ADAM MICKIEWICZ UNIVERSITY IN POZNAN
FACULTY OF PSYCHOLOGY AND COGNITIVE SCIENCE

Developing Human-centred Models for Improved Recognition of Emotions in Text

Indywidualne aspekty rozpoznawania emocji w automatycznym przetwarzaniu tekstu

Ewelina Gajewska

445841

A Thesis Presented for the Degree of
Master of Arts in Cognitive Science

Thesis supervised by

Barbara Konat, PhD

.....

Poznań, July 2023

Summary

The task of automated text classification involves mapping textual units into a predefined taxonomy of interest. Emergence of novel deep learning algorithms allows for development of automatic detection systems, whose accuracy approaches the human standard. Development of supervised machine learning (ML) models – the most common method employed by computer scientists – requires high quality annotated data. However, some phenomena such as emotion recognition are highly subjective in nature and characterised by systematically achieved low inter-annotator-agreement coefficients in the data annotation process. Recently, a human-centred approach has been proposed in natural language processing (NLP) to account for those natural differences of opinion in development of ML models. In the current thesis, I propose to extend previous methods of modelling individual perspective in regard to perception of subjective phenomena such as emotion recognition. I demonstrate that the personalised approach that implements the proposed Personalisation Metric (PM) outperforms both traditional majority-based accounts as well as alternative solutions to human-centred ML. It achieves not only superior technical performance (measured by F_1 score) but also qualitative outcomes as the results returned by these automatic ML models are tailored to each individual here. The method is validated in two studies – small-scale study and large-scale study. Lastly, I show that the method proposed in the current work could be successfully adopted to mining other subjective tasks such as abusive and offensive language detection. The proposed solution could be employed for development of more accurate classification models that account for the opinions of individuals, including recommendation systems.

Keywords: emotion recognition, human-centred AI, machine learning, natural language processing, recommendation systems

Streszczenie

Automatyczne metody klasyfikacji tekstu pozwalają na przyporządkowanie zbioru jednostek tekstu do z góry określonych kategorii, np. kategorii emocji, na podstawie wyróżnionych przez algorytm cech dystynktywnych tych kategorii. Zaawansowane metody uczenia maszynowego, zaproponowane przez środowisko przetwarzania języka naturalnego, osiągają wyniki bliskie standardom osiąganym przez człowieka. Do wyuczenia takiego algorytmu maszynowego konieczne jest posiadanie zaanotowanego korpusu jednostek tekstu, co najczęściej osiąga się poprzez anotację takiego korpusu przez kilku anotatorów i agregację wyników zasadą większości (ang. *majority-voting*). Jednakże, anotowanie emocji w tekście charakteryzuje się wysoką subiektywnością opinii, gdzie trudno jest uzyskać zgodę wśród grupy anotatorów, co odwzorowane jest w systematycznie osiąganych niskich wynikach zgodności ocen między anotatorami. W ostatnich latach zaproponowano więc zmianę podejścia do tego typu zadań – z podejścia skoncentrowanego na danych (ang. *data-centric*) na podejście dopasowane do natury człowieka (ang. *human-centred*). Wzorując się na wynikach wypracowanych w tymże podejściu, w niniejszej pracy proponuję rozszerzyć istniejące metody modelujące różnorodność opinii w odniesieniu do rozpoznawania emocji w tekście poprzez zaproponowanie metryki kwantyfikującej tę różnorodność opinii w systemach uczenia maszynowego. Z przeprowadzonych badań wynika, że zaproponowana metoda pozwala podnieść wyniki uzyskiwane przez automatyczne klasyfikatory w odniesieniu do rozpoznawania emocji w tekście w porównaniu do zarówno tradycyjnego podejścia skoncentrowanego na danych, jak i alternatywnych sposobów modelowania indywidualnych tendencji klientów. Zaproponowane podejście pozwala więc na wypracowanie bardziej trafnych klasyfikatorów uczenia maszynowego, dodatkowo dostosowanych do indywidualnych tendencji użytkowników, które może znaleźć zastosowanie w automatycznych systemach rekomendacyjnych. Zaproponowana metoda spersonalizowanej klasyfikacji tekstu może także z powodzeniem zostać zaimplementowana do rozpoznawania podobnych zjawisk, np. detekcji mowy nienawiści w tekście.

Słowa kluczowe: przetwarzanie języka naturalnego, rozpoznawanie emocji, spersonalizowane klasyfikatory uczenia maszynowego, systemy rekomendacyjne, uczenie maszynowe

Words of Thanks

I would like to say a special thank you to my supervisor, dr Barbara Konat. The success of my dissertation would not have been possible without the guidance, encouragement and advice you have provided throughout my time as your student. Thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on research and my career has been invaluable.

Contents

1 Introduction

> The limits of my language mean the limits of my world
>
> Ludwig Wittgenstein, *Tractatus Logico Philosophicus: Logical-Philosophical Treatise*, 1922

By categorising a given stimulus one gives meaning to it. This allows one to make sense of the world and, thus, is crucial to cognition. Some stimuli in the environment are more relevant for one's well-being than others, and therefore are caught by the human mind quickly and processed with "high priority". That is because they usually require a rapid response from an individual. Many examples of those "high priority" stimuli are emotional. How a given object will be categorised by an individual is influenced not only by attributes of that object but also by the knowledge and previous experiences of the individual. In other words, both bottom-up and top-down processes take part in this mental activity.

Besides descriptive meaning, language can also convey emotional meanings and therefore induce affective reactions in people. The same word can ignite different categories of emotional experience in different people, similarly to visual stimuli (Wierzba et al., 2015; Saganowski et al., 2022). Affective states can also have an impact on cognitive processes, including processing of words and their meanings. In my study I investigate those individual differences in recognition of emotions expressed in text and its influence on the performance of machine learning models. Therefore, my research fills the gap in knowledge on mining the subjectivity of emotion recognition in text, and provides methods for the implementation of personalised emotion recognition classifiers.

