as the model used for prediction was trained on lower-cased data. Also, I removed HTML characters, URLs, symbols that are used on Twitter for replies and usernames that follow them and replaced emojis with their corresponding word, using the python module demoji [4]. These steps were necessary because the training instances contained plain text. Moreover, I had to ignore Reddit comments that were removed from the users, as they are still on the comment tree but with an empty body. Besides looking on the comments themselves, I investigated the crawled articles and I removed the duplicate instances of articles on the same subreddit, for which I had duplicate comments and tweets. Those duplicate articles were created because some posts remained on the hot page for consecutive days.

For more details about the dataset, I completed a data sheet that describes the dataset following the methodology of Gebru et al. (2018), which can be found at the appendix.

[2] https://praw.readthedocs.io/en/latest/#
[3] https://docs.tweepy.org/en/latest/index.html
[4] https://pypi.org/project/demoji/

4.2 Dataset Statistics

In this section I present statistics about the dataset. An analytical view of the created dataset plays a crucial role on deciding some of the next steps for answering the research question. First, I present statistics about the articles and then examine Reddit and Twitter separately.

4.2.1 News stories

Number of articles After the preprocessing step of removing the duplicate instances, the dataset includes 406 articles and the Reddit comments and tweets about them. The articles have been collected from most dates between the 07/04/2021 to 10/05/2021, as previously mentioned. Figure 4.1 presents the number of articles per date. The maximum number of articles crawled on a single day is 21 in 20/04/2020. From the 406 articles, 171 instances were posted on r/news, 101 on r/worldnews and 134 on r/politics. It is interesting that the subreddit of r/politics is not on the last place in terms of articles, as each new user of Reddit is automatically subscribed on r/news and r/worldnews. This shows that even with a smaller audience, r/politics has a more constant user interaction, as there is the threshold of 1000 comments. Out of the 406 articles crawled only 1 news story is represented twice, so the number of unique articles is 405.

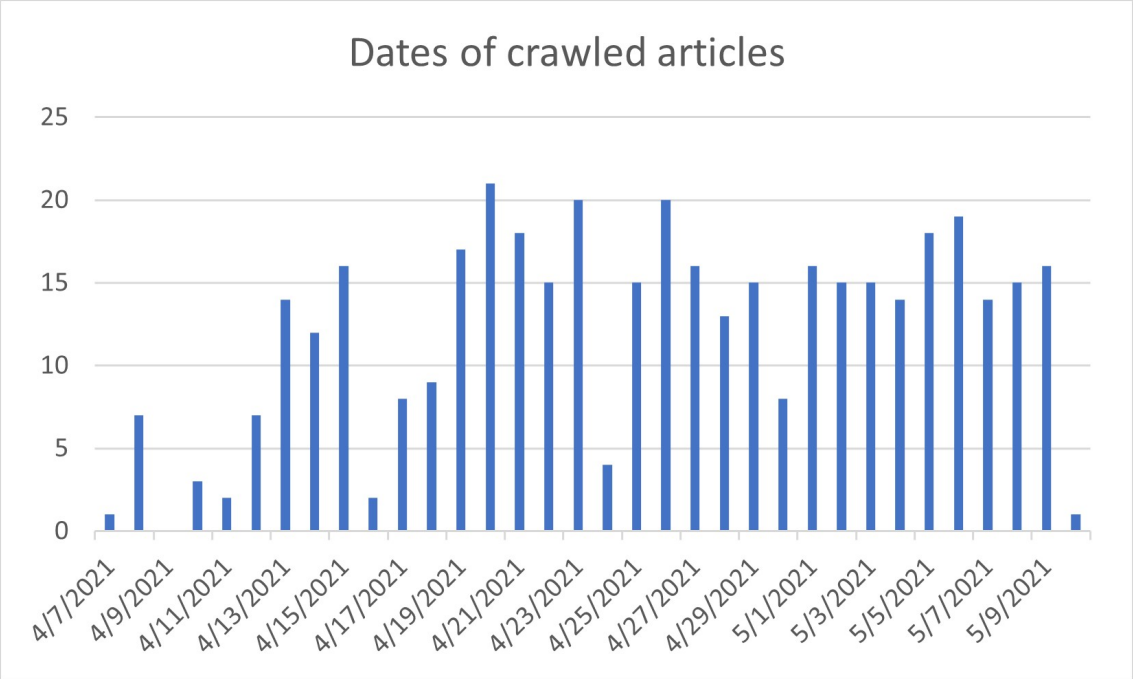

Figure 4.1: A column chart with the number of crawled articles per date between the period of 07/04/2021 to 10/05/2021.

News sources Before continuing with investigating the comments and tweets about the news stories, it is interesting to examine which news sources are represented in the data. Figure 4.2 presents a bar chart showcasing the 13 most represented news sources, with CNN topping the list with 23 instances and BBC last

in the list with 9. Out of the 406 articles, 204 come from these top 9 news sources, which means that the dataset contains a variety of news sources and is not biased towards a specific one. To be specific, the number of unique news sources is 143. The low number of articles per news source, even for CNN, does not allow for a comparison between news sources though.