The task of automated text classification involves mapping textual units (paragraphs, documents, sentences) into a predefined taxonomy (of emotions, for example). An automated system needs to learn distinct patterns in data for each of the predefined categories and then predict the correct category for a new instance of text. This is accomplished by comparing the new data point with patterns learned during training and choosing the most similar one. Development of supervised machine learning (ML) models requires high quality annotated data. The common understanding of high quality corpora or the so-called "gold standard", however, implies aggregation of labels obtained in the annotation process. The idea here is that the majority of people cannot be wrong. Therefore, corpora creators make use of repeated labelling – acquisition of multiple labels for some or all data points – and label aggregation techniques, such as majority voting or simple averaging, for obtaining the final corpora.

Although this traditional approach accounts for random errors in the annotation process, it eliminates potentially valuable information regarding the perception of subjective phenomena, such as emotional expression. Thus, in the current work, I argue for the change of paradigm in automated text classification in regard to subjective phenomena. The alternative, individualised, approach accounts for user characteristics that influence the perception

of subjective phenomena by modelling user factors in ML models. Such a personalised approach ensures that the outcomes of automated systems are tailored to each user. Therefore, the proposed approach could improve not only the technical performance of automated natural language processing (NLP) systems, but also the experience of end-users. In addition, it could be achieved with relatively low effort and simple techniques.

Subjective tasks comprise any phenomena where objective validation is not easily applicable and, thus, evaluation is conducted from a subjective perspective. Yet, the dominant approach to the annotation and automated classification of those subjective tasks is based on majority-voted aggregation of labels. Although it reduces the diversity of opinions into more consistent annotation, we do not know whose view is represented in the end and modelled in ML systems. We know it is the majority's view but who comprises the majority?

Manual annotation is conducted either by a few experts in the field, such as psychologists or linguists, or by a larger group of less experienced participants from the general population. However, corpora creators usually do not gather any information about raters and thus cannot confirm their group of raters comprises a representative sample. This is an important aspect because systematic associations between sociodemographic characteristics and annotation outcomes have been found in social sciences as well as NLP research. For instance, Waseem (2016) found significant differences between expert (feminist and anti-racism activists) and amateur annotations on hate speech. In turn, Sap et al. (2022) investigated the relation between socio-demographic factors and the ratings on toxic language. The authors found that more conservative annotators were more likely to rate African American English dialect as toxic. And Jakobsen et al. (2022) demonstrate that raters tend to disagree, not because of the annotation guidelines or task complexity, but different socio-demographic backgrounds of raters (e.g., gender and political leaning for argument annotation in that study).

As a result, corpora creators often report low inter-annotator agreement coefficients because of the subjective nature of the task. If we have five annotators, and two of them choose label a and the other three choose label b (i.e., 2 vs. 3 split), there is not such a clear distinction between a correct and incorrect label as in the case of 4 vs. 1. What is more, Alm (2011) explicitly states that achieving a single "ground truth label" is not possible, nor essential in subjective tasks. She argues against the reduction of variability in annotations and calls for modelling those subjective interpretations instead.

Because in the traditional approach annotated corpora represents the view of the majority on a given issue, one might say some kind of bias is present in the data – a majority bias. Perspectives of minority groups is not included. It makes the problem even bigger if this group is the target of hate speech, for example, and corpora creators are not able to account for this fact. Existing ML models repetitively fail to capture diverse views of Internet users on toxic content (Kumar et al., 2021b). What is more, Kumar et al. (2021b) found that individuals belonging to certain groups characterised by a high risk of being a target of harassment, such as LGBTQ+ people, rate more social media comments as toxic compared to control groups. Thus, ML systems trained on biased datasets could inherit and further propagate those biases. Besides that, the outcomes of machine learning prediction are well suited for "the average Jane", however, disappointing for individual users. Thus, this approach results

in low quality predictions for individual users and poor user experience. So, if the idea of majority-based "gold labels" is flawed, what is the alternative?

Problems with the traditional majority-based approach have caught the attention of researchers, though. Recently, an alternative paradigm to data annotation and subjective NLP tasks was proposed (Röttger et al., 2021). The alternative (referred to as descriptive or perspectivist) paradigm encourages raters to share their subjective views regarding the task and, therefore, allows for studying and modelling those different beliefs in ML systems. The alternative approach thus challenges the traditional one by asking *is disagreement a random noise or a signal?*

An individualised perspective has also been acknowledged by the NLP community (Davani et al., 2022; Miłkowski et al., 2021; Abercrombie et al., 2022). A recent workshop on the perspectivist approach gathered scholars studying a wide range of phenomena, such as detection of emotions, aggression, hate speech and argumentation (Abercrombie et al., 2022). Several works compare traditional and alternative approaches to automatic detection of subjective phenomena. For example, Kocoń et al. (2021) investigate several ML models designed in a traditional and an alternative (perspectivist) fashion. The authors report that the latter approach achieves better results than the former regarding offensive language detection. Superior performance was also achieved by the personalised models in (Miłkowski et al., 2021), where individual perception of emotion intensity elicited by text is measured with a custom metric based on a sample of annotation behaviour of each rater. (Davani et al., 2022) in turn conceptualise the personalised emotion classification as a multi-task problem for a ML system (several outputs are returned from a model, each of which is tailored to a separate user). It also outperforms the traditional majority-based classification.

The alternative approach, then, is easily applicable not only to the data annotation process, but also to the development of ML models. It addresses several limitations of the traditional approach as well. First, it reduces bias in the annotated corpora as multiple categories are allowed to be assigned to the same item. As a result, automated systems account for differences in opinions instead of forcing the majority belief. Second, it does not involve any additional effort from corpora creators in terms of data collection – several works have demonstrated that annotation behaviour of individual raters is sufficient for modelling additional features required in the development of personalised ML models (Miłkowski et al., 2021; Davani et al., 2022). Third, it improves the quality of automated system outcomes as they are tailored to predict an individual's perspective on a given task, not the majority one. Finally, it addresses societal concerns about bias and fairness in Artificial Intelligence (AI) research.