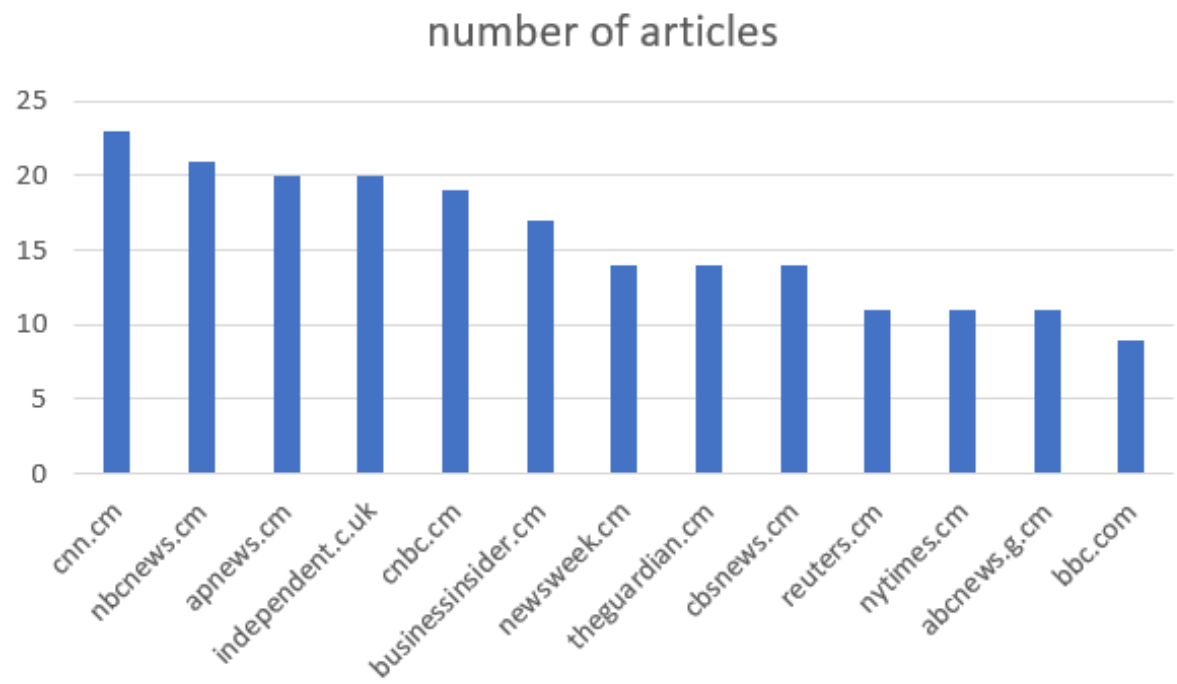

Figure 4.2: A bar chart with the 13 most frequent news sources in the dataset.

4.2.2 Reddit

Comments on Reddit Let us continue with the analysis of Reddit comments. First, I investigated the number of all the crawled comments, both immediate comments and user replies to these comments. The number of comments per article is close between all subreddits as it is evident from Tables 4.1 and 4.2, with r/news having the most comments crawled per article (893.27) and r/politics having the least (742.45). On the other hand, r/politics has a considerable smaller standard deviation on comments per article, 395.97 compared to 612.41 and 546.54. This can be explained from the fact that r/politics subscribers chose to subscribe and follow it. In the subreddits r/news and r/worldnews users are already members when they join the platform and they might be less engaged with them, as previously mentioned. For example, subscribers of r/news and r/worldnews might only interact with the posts that reach their front-page and not follow all posts from them. To avoid confusion, the threshold of 1000 comments for a news story to be selected was placed on the total number of comments, but the crawling pipeline crawled only the first two levels of the comment tree.

Immediate replies versus all crawled comments Table 4.2 presents the number of the immediate replies only. In this situation, again r/politics has less com-

ments per article, but the difference between the subreddits is smaller than before, 493.47 comments per article compared to 545.69 and 584 from r/news and r/worldnews. By looking the standard deviation of the comments per article, r/politics again has a user engagement with the lowest standard deviation of the three at 299.67. In all three subreddits, the crawled comments are immediate comments and not replies, as the average number including all subreddits dropped from 839.95 to 537.99. This number of immediate comments gives me confidence to continue with only them for comparing Reddit and Twitter, as the dataset that I used for training the model of the first research question included users comments on the articles and the additional replies might infuse noise to the analysis.

Subreddit	articles	comments	per article	std
News	171	152750	893.27	612.41
Worldnews	101	88783	879.03	546.54
Politics	134	99489	742.45	395.97
All	406	341022	839.95	536.1

Table 4.1: A table with the number of articles and the number of all comments (both immediate comments and replies to the comments) per article.

Subreddit	articles	comments	per article	std
News	171	93314	545.69	469.94
Worldnews	101	58984	584	438.94
Politics	134	66126	493.47	299.67
All	406	218424	537.99	413.49

Table 4.2: A table with the number of articles and the number of the comments on the article (excluding replies to the comments) per article.

Length of comments on Reddit Let us now explore the length of the immediate comments only per subreddit. Table 4.3 presents the mean number of tokens (words) of the comments and their standard deviation. The comments of the worldnews subreddit show a different behaviour compared to the other two subreddits by having a mean of 21.79 compared to approximately 26. In general, the standard deviation, which is similar across all subreddits, is high, but that is logical as Reddit does not enforce a character limit on the user comments. So there might exist one word comments or comments that last multiple paragraphs, which is verified in Table 4.4. For all subreddits the shortest comment has a length of 1, which is expected. The longest comment in the dataset belongs to r/politics with a length of 1704 words, which is a shockingly long comment.

4.2.3 Twitter

Crawled tweets Continuing with Twitter, Table 4.5 displays the total number of tweets, the mean and standard deviation per subreddit. The r/news subreddit has

Subreddit	mean length	std
News	26.39	37.11
Worldnews	21.79	37.14
Politics	26.01	38.25
All	25.06	37.52

Table 4.3: A table with the mean length of the immediate comments on Reddit per subreddit and the standard deviation of it.

Subreddit	max length	min length
News	1293	1
Worldnews	1605	1
Politics	1704	1

Table 4.4: A table with the length of the longest and shortest comment per subreddit.

the highest mean of tweets per article with 41.68, followed by r/politics with 37.44, which is the opposite situation from the Reddit comments in which r/worldnews had the most immediate comments per article. Additionally, r/news has the highest standard deviation by a wide margin. Looking at the bigger picture, the high standard deviation, for all sources, is a result of the decision to have high precision on the crawled tweets. The formulated queries were very specific and unfortunately in some cases the number of crawled tweets is low. In comparison with Reddit, the number of tweets is far lower than the number of Reddit comments, even with the exclusion of the reply comments. The Reddit comments were 218424 compared to 15618 tweets. This is again a product of the high precision. Nevertheless, I will compare the two platforms by comparing the the scores of each article, so the difference of the number of comments per article will not affect the results.

Subreddit	tweets	per article	std
News	7128	41.68	48.55
Worldnews	3472	34.37	33.11
Politics	5018	37.44	34.73
All	15618	38.46	40.81

Table 4.5: A table with the number of all tweets, the mean and the std per subreddit.