The current work comprises of two studies – classification task of emotion recognition in text. However, prediction is conducted for an individual rater (i.e., the same piece of text could by annotated with different labels by different annotators and a machine learning model needs to learn and predict those different labels depending on who the prediction is employed for) instead of the average person. Therefore, the task seems to reflect to a higher degree real-life applications of machine learning algorithms – in the end they are employed to detect sentiment or emotions for a particular person. A model developed in the standard

majority-based approach in this case returns the same label (e.g., a category of emotion) for each individual, even though perception of emotions differs between people, which is reflected by systematically achieved low inter-annotator agreement coefficients in the data annotation process (Litman & Forbes-Riley, 2004; Zahiri & Choi, 2017; Demszky et al., 2020) as well as research in psychology (Brosch et al., 2010; Smith et al., 2018). Thus, the task employed in the current work could be formulated as follows: recognition of emotions in text on the individual rater level. As such, the proposed personalised approach to text classification constitutes an alternative to the majority-based method. Although the latter approach works well for the majority of people, the proposed alternative acknowledges those individuals that do not have enough representation in the majority perspective to work satisfactorily for them. Thus, the alternative is designed to be inclusive of different perspectives (in regard to the emotion recognition, offensive language detection or any other subjective phenomenon). This kind of approaches seems to be of special importance in regard to hate and offensive language observed in social media platforms due to the limitations of hate keyword-search approach as well as a mix of crowdsource labelling and machine learning techniques (Waseem, 2016; Mondal et al., 2017). The proposed solution to personalised text classification for subjective phenomena is tested in two studies (small-scale and large-scale studies), with the use of different data and raters.

The perspectivist approach to text classification has gained the attention of scientific community. There has been two editions of *Learning with disagreements* shared task (Uma et al., 2021a; Leonardelli et al., 2023) organised by the International Workshop on Semantic Evaluation (SemEval) – "a series of international natural language processing (NLP) research workshops whose mission is to advance the current state of the art in semantic analysis"[1]. However, the objective of these tasks was to predict the one final label (of offensiveness, for example) based on the provided disaggregated annotation, and therefore, information about disagreement among raters.

In my work I demonstrate that including those subjective differences in NLP and AI research may be a beneficial shift of perspective that can lead to the development of better technological solutions (tools and services) which take personal vulnerabilities into account and adjust to them accordingly. Therefore, my study aims to answer the following research questions:

- How could subjective factors of emotion recognition be modelled in algorithms?
- Does inclusion of those factors in deep learning models lead to the improvement of their performance?

Contributions of the present work are threefold:

- I propose a novel solution to emotion recognition in text that follows the recently introduced human-centred approach in NLP
- I provide state-of-the-art model developed for personalised emotion recognition

[1]Definition provided by the organisers on the official SemEval webpage: `https://semeval.github.io/`

- I demonstrate the value of so-called noisy labels (disaggregated data) for training machine learning classifiers

This novel approach to text classification, referred to as human-centred (Kocoń et al., 2021) or perspectivist (Abercrombie et al., 2022) in the NLP literature, runs counter to the dominant data-centred approach in the AI research. Scientific goals of my work go beyond the analysis of disagreements in emotion annotation. My work provides evidence for the change of paradigm supplementing previous work in the area of human-centred (Kocoń et al., 2021) or perspectivist (Abercrombie et al., 2022) approach to text classification. I propose the Personalisation Metric (PM), following related works (Kocoń et al., 2021; Miłkowski et al., 2021), as a measure of the subjectivity of emotion recognition in text that is suitable for modelling this subjectivity of opinions in ML models. It is tested in two studies which prove superiority of the human-centred approach in regard to the classification of subjective phenomena (emotions) in text. The proposed method could be utilised in the development of recommendation systems and more accurate automated sentiment and emotion analysis systems. Furthermore, it could be successfully transferred to other classification tasks such as hate speech detection.

The thesis is structured as follows: in the following section on psychological models of emotions 1.1 I present three prominent theories of emotions. First, two categorical accounts on the nature of emotions are described – a taxonomy of basic emotions proposed by Paul Ekman in section 1.1.1 and a psychoevolutionary theory of primary emotions by Robert Plutchik in section 1.1.2. Second, the dimensional account of emotions represented by James Russell's circumplex model is presented in section 1.1.3. Brief review of emotion research in natural language processing field, including state-of-the-art solutions is gathered in section 1.2. A description of two major approaches to sentiment analysis and emotion recognition in text are offered in sections 1.2.1 on lexicon-based approach and 1.2.2 regarding machine learning methods. Section 1.3 gives overview of the idea of and research conducted in the human-centred approach in NLP. Section three on the methodology of studies conducted as a part of the current thesis is divided into two subsections: section 2.1 describes in detail the proposed Personalisation Metric as a measure of individual opinion on the issue of interest (here, a tendency to recognise emotions), and section 2.2 summarises the architecture of a deep learning model proposed for the purposes of personalised emotion classification in the current work. The next two sections describe two studies conducted as a part of the current thesis, that validate the proposed metric, model and approach to emotion recognition in text. Specifically, section 3 present material, experimental conditions and result of the first, small-scale study. Then, in the same manner the second, large-scale study is described in section 4. Finally, the last section 5 provides a summary of work conducted as a part of the current thesis, its main findings and future work.