Tweets similar to the title After examining individual tweets I noticed that there were tweets that had as their text the title of the article only, without any more input from the user. I decided to explore this further. Table 4.6 presents the number of tweets that are dissimilar to their titles. For the process of judging which tweet is similar to the title of the article that they speak about, I used cosine similarity [5]. In cosine similarity, each document (tweet) gets compared to the title

[5]https://en.wikipedia.org/wiki/Cosine_similarity

by their vectors of words frequency, which are created by the words they both have with the process of vectorization. Mathematically, cosine similarity measures the angle of the two vectors in the multidimensional space. I decided to set a threshold of 0.8 of cosine similarity after experimenting with the parameter. The number of tweets that are not similar to the title is 12360, so over three thousand tweets were similar to the title.

Discarding of tweets that are similar to the title For this situation of the similar tweets, I decided to exclude them for the comparison of the platforms. People might have decided to use the exact title for their tweets, as it might express how they feel about the situation, but using these tweets would infuse noise to the analysis, as their population size will affect the empathy and distress scores. In addition, the training dataset contained the comments of the users, thus these instances can not be used.

Subreddit	tweets	per article	std
News	5812	34.18	36.8
Worldnews	2643	26.96	28.65
Politics	3905	29.58	26.77
All	12360	30.9	31.92

Table 4.6: A table with the number of tweets that are dissimilar to the title, the mean and the std per subreddit

Length of tweets In table 4.7, the mean length of the dissimilar tweets are presented. For all tweets there is a stable mean approximately of 20 and a similar standard deviation, no matter the origin of the news story. This is logical as Twitter has a character limit of 240. This comes in contrast to the situation from Reddit, as the tweets do not show different behaviour between subreddits. By comparing the mean length between Reddit and Twitter, Reddit has a higher mean which is logical as Reddit does not have that restriction. It will be interesting to see how this affects the EC and PD scores, as on the training set there was a small correlation between both scores and the length of the text, showcased in 3.1.

Subreddit	mean	std
News	19.85	13.24
Worldnews	19.84	13.24
Politics	20.21	13.37
All	19.6	13.17

Table 4.7: A table with the mean length of the dissimilar tweets per subreddit and the standard deviation.

As previously mentioned, for more information about the created dataset there is a data sheet in the appendix of the thesis.

4.3 Empathetic concern and personal distress prediction

In this section I present the results of the EC and PD prediction scores on the created dataset. For both Reddit comments and tweets, I first predicted the EC and PD scores of each instance separately, using the BERT balanced model produced on the previous chapter, and then I averaged these per article. Regarding the applicability of the BERT balanced model on this dataset, there are some assumptions that need to be made. First, the training instances had a character limit, which is greater than the one on tweets, but Reddit comments do not have one. So, I have to assume that these differences do not affect the predictions. Additionally, Twitter users use emojis, which were transformed into text during the pre-processing phase, because the training data did not contain similar tokens (emojis). I assume that the context of the original comment by the commenter does not change with this emoji transformation and there is no loss of information. Moving forward, the observations that occurred during the analysis of the dataset led me on comparing the immediate comments of Reddit against dissimilar tweets only. Moreover, I looked into the differences between the three subreddits of the dataset.

Reddit comments The empathetic concern and personal distress scores of the immediate comments from Reddit are presented in Table 4.8. On a holistic level the EC and PD scores could be considered on the low spectrum, with average predictions of 3.15 and 3.23 respectively, as the BERT balanced models could predict scores between 1 and 7. In addition, the standard deviation of both scores is low as well, making seem that all three subreddits have stable levels of empathy and distress. The EC and PD mean and standard deviation of the BERT balanced model on the test set is also presented in that Table. Both the mean and the std on both scores are substantially higher than the Reddit scores. By looking at each subreddit separately, r/worldnews shows higher empathy and distress scores compared to the other two and it is the only subreddit that the EC score was higher than PD. Plausible explanations for these differences are the different subscribers of each subreddit and/or the different news stories that were discussed on each subreddit.

Not similar tweets compared to comments from Reddit Let us now examine the EC and PD scores of the dissimilar tweets, these are also presented on 4.8. For both empathy and distress the predicted scores are higher than those from Reddit, on both a holistic level and on each subreddit. Moreover, the standard deviation of the PD scores on Twitter is higher than the standard deviation of EC scores and of PD and EC from Reddit comments. Still, the standard deviation observed for both EC and PD scores on both platforms is far lower to the standard deviation observed on the test set. By looking only at the scores of Twitter, there is a different situation compared to Reddit. The tweets concerning posts of r/politics have a higher mean of EC compared to PD (3.53 to 3.69), the tweets concerning r/news have the same means and the tweets of r/worldnews have a higher PD to EC (3.9 to 3.87). In contrast, on Reddit only on r/worldnews the PD scores had a higher mean than the EC ones. Continuing with the comparison of Reddit and Twitter, the average EC of all 406 articles is 3.72 on Twitter compared to 3.15 on Reddit and the PD mean is 3.69 compared to 3.23 on Reddit. This trend is seen on

source	subreddit	EC mean	EC std	PD mean	PD std
Reddit	News	3.12	0.22	3.27	0.26
	Worldnews	3.33	0.18	3.32	0.25
	Politics	3.05	0.17	3.12	0.21
	All	3.15	0.22	3.23	0.25
Twitter	News	3.78	0.36	3.78	0.65
	Worldnews	3.87	0.39	3.9	0.62
	Politics	3.53	0.3	3.43	0.64
	All	3.72	0.37	3.69	0.67
Test set		3.82	1.19	3.96	1.31

Table 4.8: A table with the average EC and PD predicted scores from Reddit comments and tweets per subreddit. The last row includes the mean EC and PD scores with the standard deviation of the BERT balanced model on the test set, for comparability reasons. Reddit rows are in blue color, Twitter rows are in orange and the test set is in green.

all three subreddits as well. The tweets about the articles of r/worldnews are the most emotionally charged (higher EC and PD), as these are the only ones that had average PD scores of 3.9 and EC scores slightly lower 3.9. The same difference of scores between subreddits is also observed on Reddit comments, with r/worldnews on Top, followed by r/news and then by r/politics. This constant difference between the three article sources on both Reddit and Twitter indicates that the discussed articles played a role to the scores.

Significance of differences With the intention of validating these differences as significant, I performed a paired t-test [6] for each subreddit and on the whole dataset. I chose the paired t-test as instances from both sets are connected by the same articles, so there are pairs of predictions that should be examined together and not as independent. The t-test results are showcased in table 4.9 and the difference in every comparison is significant, as the p-values on all comparisons are far lower than 0.05.

subreddit	*EC: t-statistic*	*p-value*	*PD: t-statistic*	*p-value*
News	23.9	$4.1^{-}56$	11.51	$5.1e^{-}23$
Worldnews	14.06	$3.7e^{-}25$	9.8	$2.6e^{-}16$
Politics	18.34	$5.1e^{-}38$	6.32	$3.8e^{-}09$
All	32.26	$2.9e^{-}113$	15.82	$3.9e^{-}44$

Table 4.9: Paired T-test results for comparison between Twitter and Reddit per subreddit and on a holistic level.

How length affected the predictions In the previous chapter during the analysis of the training set, there was a small positive correlation between both the EC

[6] https://en.wikipedia.org/wiki/Student%27s_t-test)

and PD scores with the length of the comment, 0.136 and 0.153 respectively. This is the reason why I examined the length of both tweets and Reddit comments. Interestingly, the tweets might have a lower length than Reddit comments but this did not translate to a higher EC and PD. A plausible explanation is that users of Twitter have adapted to express their feelings with the restricted amount of characters.

EC and PD scores of the titles During the analysis of the created dataset, I decided to exclude from the study the tweets that are very similar to the title. It would be interesting to see how these tweets would have affected the prediction scores, as their population size would inflict bias. Table 4.10 includes these scores for the twelve websites that have eleven or more articles in the dataset. The number of articles does not allow to make concrete comparisons between the websites, but it is clear that the titles for each news story carry higher EC and PD scores than the comments of users. This is logical, as all websites try to draw the attention from the audience to make people visit their page and read the article, which could lead on titles with exaggerated expressions. By looking the websites one by one, the Independent and CNN are the two with the highest level of PD at 4.74. On the scope of EC, Abcnews is the website with the highest score at 4.52 and it is the only news source with a higher score on empathy compared to distress.

News source	*EC*	*PD*
cnn.com	4.23	4.74
nbcnews.com	4.2	4.43
apnews.com	4.24	4.59
independent.co.uk	4.07	4.74
businessinsider.com	3.86	4.47
newsweek.com	3.88	4.4
theguardian.com	4.0	4.31
cbsnews.com	3.9	4.7
reuters.com	4.23	4.69
nytimes.com	4.08	4.49
abcnews.go.com	4.52	4.27

Table 4.10: The EC and PD scores of the tweets that are similar to the titles of the news stories, grouped per news source

4.4 Discussion

Before continuing with the next and final chapter I summarize and discuss the findings of this chapter. The research question examined was "Do the comments reacting to news stories on Twitter express less personal distress and more empathetic concern compared to the comments on the same news articles on popular news subreddits?"

Data set creation The goal of this question was to identify and compare the levels of EC and PD on the comments for the platforms of Reddit and Twitter

about news stories, but a dataset fitting this description was not available. Thus, I proceeded with the creation of one, with comments from both platforms on the same news stories without any annotation. When someone creates a dataset with social data there are ethical matters that need to be discussed. For the acquisition of the data, I used the suggested way to crawl the data (API connection) and I followed the terms and conditions of both websites. But for the comments and tweets that I crawled I did not ask for the consent of each user, as it would be ideal. The comments might be public, but not all users would be comfortable with studies on their comments. For this reason I will try to preserve the anonymity of the users by not releasing sensitive information. More specifically, I will not release the ids of neither the articles or the user comments, so people can not locate the users' profiles. Additionally, I will only give access to the data upon request, if the study meets my expectations. A datasheet is attached on the appendix containing more information about this dataset.

Results For answering the research question I decided to only use the immediate replies from Reddit and only the dissimilar tweets (from the title). I used the BERT balanced model to predict the EC and PD scores of these instances and then averaged the scores per article. These predictions showed that the empathetic concern and personal distress scores of Twitter are significant higher than the scores from Reddit, on both a holistic level and per subreddit. The significance of the results was validated with a paired t-test per subreddit. These results validate only half of my hypothesis. My hypothesis was that users on Twitter express less personal distress and more empathetic concern compared to the comments on the same news articles on popular news subreddits. Based on the results though I conclude that **users on Twitter express more empathetic concern and personal distress than users on Reddit when they are commenting on the same news stories**. In addition, when examining Twitter and Reddit on a subreddit level, I identified that there is a constant difference of scores between the three subreddits on both platforms. The comments of r/worldnews are more emotional charged, followed by r/news and finally r/politics. The most probable explanation is the type of news stories posted in each subreddit. All these results have implications and limitations.

Implications The model that I used for predicting the EC and PD scores on these platforms, was trained to predicted the self-felt EC and PD scores. With this in mind, the results showcase that the users of Twitter have higher levels of EC and PD compared to the users of Reddit, after reading the same news story. A possible explanation about these results is the different characteristics of these platforms, by assuming that the users of these platforms have similar characteristics. To the best of my knowledge, there has not be a study showcasing that there is a significant difference on the empathy levels of users between social media platforms. These results should encourage further research on the effects of social media on our psychology and if these differences stand users should be explicitly informed. Social media play a central role in our lives and even with evidence of negative effects, from depression (Scherr and Brunet, 2017), to social anxiety (Primack et al., 2017), to targeted manipulation for influencing election results (Timberg, 2017), the public is not properly informed and the websites are not properly monitored. This work could be a step in the right direction.

Limitations For this study to be completed there were assumptions made, which bring limitations. Due to the differences of the way Twitter and Reddit function as social media platforms, the dataset consists of significant more instances for Reddit compared to Twitter. The number of tweets per article is less than 10% of the comments on Reddit (Tables 4.2 and 4.5). On Twitter, one has to create queries to crawl the tweets that they want. The queries that I created were strict, as I was aiming for high precision on the crawled tweets. So not all tweets about the articles were collected, which is not the case for Reddit comments.