1.1 Psychological Models of Emotions

A number of prominent psychologists, physiologists, neuroscientists and philosophers throughout decades proposed different theories regarding the nature of emotions. In 1884 William James stated, contrary to the dominant view on emotions at that time, that the bodily changes follow the perception of a stimulus and the emotion is the feeling of those changes as they occur (James, 1884). Then, in 1980s Klaus Scherer in his component process model defined emotion as an episode of interdependent changes in several systems of an organism in response to the evaluation of an event as relevant to the well-being of the organism (Scherer, 1987). In 1990s Paul Ekman based on his studies of facial expression of emotions in different cultures differentiated a set of basic emotions that were meant to be universal, innate and hardwired, and provided several arguments in support of his theory (Ekman, 1999). At the same time, Andrew Ortony and others argued for a cognitive account on the nature of emotions (Ortony et al., 1990; Ortony & Turner, 1990). Many handbooks in psychology also offer slightly different definitions of emotion, for example in (Lewis et al., 2010, p. 71) authors cite the following concept: "emotions are valenced reactions, or conscious experiences of events with valenced meanings". In turn Frijda (2017) defines emotions as "states of action readiness, and feelings of readiness that bear on the aim of achieving or maintaining, or terminating or decreasing one's relationship to a particular object or event (...)". Nonetheless, other theorists tend to describe emotions as behavioural patterns, states of bodily arousal, or kinds of judgments (Lewis et al., 2010; Frijda, 2017).

What precisely emotion categories are and how many of them exist is the subject of the ongoing debate among psychologists and cognitive scientists (Barrett, 2006a). Some of them distinguish a set of biologically-based basic, primary or modal emotions (Ekman, 1999; Plutchik, 2003; Scherer, 2009) while others, representing the constructionist perspective, claim that there is no universal (and finite) set of emotion categories because they are culturally dependent man-made concepts (Barrett, 2006b). Theories of basic emotions distinguish a set of distinct emotion categories, each of which has a specific role in human psychology. Usually, this set comprises five to eight primary emotions. Emotions such as fear, sadness, happiness and anger are distinguished in most classifications. On the other hand, hate, love, or boredom are recognised rarely. Researchers who represent this approach also emphasise the fact that evolution shaped the current functions and features of emotions. Some acknowledge there might exist non-basic emotions as well, however these are then only combinations of basic categories of emotion (Brosch et al., 2010). Additionally, these theories provide slightly different definitions of what can be regarded as an emotional stimulus. Nonetheless, common ground can also be found between these separate theoretical approaches. In general, scientists in psychology and cognitive science agree that the emotional stimulus represents a special type of stimulus — it is the one of high relevance for the well-being of an individual (Brosch et al., 2010; Frijda, 2017).

Researchers who represent the dimensional account of emotions hold that each emotion can be described by two or three key features (dimensions) instead, i.e. by the level of valence and arousal in a two-dimensional model, and by the level of valence, arousal and dominance in a three-dimensional model (Russell, 1980).

Emotion theorists also point out that specific emotions tend to be elicited by particular kinds of event. Sadness is usually experienced in response to personal loss and anger is elicited by some type of an offense or frustration (Frijda, 2017). Emotional response is regarded as the adaptive reaction to a certain stimulus or event that consists of three parts: the behavioural action, bodily and neurobiological reaction, and subjective feeling. Nonetheless, Frijda (2017) emphasises stimuli are not positive or negative, pleasant or unpleasant themselves. It is the interpretation or appraisal of an individual as well as situational context that gives meaning to a particular stimulus. As a result one object can elicit similar responses in individuals, while the other can induce quite opposite reactions in different people. What is more, the same stimulus or event in different circumstances can induce different (sometimes even opposite) emotions in the same person. Special interest of those aspects is paid within the field of sociology of emotions. This line of research view human activity, including the experience and expressions of emotions, as constrained by her or his location in social structures guided by a particular culture (Lewis et al., 2010).

Researchers in their attempts to give an answer to the chief question *what is the nature of emotion?* study different aspects of emotions. One strain of research focuses on their antecedents, while the others concentrate on their functions, the relation between emotions and cognition, as well as their expressions and neurological bases (Scherer & Ekman, 2014).

The models of emotion are a fundamental part of computational approaches as they provide information on how emotions are expressed in language as well as how one category of emotion differs from the other. The quality of data annotation and ML models therefore depends on those theoretical backgrounds adopted by a researcher.

Although there exists three major traditions of studying the nature of emotions in psychology and cognitive science: basic emotion theories, dimensional accounts, and appraisal theories — the first two are dominant approaches in natural language processing field (Calvo & Mac Kim, 2013; Shuman & Scherer, 2014; Hofmann et al., 2021). Therefore, in what follows I describe in detail Paul Ekman's and Robert Plutchik's theories — which belong to a more general class of categorical models of emotions, as well as James Russell's dimensional account of emotions.

1.1.1 Ekman's Model of Basic Emotions

Paul Ekman formulated his theory of basic emotions based on Darwin's and Tomkins's evolutionary accounts (Ekman, 1992, p. 169-170). The researcher argues evolution played a major role in shaping emotions, including their function, structure, and common as well as specific attributes. Ekman proposes a finite set of discrete emotional categories, although he changed the number of emotions included in his model over the years — starting from 5 through 6 to considering even 13 possible candidates for basic emotions. Importantly, the researcher specifies what he means by calling some emotions "basic" — it is, first, the unique set of attributes for each category of emotion which include the type of antecedent events, behavioural response, pattern of expression and physiological activity, and second, the fundamental function (adaptive value) shaped by evolution in dealing with life tasks, in particular with interpersonal encounters. Moreover, Ekman proposed a list of 9 characteristics that

helps in distinguishing basic emotions from one another and from other affective states such as moods and emotional attitudes (Ekman, 1992). The first one concerns facial expressions which are universal across cultures and at the same time unique for each emotional category. The researcher emphasises here the fundamental role of those expressions as they communicate information to other humans about, for example, the antecedent event that an individual came across and her/his internal states and therefore they take part in the regulation of interpersonal relationships as well. Second is the presence of those emotions in other primates. Third are emotion specific patterns of physiological activity. Fourth item on the list covers common events that precede and evoke each basic emotion. Fifth is the coherence (interrelation) between organismic subsystems in emotional response. Another two characteristics for basic emotions are their quick, almost automatic onset and brief duration counted in seconds rather than minutes and hours. The eighth one concerns the mechanism of appraisal that works automatically and unconsciously. And the last one points to an involuntary aspect of the experience of emotion.