As previously mentioned, I had to assume that the users have similar characteristics. In fact, I did not track the users and I did not try to identify common users between the two platforms. So these results might be originated by the difference of their population base, but in that case there should be an explanation why people with higher empathetic capabilities join Twitter than Reddit. Since this can not be answered, a future study on the topic should investigate people who use both platforms or similar users.

Finally, there are limitations produced by the differences between the training data and these platforms. In all three cases (training, Reddit, Twitter) there is a difference on the character limit of the comments. Twitter enforces a 280 character limit, the training data had a restriction of 800 characters and Reddit does not have such a threshold. Studying the training data, there was a small positive correlation between EC and PD with the comment's length. But during the prediction, Twitter, which has shorter comments, has a higher EC and PD compared to Reddit. In reality, we can not be sure of the effects of this difference. In addition, Twitter users use a lot of emojis, which were not included in the training set. During the preprocessing phase, this limitation was handled with the transformation of the emojis to their respected words. This transformation allowed me to continue with the experiments, but I do not know the effect it had on the scores.

Chapter 5

Conclusion

5.1 First research question

During my thesis my main research question was to investigate if the Transformer architecture is capable of predicting the empathetic concern (EC) and personal distress (PD) scores of people that react to news articles. Based on the experimental results, it can be concluded that the Transformer architecture is capable of predicting these scores with a higher correlation compared to the baselines. In more detail, almost every Transformer model created during these experiments had an improvement of approximately 10% on each task (empathetic concern and personal distress) compared to the best original model (CNN) created by Buechel et al. (2018). The improvement upon the other two baselines (Ridge regression, FNN) is even greater. The selected model from this question (BERT balanced) achieved a Pearson's r correlation of 0.437 on EC and of 0.483 on PD. These results might appear low, but I argue that they are quite strong due to the sample size and the way the data were constructed. Guo et al. (2020) performed benchmarks on text classification using three pre-trained Transformer-based models on 25 different social media datasets. This task might differ from mine (classification versus regression), but they are both supervised and the sources of data are similar. The performance of the models dropped rapidly depending on the size of the datasets. Interestingly enough out of the 25 datasets none had a lower number of instances than the dataset used here (1860). So, it is evident that the performance of the models in this study is restricted by the low number of instances. In addition, the EC and PD scores were collected with self-evaluation of the users, which also raises the difficulty of the task, as people have different ways to express themselves. Overall, the Transformer architecture, since its release, has become the state of the art for most NLP tasks. There is strong evidence that the prediction of the self-felt EC and PD on social media comments is one more NLP task that Transformers perform better than other architectures.

Data augmentation. In the process of answering the first research question I experimented with data augmentation methods on text. The methods were back-translation (BT), easy data augmentation (EDA) and contextual augmentation (CE). Unfortunately, these methods did not improve the prediction capabilities of the models on the regression tasks, supporting the work of Longpre et al. (2020). They examined pre-trained Transformer models with various data augmentation methods, including EDA and BT, on different tasks. Their conclusions were that

on pre-trained Transformer models these methods do not improve the performance, even on very small datasets. My results are in agreement with their work.

Future studies. As previously mentioned, the most plausible explanation for the low performance of the dataset is the small number of instances. Future search could address this situation with the creation of a larger dataset or with the enlargement of the existing one. Another research direction could be the study of correlation between a model trained using self-felt data, like the data used here, and a model trained on other-annotated data. Sharma et al. (2020), whose dataset was annotated by others, claim that one aspect of their empathy capture framework (EPITOME), highly correlates with the EC used by Buechel et al. (2018). This dataset, not only has a different empathy framework but also the data was gathered from social media platforms dedicated to mental health support, where people are more inclined to give empathetic replies and the discussion topics are different. So, for studying the correlation between the self-felt and the other perceived empathy assumptions are needed to be made for using different frameworks and for the differences of the data. With these observations, I would suggest two ideas. First, the creation of a dataset with comments on news stories annotated by a third party using the empathy theory of Batson (1987) or the EPITOME framework. Second, the annotation of the existing dataset of Buechel et al. (2018) by a third party, which would isolated the difference of the self-felt and other-perceived empathy.

5.2 Second research question

The second research question of my thesis investigated the difference between Twitter and Reddit on the self-felt empathetic concern and personal distress of people, after they have read the same news story. A dataset fitting this description was not available, so for completing this research question I created one. This dataset contains for 406 articles their Reddit comments and tweets. In total the immediate Reddit comments are 218,424 and the tweets are 15,618. After the creation of this dataset I continued with the predictions using the BERT balanced model from the first research question. By analyzing the predictions per article, I observed that the comments on Twitter have significantly higher levels of EC and PD. During these experiments some assumptions needed to be made, limiting the applicability of the results. First, due to the nature of the platforms not all tweets about the articles were crawled, which is not the case for Reddit. Second, there was no study of the people behind the comments so the results could be affected by the population. Third, there were differences between the training samples, the comments of Reddit and tweets. These were the limit of characters, which were 240 characters for tweets, 800 characters for training data (imposed during the creation) and unlimited for Reddit. One additional difference is the usage of emojis. By assuming these did not affect the significant difference between the two platforms, there is a strong indication that people on Twitter have higher levels of EC and PD compared to users on Reddit. To pinpoint the reasons behind this difference further research could be made.

Future studies. Many research opportunities arise from this study. First, the creation of a dataset with the inclusion of more tweets would allow more robust

results. For this to happen more open queries need to be formulated and more time should be allowed to the crawler, because of the time limit from Twitter API. With open queries I mean the usage of bigram or trigram terms from the article that could also return tweets that are not connect to it. With this methodology you lose precision for an increase on recall. Second, more information about the characteristics of the people whose data was captured would also benefit a future study. This would allow for the researchers to observe if there are characteristics of the population that lead to differences or that this is not a factor. Looking at another direction, the EC and PD models could be used to investigate the relationship between other social media platforms on comments about news articles, as Twitter and Reddit are not the only ones used to follow the news. Finally, one could study how different news agencies affect people. For this study, it would be required to use one social media platform to extract comments and search for comments on news articles with the same topic.

Appendix A

Datasheet

Below I answer the datasheet questions created by Gebru et al. (2018). They created this framework with the idea of researchers having a standardized way to document the datasets they create.

Motivation The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

a. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created for my thesis with the title: text-based empathy detection on social media. While the data are not annotated, I predicted the empathy score of these instances with a trained model.

b. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Nikolaos Bentis (nikobent12@gmail.com) for my thesis at Utrecht university.

c. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

None

d. Any other comments?

The date were crawled using Twitter's and Reddit's APIs.

Composition Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

a. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of

instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are tweets and Reddit comments about specific articles. Each text (tweets and Reddit comments) includes the article id to connect id with. Reddit comments include only the replies to the post and not replies to the replies. The language were restricted to English.

b. How many instances are there in total (of each type, if appropriate)?

For 406 articles, I have crawled 218,424 Reddit comments and 15,618 tweets.

c. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

For Reddit comments includes all immediate replies but not all comments for the article. For Twitter it includes a subset of all tweets about the article, we can not know for sure how many tweets are left uncrawled. But my queries on twitter aimed for precision (title or url of the article) so tweets without these information were left uncrawled.

d. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance includes the pre-processed text, the article id and the subreddit that it is about (even for twitter, from which subreddit did the article originated).

e. Is there a label or target associated with each instance? If so, please provide a description.

No.

f. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

user ids, names and any other information provided by the APIs. Except the text.

g. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The article ids connect the Reddit comments and tweets.

h. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No.

i. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Tweets that are similar to the titles of the articles are noise.

j. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The data were crawled from Twitter and Reddit. The data also exist in the archives of these datasets

k. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No.

l. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

It includes people's comments. So the language in some cases is offensive.

m. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

n. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

o. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Yes, someone using Twitter's API can search the exact comment that it is crawled and identify the user that posted it. I decided to not alter the comments for the purposes of my research. That is the reason why the dataset is given upon request

p. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

q. Any other comments?

No.

Collection process The answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

a. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data consists of text that was preprocessed.

b. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

Tweepy and Praw.

c. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

For Twitter due to query construction I only have a subset of all tweets about the article.

d. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

I did it alone.

e. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g. recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

This process was executed daily between the dates of 07/04/2021 and 10/05/2021, except the days were the procedure failed. The data were about these dates.

f. Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

g. Does the dataset relate to people?

Yes.

h. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

I crawled the data from Twitter and Reddit

i. Were the individuals in question notified about the data collection?

No.

j. Did the individuals in question consent to the collection and use of their data?

Yes, by agreeing to the terms of service from both Twitter and Reddit.

k. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

No.

l. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

No.

m. Any other comments?

No.

Preprocessing/cleaning/labeling The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

a. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

I removed empty comments, replaced emojis with their associated words, removed html characters. From tweets removed the reply sign with the username (i.e. @nikobent)

b. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

No.

c. Is the software used to preprocess/clean/label the instances available?

Yes, demoji for emojis on Twitter. Everything else using common libraries from python for text preprocessing.

d. Any other comments?

No

Uses These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

a. Has the dataset been used for any tasks already?

For my thesis with the title: Text-based empathy detection on social media. I used a trained model to predict the empathy score on the instances.

b. Is there a repository that links to any or all papers or systems that use the dataset?

No.

c. What (other) tasks could the dataset be used for?

For any study that intends to compare these two platforms or for studies on the crawled subreddits. As there are no labels or scores for the instances, this dataset can be used mostly for predictions.

d. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks). If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No.

e. Are there tasks for which the dataset should not be used?

Any task that might have negative effects on people. For this reason I will not give access to the data to every request.

f. Any other comments?

No.

Distribution.

a. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**

   I will distribute the dataset upon request.

b. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**

   Contact me at my email: nikobent12@gmail.com or at my github: nikobent

c. **When will the dataset be distributed?**

   After September of 2021.

d. **ill the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**

   No.

e. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

   No.

f. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

   No.

g. **Any other comments?**

   No.

Maintenance These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

a. **Who is supporting/hosting/maintaining the dataset?**

   Nikolaos Bentis.

b. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

   nikobent12@gmail.com

c. **Is there an erratum?**

   No.

d. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**

   No.

e. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

No.

f. Will older versions of the dataset continue to be supported/hosted/maintained?

No.

g. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

There is not one in place, but whoever wants to contribute on the dataset is free to contact me.

h. Any other comments?

No.

Bibliography

M. Abdul-Mageed, A. Buffone, H. Peng, S. Giorgi, J. Eichstaedt, and L. Ungar. Recognizing pathogenic empathy in social media. In *Proceedings of the 11th International Conference on Web and Social Media*, Icwsm, 448-451, 2017. ICWSM 2017.

F. Alam, M. Danieli, and G. Riccardi. Annotating and modeling empathy in spoken conversations. *Computer Speech and Language*, 50:40, 2018. URL `https://doi.org/10.1016/j.csl.2017.12.003`.

N. Alswaidan and M. E. B. Menai. *A survey of state-of-the-art approaches for emotion recognition in text*, volume 62. Springer, London, 2020. URL `https://doi.org/10.1007/s10115-020-01449-0`.

R. L. Barker. *The Social Work Dictionary.* Washington, DC, NASW Press, 2008.

C. D. Batson. Distress and empathy: Two qualitatively distinct, vicarious emotions with different motivational consequences. *Journal of Personality*, 55:1, 1987.

H. Becker. Some forms of sympathy: A phenomenological analysis. *The Journal of Abnormal and Social Psychology*, 26:58–68, 1931. doi: 10.1037/h0072609.

S. Buechel, A. Buffone, B. Slaff, L. Ungar, and J. Sedoc. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2017, 4758-4765, 2018. EMNLP 2018.

D. Cohen and D. A. Cohen. *Empathy in conduct disordered youth.* Differences, 1992.

B. M. Cuff, S. J. Brown, L. Taylor, and D. J. Howat. Empathy: A review of the concept. *Emotion Review*, 8(2):144–153, 2014. ISSN 17540739. doi: 10.1177/1754073914558466.