Ekman's classification differentiates at least five emotion families based primarily on his cross-cultural studies of facial expressions and research on autonomic nervous system response — anger, fear, sadness, joy and disgust. In later works the researcher included also surprise and considered adding contempt, shame, guilt, and embarrassment, however he emphasised that the evidence is not as firm as in the case of the first five emotions on the list. That is because facial expressions of surprise were not recognised by members of some cultures and the other ones do not fulfil all the requirements of key characteristics of basic emotions laid down in (Ekman, 1992).

The researcher also gives brief summary of each emotion. Anger arises when an individual is blocked from pursuing a goal or feels treated unfairly. It could be characterised by different levels of intensity starting from annoyance and finishing with fury. The emotion of fear arises with the threat of harm. It helps to mobilise the organism to deal with potential danger. There are several levels of intensity of this emotion as well. The least intense is the feeling of nervousness and the most intense include panic and terror. Sadness is experienced following the loss of someone or something of great importance. This group of emotional states covers mild disappointment as well as extreme sorrow and anguish. Joy is related to pleasurable states such as the feeling of relief or excitement. Disgust (also referred to as contempt) is a feeling of (moral) superiority over another person or a dislike for objects that are aversive to the senses. Here, contempt is believed to be one of the social or moral emotions and disgust is felt rather towards physically repulsive objects. Lastly, surprise is aroused by a sudden and unexpected stimuli such as sounds or movements.

1.1.2 Plutchik's Wheel of Emotions

In the 1980s Robert Plutchik proposed his psychoevolutionary theory of emotions (Plutchik, 1982). He placed emotions in the centre of life of all creatures at different phylogenetic levels and recognised their role in social functioning. Plutchik differentiated eight primary emotions: joy, trust, fear, surprise, sadness, anticipation, anger and disgust, based on eight basic behavioural patterns. In his model, also called *the Wheel of Emotions*, he distinguished different

levels of intensity of those emotions. This is because he observed "that rage might be manifested differently from irritation, or terror differently from apprehension" (Plutchik, 1982, p. 531). What is more, the researcher emphasised that pure emotions are rarely encountered – usually, they are experienced as mixed.

Therefore, his model postulates that besides basic emotions there are complex emotions, as well, called dyads and triads. That is, the author defined rules for forming complex emotions as the combinations of, respectively, two or three basic emotions. There are three types of dyads: primary, secondary and tertiary. Primary dyads are the emotions formed of two adjacent basic emotions. Secondary and tertiary dyads are build from emotions that are one and two steps apart, respectively, in the wheel model. For example, awe is formed as a mixture of fear and surprise (primary dyad), envy is a combination of sadness and anger (secondary dyad), and dominance is formed from the emotions of anger and trust (tertiary dyad) (Jingar & Lindgren, 2019). In addition, Plutchik proposed that emotions can be characterised by the degree of similarity to one another and by polarity, i.e. the idea that each emotion has its opposite. Here, trust is the contrary emotion for disgust, fear is the opposite of anger, surprise is the opposite to anticipation, and sadness opposes joy. As a result, similar emotions are placed next to each other in the Wheel of Emotions and contrary emotions opposite each other in the model. Moreover, Plutchik discovered that primary emotions exist in varying degrees of intensity. Less intensive emotional states include: admiration (trust), terror (fear), amazement (surprise), grief (sadness), loathing (disgust), rage (anger), vigilance (anticipation), and ecstasy (joy). Regarding more intensive states, the model recognises the following emotions: acceptance, apprehension, distraction, pensiveness, boredom, annoyance, interest, and serenity, respectively. Plutchik's assumptions about the nature of emotions and their interrelations were tested and validated in a series of studies conducted by him and his collaborators. His structural model is shown in Figure 1.

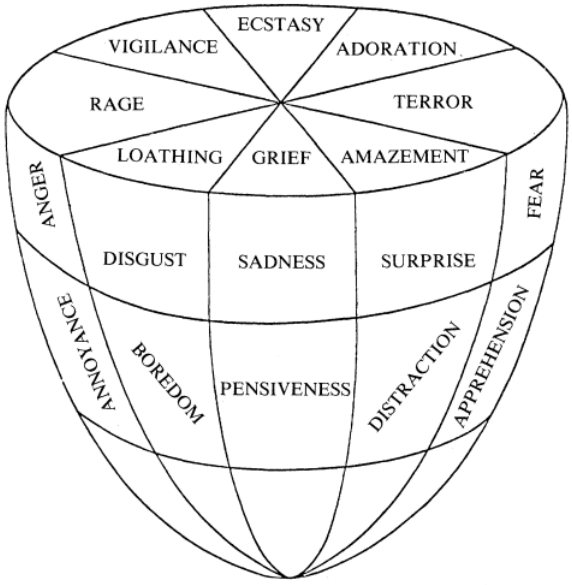

Fig. 1 Plutchik's multi-dimensional model. Adapted from (Plutchik, 1982, p. 539)

Plutchik's model acknowledges previous findings on the nature of emotions in the fields of psychology and physiology. In particular, it refers to works by Charles Darwin, William James, Walter Cannon and Sigmund Freud. Essentially, four traditions in emotion research – the evolutionary, the psychophysiological, the neurological and the dynamic – represented by those researchers were taken into account by Plutchik when developing his own theory.

Plutchik postulated that emotions are complex constructs and one can use different types of languages to describe them – the subjective, the behavioural and the functional. In (Plutchik, 1982, p. 543) he specifies that "the word emotion refers to [the] complex chain of reactions which has an adaptive value for the individual in dealing with various kinds of life crises or survival problems". That is, emotions comprise several components – events, cognitions, feelings, behaviours, and functions. Plutchik underlines that in a day-to-day language people use the term "emotion" to refer solely to the feeling state, however the nature of emotions is much more complex (Plutchik, 2001). The researcher characterises each of the postulated primary emotions by its prototypical stimulus event, cognitive evaluation of the stimulus, the feeling state, manifested behaviour, and effect. The summary of this conceptualisation is presented in Table 1.