S. B. Daily, M. T. James, D. Cherry, J. Porter, S. Darnell, J. Isaac, and T. Roy. Affective computing: Historical foundations, current applications, and future trends. 2017.

B. I. Davidson and A. N. Joinson. Shape Shifting Across Social Media. *Social Media and Society*, 7(1), 2021. ISSN 20563051. doi: 10.1177/2056305121990632.

M. H. Davis. Self report measures for love and compassion research: Empathy interpersonal reactivity index (iri). *JSAS Catalog of Selected Documents in Psychology*, 10(85):3, 1980a. URL `http://fetzer.org/sites/default/files/images/stories/pdf/selfmeasures/EMPATHY-InterpersonalReactivityIndex.pdf`.

M. H. Davis. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, 10:85, 1980b.

J. Decety and P. L. Jackson. *The functional architecture of human empathy.*, volume 3. 2004. ISBN 1534582304267. doi: 10.1177/1534582304267187.

J. Decety and C. Lamm. Human empathy through the lens of social neuroscience. *TheScientificWorldJournal*, 6:1146, 2006. URL `https://doi.org/10.1100/tsw.2006.221`.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT*, 1:2019–2019, 2019.

N. Eisenberg and J. E. Strayer. *Cambridge studies in social and emotional development.* Cambridge University Press, Empathy and its development, 1987.

N. Eisenberg, R. A. Fabes, and T. L. Spinrad. Prosocial development. In W. D. Eisenberg and R. M. Lerner, editors, *N*, pages 646–718. Wiley, Handbook of child psychology vol. 3: Social, emotional and personality development . Hoboken, NJ, 2006.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, and K. Crawford. Datasheets for Datasets. 2018. URL `http://arxiv.org/abs/1803.09010`.

P. G. Georgiou, M. P. Black, A. C. Lammert, B. R. Baucom, and S. S. Narayanan. "that's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6974, 2011. URL `https://doi.org/10.1007/978-3-642-24600-5_12`.

K. E. Gerdes, E. A. Segal, and C. A. Lietz. Conceptualising and measuring empathy. *British Journal of Social Work*, 40:7, 2010. URL `https://doi.org/10.1093/bjsw/bcq048`.

J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. Narayanan. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 2015-Janua, 1947-1951, 2015. INTERSPEECH.

Y. Guo, X. Dong, M. A. Al-Garadi, A. Sarker, C. Paris, and D. M. Aliod. Benchmarking of Transformer-Based Pre-Trained Models on Social Media Text Classification Datasets. *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, (February):86–91, 2020.

G. Hein and T. Singer. I feel how you feel but not always: the empathic brain and its modulation. *Current Opinion in Neurobiology*, 18:2, 2008. URL `https://doi.org/10.1016/j.conb.2008.07.012`.

M. Hosseini and C. Caragea. *It Takes Two to Empathize: One to Seek and One to Provide. In: Proceedings of the 35th American Association for Artificial Intelligence (AAAI 2021).* Virtual conference, 2021.

W. Ickes. *Empathic accuracy.* Guilford Press, New York, 1997.

W. Ickes. *Everyday mind reading.* Prometheus, New York, 2003.

W. Ickes, L. Stinson, V. Bissonnette, and S. Garcia. Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal ofPersonality and Social Psychology*, 59:4, 1990.

P. L. Jackson, E. Brunet, A. N. Meltzoff, and J. Decety. Empathy examined through the neural mechanisms involved in imagining how i feel versus how you feel pain. *Neuropsychologia*, 44:752–761, 2006. doi: 10.1016/j.neuropsychologia.2005.07.015.

A. Joshi, P. Bhattacharyya, and M. J. Carman. Automatic sarcasm detection: A survey. acm comput. *Surv.*, 50:5, November 2017. doi: https://doi.org/10.1145/3124420.

H. Khanpour, C. Caragea, and P. Biyani. Identifying empathetic messages in online health communities. *Aclweb.Org*, 246, 2017. URL `https://csn.cancer.org`.

S. Kobayashi. Contextual augmentation: Data augmentation bywords with paradigmatic relations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2:452–457, 2018a. doi: 10.18653/v1/n18-2072.

S. Kobayashi. Contextual augmentation: Data augmentation bywords with paradigmatic relations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2:452–457, 2018b. doi: 10.18653/v1/n18-2072.

B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel, and H. Prendinger. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, 115, March 2018. URL `https://doi.org/10.1016/j.dss.2018.09.002`.

D. Li, Y. Zhang, H. Peng, L. Chen, C. Brockett, M. T. Sun, and B. Dolan. Contextualized perturbation for textual adversarial attack. *ArXiv, Figure*, 1, 2020.

M. Litvak, J. Otterbacher, C. S. Ang, and D. Atkins. Social and linguistic behavior and its correlation to trait empathy. *PEOPLES (Workshop on Computational Modeling of Peoples Opinions, Personality and Emotions in Social Media)*, 128, 2016. URL `http://www.anthology.aclweb.org/W/W16/W16-43.pdf#page=142`.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, 1, 2019.

S. Longpre, Y. Wang, and C. DuBois. How Effective is Task-Agnostic Data Augmentation for Pretrained Transformers? pages 4401–4411, 2020. doi: 10.18653/v1/2020.findings-emnlp.394.

J. D. Mayer, R. D. Roberts, and S. G. Barsade. Human abilities: Emotional intelligence. *Annual Review of Psychology*, 59(1):507–536, 2008. doi: 10.1146/annurev.psych.59.103006.093646. URL `https://doi.org/10.1146/annurev.psych.59.103006.093646`. PMID: 17937602.

A. Mehrabian and N. Epstein. A measure of emotional empathy. *Journal of Personality*, 40:4, 1972.

W. Miller and T. Moyers. Manual for the motivational interviewing skill code (misc). *Albuquerque: Center on , version*, 2(1):50, 2008. URL `http://casaa.unm.edu/download/misc.pdf`.

T. B. Moyers, J. K. Manuel, and D. Ernst. Motivational interviewing treatment integrity coding manual 4.1. *Unpublished Manual, June*, 1, 2014.