Table 1 Conceptualisation of emotions as chains of events by Plutchik (2001, p. 348)

Stimulus Event	Cognition	Feeling State	Behaviour	Effect
gain of valued object	"possess"	joy	retain	gain resources
threat	"danger"	fear	escape	safety
obstacle	"enemy"	anger	attack	destroy obstacle
loss of valued object	"abandonment"	sadness	cry	reattach to lost object
unpalatable object	"poison"	disgust	vomit	eject poison
member of one's group	"friend"	acceptance (trust)	groom	mutual support
new territory	"examine"	expectation (anticipation)	map	knowledge of territory
unexpected event	"what is it?"	surprise	stop	gain time to orient

From the psychoevolutionary perspective adopted by Plutchik, emotions are triggered by stimuli significant for the organism. Those include categories such as prey, enemy and mate. Universality of emotions is observed due to the presence of similar survival-related problems that all organisms face in their environments. As a result, Plutchik proposed there are two main functions of emotions. The first is the communication of one's intentions to others. The second is the increase of an organism's chances of survival.

1.1.3 Russell's Circumplex Model

Dimensional perspective on the nature of emotions is contrary to the theories of discrete emotions presented in previous sections. Instead of treating each emotion category as independent from one another, the dimensional model suggests that different affective dimensions are interrelated (Russell, 1980). Russell assumes that so-called primitives are the building blocks of each emotional process. As a result, his model comprises two bipolar dimensions — valence ranging from pleasure to displeasure, and arousal ranging from activation to deactivation (Russell, 2003). An integral blend of these two dimensions, in turn, form a *core affect*, i.e. "a neurophysiological state that is consciously accessible as a simple, non-reflective feeling" (Russell, 2003, p. 147). In this regard, it is a simple feeling that cannot be reduced to anything else. This subjective experience does not need to be labelled or interpreted in any sense as well. Core affect is also the first primitive in Russell's framework. The second primitive is the perception of the pleasant–unpleasant and activating–deactivating attributes (affective quality) of a stimulus.

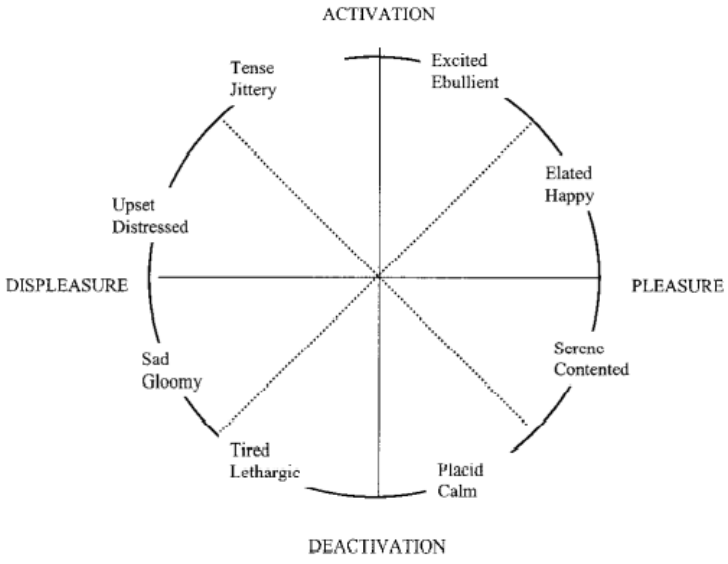

Fig. 2 Russell's dimensional model. Adapted from (Russell, 2003, p. 148)

In this model every feeling can be described as a single point in the two-dimensional space presented in Figure 2. These dimensions are placed on a continuous scale with many levels in between — they go from a neutral level near the centre point to the extreme at the periphery. For instance, arousal "ranges from sleep, then drowsiness, through various stages of alertness to frenetic excitement" (Russell, 2003, p. 148). Russell developed his account based on the results of studies on affective structure of language. This body of research indicated that in most cases participants rated emotional terms on two dimensions — evaluation (pleasure-displeasure) and activity (arousal). Some researchers reported also the third dimension such as potency or dominance (Russell, 1980).

Later, Russell validated his circular model in a series of studies with human subjects. Results confirmed his hypotheses that, first, affective dimensions are bipolar, second, two

orthogonal dimensions can be easily distinguished, i.e. pleasure-displeasure (valence) on the horizontal axis and degree of arousal on the vertical axis, and third, many affective states spread out in the four quadrants of the model (Russell, 1980). As a result, Russell claimed that instead of viewing affect as a set of monopolar and independent factors, new evidence suggests to place affect in a two-dimensional space where individual affective states are in relation to one another.

1.2 Emotion Recognition in Natural Language Processing

A basic assumption of cognitive linguistics is that language and cognition interact. Formulated in cognitive linguistics the Cognitive Commitment states that principles of language should accord with the discoveries about human cognition from other disciplines, in particular from philosophy, psychology, AI and neuroscience (Lakoff, 1990). Human cognitive processes influence the structure of human language and language has an influence on human cognition. Humans have the unique ability to not only experience complex emotions but also communicate this experience to others with the use of language. Individuals' choice of words is regarded as a marker for various psychological phenomena. Emotional tone of language is one of such markers. Language use conveys information about the speakers, the audience, as well as the situational context (Pennebaker et al., 2003). With the rise of computer-mediated research, quantitative approaches to text analysis gain popularity over traditional, manual qualitative studies. Nowadays, statistical analyses on the basis of word counts are commonly employed by social scientist to investigate linguistic manifestations of social phenomena such as endorsement of moral values (Hopp et al., 2021) or cyberbullying (Gitari et al., 2015; Bassignana et al., 2018). People can also express their emotional experiences linguistically, which in turn could be detected and analysed with the use of statistical and machine learning techniques (Pennebaker et al., 2003; Deng & Ren, 2021).

Emotions could be verbalised in three ways: by language expression, nomination and description. Different lexical units are employed to express, denote and describe emotions, respectively. There are literal (e.g., "be annoyed") as well as figurative ways (e.g., "go nuts") of describing one's own emotional states. Expressive language in turn comprises direct phrases (emotional interjections) such as "haha" which reflects joy or positive surprise and "meh" which indicates the feeling of disappointment. In text data, emotional states of the writers could be additionally signalled by the use of punctuation marks such as exclamations or a period in a message-final position (Houghton et al., 2018). On the other hand, emotive language is often used by speakers in persuasive texts in order to appeal to the reader's emotion. It is based on the concept of emotive or emotion-bearing words such as "death" and "wonderful" which have negative and positive connotations, respectively, attached to them. Therefore, by the use of those emotion-bearing words a speaker could influence the emotional state of the reader.

Research on emotion detection in the Natural Language Processing (NLP) field focuses on data with expressive and emotive language use. Therefore, the task usually comprises recognition of emotion expressed by the author in a given text. Although, computational approaches focus on detection of the emotion-bearing words instead of emotional interjections

as useful features for the task. Novel techniques such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) allow to detect emotions implicitly expressed in a text by taking into account the meaning of a whole sentence, instead of focusing on the recognition of single (emotive) words.

Sentiment analysis refers to "the process that can identify and categorise opinions expressed in a piece of text, especially in determining whether the writer's attitude towards a particular topic is positive, negative, or neutral" (Rodriguez et al., 2019, p. 170)[2]. Opinion mining in textual data has become a topic of interest in recent years, both for academia and industry (Tsapatsoulis & Djouvas, 2019). Dynamic growth in the volume of research comes from the widespread availability of big data (e.g., reviews from online shopping websites and posts on social media platforms) as well as introduction of novel deep learning architectures. As a result, a new survey or review of state-of-the-art techniques in the field is published almost every year (Marrero-Fernández et al., 2014; Seyeditabari et al., 2018; Al-Saqqa et al., 2018; Alswaidan & Menai, 2020; Deng & Ren, 2021).

Different technologies for sentiment analysis are being developed by private companies, for example, to investigate customers' reviews about their products in order to improve the customer experience or survey public opinions about candidates in political campaigns based on activity and comments posted on social media sites. Analysis of engagement between political leaders and voters is an additional factor utilised in some studies for mining emotions and stance towards candidates in an attempt to predict the outcomes of elections (Anstead & O'Loughlin, 2015; Gallagher et al., 2019; Chauhan et al., 2021). Ren & Nickerson (2014) discovered that reviews with emotion-expressing language can influence not only shopping experiences but also customer behaviour and sales for some products. Sentiment analysis became an important part of systems designed for investigating public attitudes towards controversial topics such as refugees in Europe (Chen et al., 2022b) and vaccination (Aljedaani et al., 2022). Features related to emotions expressed in text proved to be effective in hate speech detection tools (Rodriguez et al., 2019). Sentiment mining is also useful for a financial sector, for example, in stock market forecasting or prediction of price for popular cryptocurrencies using financial news from Twitter and headlines of related news articles (Pimprikar et al., 2017; Mohapatra et al., 2019; Liapis et al., 2021). Thus, it is becoming an important part of methodology employed in studies regarding social phenomena as an alternative or an auxiliary tool to traditional surveys in social sciences. Sentiment mining applied in early warning systems on social media was effective in the prediction of disease outbreak and tracking a disease spreading (Diaz-Aviles et al., 2012). It is becoming a crucial part of human-computer interaction systems such as dialog chatbots or social robots as well. Incorporating emotion understanding modules in such technologies could make the interaction more human-like and thus improving user experience.

Analyses of sentiment and emotions in text can be conducted at different levels — ranging from the most general document-level through sentence-level to the fine-grained phrase- and

[2]Thereafter I will use "sentiment analysis", "emotion analysis" and "opinion mining" interchangeably as they refer to similar concepts in the NLP field, though emotion analysis is recognised as a more complex task with a higher number of categories a model has to predict (Rodriguez et al., 2019).

aspect-level analysis (Liu & Zhang, 2012). The aim of document-level sentiment analysis is to assigned a single label to the whole text content, for example product review. Here, one must assume that a review expresses opinion about a single subject and not comparison of two products or entities. Sentence-level sentiment or emotions detection splits a given text into sentences and considers each of them separately in the classification. Finally, the aspect-level analysis is based on the idea that each subjective text (opinion) consists of sentiment (or emotion) and its target, therefore one needs to detect a target entity first, before recognition of sentiment or emotion in the text. It is the most detailed, although the most challenging, level of analysis as it allows to discover both negative and positive features of a product or service in one customer review, among others (Chachra et al., 2017).

Simple sentiment analysis comes down to classifying a piece of text to either a *positive* or *negative* category. Sometimes a third — *neutral* — label is included as well, which indicates that a given unit of text is neither positive nor negative. On top of that, some models/tools differentiate between 5 categories of polarity scores — from labels indicating strongly negative connotations through weakly negative, neutral, weakly positive to strongly positive (Wilson et al., 2004). Therefore, these tools are able to detect the strength of sentiment in text as well. Some authors incorporate also *ambiguous* class into a set of sentiment categories, which signal that a given word, sentence or document conveys some positive emotions as well as some negative ones (Kocoń et al., 2018; Addi et al., 2020; López-Chau et al., 2020). Wang et al. (2018) acknowledge that the field focuses on on the binary or ternary sentiment classification, that is allows prediction of only one category for each text. Therefore, there is a large space of exploration for machine learning classifiers that account not only for the prediction of several categories of emotions but also multi-emotion detection in a single text.