S. Mukherjee and P. Bala. Detecting sarcasm in customer tweets: An nlp based approach. *Industrial Management Data Systems*, 117(10):6–2016, 2017.

J. Pennebaker, M. Francis, and R. Booth. Linguistic inquiry and word count (LIWC, 1999.

V. Perez-Rosas, R. Mihalcea, K. Resnicow, S. Singh, and L. An. Understanding and predicting empathic behavior in counseling therapy. *ACL*, 1(1426):2017–55, 2017. URL `https://doi.org/10.18653/v1/P17-1131`.

R. W. Picard. *Affective Computing, first ed.* MIT Press, Cambridge, MA, 1997.

S. D. Preston. A perception-action model for empathy. In T. F. D. Farrow and P. W. R. Woodruff, editors, *Empathy in mental illness*, pages 428–447. Cambridge University Press, Cambridge, 2007. doi: 10.1017/CBO9780511543753.024.

S. D. Preston and F. B. de Waal. Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, 25:1–20, 2002. doi: 10.1017/S0140525X02350015.

B. A. Primack, A. Shensa, C. G. Escobar-Viera, E. L. Barrett, J. E. Sidani, J. B. Colditz, and A. E. James. Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among U.S. young adults. *Computers in Human Behavior*, 69:1–9, apr 2017. ISSN 07475632. doi: 10.1016/j.chb.2016.11.013.

S. Priya. Where should one get news updates: Twitter or reddit. *Online Social Networks and Media*, 9:17–29, 12 2018. doi: 10.1016/j.osnem.2018.11.001.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016. URL `http://arxiv.org/abs/1606.05250`.

H. Rashkin, E. M. Smith, M. Li, and Y. L. Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *ACL*, 5370:2019–57, 2020. URL `https://doi.org/10.18653/v1/p19-1534`.

A. Reyes, P. Rosso, and D. Buscaldi. From humor recognition to irony detection: The figurative language of social media, data knowledge engineering. *Volume*, 74:169–023, 2012. URL https://doi.org/10.1016/j.datak.2012.02.005.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

S. Scherr and A. Brunet. Differential Influences of Depression and Personality Traits on the Use of Facebook. *Social Media and Society*, 3(1), 2017. ISSN 20563051. doi: 10.1177/2056305117698495.

J. Sedoc, S. Buechel, Y. Nachmany, A. Buffone, and L. Ungar. *Learning Word Ratings for Empathy and Distress from Document-Level User Responses*, 1957, 2019. URL http://arxiv.org/abs/1912.01079.

R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. 54th annual meeting of the association for computational linguistics. *ACL*, 1:86, 2016. URL https://doi.org/10.18653/v1/p16-1009.

S. G. Shamay-Tsoori, J. Aharon-Peretz, and D. Perry. Two systems for empathy: A double dissociation between emotional and cognitive empathy ininferior frontal gyrus versus ventromedial prefrontal lesions. *BRAIN*, 132:617–627, 2009. doi: 10.1093/brain/awn279.

A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support, 2020. URL http://arxiv.org/abs/2009.08441.

E. Shearer and J. Gottfried. News use scross docial media platforms 2017. *Pew Research Center*, page 17, 2017. URL https://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/.

T. Singer and C. Lamm. The social neuroscience of empathy. *Annals of the New York Academy of Sciences*, 1156:81–96, 2009. doi: 10.1111/j.1749-6632.2009.04418.x.

M. Strauss, C. Reynolds, S. Hughes, K. Park, G. McDarby, R. Picard, J. Tao, and T. Tan. Affective computing and intelligent interaction. 3784(march). 699, 2005. URL https://doi.org/10.1007/11573548.

A. Suchman, K. Markakis, H. Beckman, and R. A. Frankel. *model of empathic communication in the medical interview*, 277:8, 1997.

J. Tao and T. Tan. Affective computing: A review". affective computing and intelligent interaction. lncs 3784. springer. pp. 981995. 2005.

J. M. Taylor. Computational detection of humor: A dream or a nightmare? the ontological semantics approach. In I. Milan, editor, *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 429–432, 2009.

C. Timberg. Russian propaganda may have been shared hundreds of millions of times, new research says. *The Washington Post*, 2017. doi: https://www.washingtonpost.com/news/the-switch/wp/2017/10/05/russian-propaganda-may-have-been-shared-hundreds-of-millions-of-times-new-research-says/.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

A. Velasquez and H. Rojas. Political Expression on Social Media: The Role of Communication Competence and Expected Outcomes. *Social Media and Society*, 3(1), 2017. ISSN 20563051. doi: 10.1177/2056305117696521.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461, 2018. URL `http://arxiv.org/abs/1804.07461`.

J. Wei and K. Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In E. ijcnlp Conference, editor, *on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the*, pages 2019–2019, 6382-6388, 2020a. Conference.

J. Wei and K. Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In E. ijcnlp Conference, editor, *on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the*, pages 6382–6388. Conference, 2020b.

Z. Wei and W. Gao. Utilizing microblogs for automatic news highlights extraction. *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 277–296, 2017. doi: 10.1142/9789813223615_0019.

A. Welivita and P. Pu. *A Taxonomy of Empathetic Response Intents in Human Social Conversations*, 4899:448, 2020. URL `http://arxiv.org/abs/2012.04080`.

L. Wispe. The distinction between sympathy and empathy: To call forth a concept, a word is needed. *Journal of Personality and Social Psychology*, 50:2, 1986. URL `https://doi.org/10.1037/0022-3514.50.2.314`.

B. Xiao, D. Can, P. G. Georgiou, D. Atkins, and S. S. Narayanan. Analyzing the language of therapist empathy in motivational interview based psychotherapy. 2012.

B. Xiao, Z. E. Imel, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan. "rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS ONE*, 10:12, 2015. URL `https://doi.org/10.1371/journal.pone.0143055`.

A. W. Yu, D. Dohan, M. T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv*, (1):1–16, 2018. ISSN 23318422.

R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. *CoRR*, abs/1808.05326, 2018. URL `http://arxiv.org/abs/1808.05326`.

N. Zhou and D. Jurgens. Condolence and empathy in online communities. *Emnlp*, 2020:609, 2020.