In addition to simple sentiment analysis, there is also emotion recognition in the field of natural language processing (NLP), which detects the presence and intensity of separate categories of emotional states in human language. In the psychological tradition there are two major approaches for studying the nature of emotions which are employed by researchers in the NLP field as well. In the first one, a dimensional model, each emotion has two or more dimensions and can be represented as a point in this multidimensional space. Here, the most fundamental and often used dimensions include valence, which ranges from negative to positive, and arousal (level of activation) which ranges from low to high. Three-dimensional representations comprise an additional dominance axis which refers to the apparent power (dominance) of a person (Grimm et al., 2007). Grimm et al. (2007) claim that this third dimension may be crucial to distinguish some emotions — for example, anger and fear. The second approach to studying emotions is the categorical account, in which scientists differentiate a varied number of discrete emotional states. Here, among many, the most popular and widely used are two classifications, described in the previous section: Robert Plutchik's (Plutchik, 2003) theory of 8 primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy (happiness), and Paul Ekman's model which distinguishes 6 categories of basic emotions: anger, fear, surprise, sadness, happiness and disgust (Ekman, 1999). Natural language processing techniques implement either a dimensional or a categorical model.

Classification task with multiple categories is always more challenging than binary predic-

tion of positive vs. negative sentiment (Bouazizi & Ohtsuki, 2019). The authors report that for a binary classifier accuracy of results could achieve over 80%, while the same task applied for the classification of seven different categories is characterised by the drop in performance to 60%. Results indicate that by adding one class to the labels set one can observe a 5% drop on average in terms of accuracy.

Mohammad (2016) enumerates several general limitations and challenges associated with automated sentiment analysis of texts, including the following:

1. sentiment of a whole sentence may be different than the value obtained just by simple summation of sentiment values of individual words

2. some terms in a given text might be associated with negations which could have an influence on the sentiment score of the whole text

3. words in different contexts can mean different things — word-meanings should be taken into account rather than single words if one wants to approximate sentiment score of a given text more accurate

4. presence of figurative language like sarcasm, irony and metaphor — it is difficult for machines to interpret it correctly

5. individual and cross-cultural differences in emotion perception and recognition

Deng & Ren (2021) add to the list the shortage of large-scale high-quality corpora, which is recognised also by others. Although many resources are publicly available, often different taxonomy of emotions is used for the annotation by different authors or the obtained distribution of annotated categories is highly imbalanced. Also, the quality of annotation could have an impact on the final results of automatic methods. As stated in Bhowmick et al. (2008, p. 58): "the accuracy of a supervised machine learning task primarily depends on the annotation quality of the data, that is used for training and cross validation (...). Inconsistency or noisy annotation may lead to the degradation of performances of supervised learning algorithms". Therefore, the usability of an annotated dataset (corpus) depends on the reliability of annotation, which is measured by various coefficients of agreement. Here, the most popular measures include Cohen's kappa (Cohen, 1960), Fleiss' kappa (Fleiss, 1971) and Krippendorff's alpha (Krippendorff, 2018). Simple percentage agreement (observed agreement) is often reported is addition to these coefficients, as well.

Repeated labelling — acquisition of multiple labels for some or all data points — is a popular method used by researchers in the NLP field for data annotation. It is a beneficial technique, in particular, in cases when annotators come from the general population (non-experts) and high quality of annotations cannot be guaranteed (Sheng et al., 2008). Although repeated labelling can introduce a certain amount of noisy labels due to, for example, lack of expertise or interest, Sheng et al. (2008) showed that there are many good reasons to employ it. It is also cheaply and easily available nowadays through online platforms such as Amazon's Mechanical Turk[3]. Following assumptions behind the central limit theorem,

[3] `https://www.mturk.com/`

corpora developers aim to uncover "ground truth" labels for text samples. In the context of the current thesis, for example, they would assume that a piece of text expresses one category of emotion (or is neutral) given a pre-defined set of emotions. By collecting annotations from several raters, it is assumed that majority of them would be able to indicate what these "ground truth" labels are.

Coefficients of agreement are then employed to determine the degree to which the annotation is reproducible, i.e. reliable. Observed agreement is the simplest method of quantifying such reliability. However, as (Cohen, 1960) points out, it is reasonable to expect a certain amount of agreement by chance. Intuitively, particularly in cases of imbalanced distribution of categories, it is relatively easier to obtain agreement for the dominant (most frequent) category compared to the scattered ones. The researcher proposes to account for this fact by finding the joint probabilities of the marginals given a matrix of agreement. The kappa coefficient then is "the proportion of agreement after chance agreement is removed from consideration" (Cohen, 1960, p. 40). By taking joint probabilities of the marginals, Cohen's kappa accounts for individual annotation tendencies of raters. For example, one rater may "overannotate" one category, therefore, the agreement for this category with the second rater would be higher compared to the other labels. Cohen's kappa is the established coefficient of agreement reported in studies in corpus linguistics. However, it is designed to computed agreement between only two raters. Fleiss' kappa is employed for reporting agreement between multiple annotators. Krippendorff's alpha is designed for multi-rater agreement calculations as well, however, it allows in addition to collect variable number of annotations for individual instances of text.

Next, the majority voting technique is generally employed in order to get a single label that (best) represents a given instance from a multi-label set. However, some useful information could be lost by the implementation of this technique such as the information about the minority class (or minority perspectives on a certain topic) and the degree of certainty of the majority class. Thus, the majority voting can induce bias in data. One of the possible strategies to reduce it is to use some form of soft (probabilistic) labels instead of hard labels or adopt multi-label annotation (Sheng, 2011).

Although majority voting is one of the most popular and simplest method to deal with disagreements in annotation, there are also a number of other techniques that estimate ground truth (so-called silver) labels. One of them is the Dawid-Skene (Dawid & Skene, 1979) probabilistic algorithm. Thus, although majority voting works Sheng et al. (2008), researchers investigate whether new methods that consider disagreements in annotation and improve performance of the model could be designed. Yet another way to obtain aggregated labels is development of a voting classifier which makes a prediction over the outcome of other models trained on data provided by individual annotators. Nevertheless, the simplest strategy to obtain high quality data is the removal of instances with substantial or any disagreement between annotators (Uma et al., 2021b).

Crucial to note is also the fact that not all disagreements are equal — in other word, there are different sources of label noise (Frénay & Verleysen, 2013). Some may count as random noise, while others may reflect different points of view on a given subject. Sometimes the

latter type of disagreement is referred to as annotator's bias, however, it can also indicate a high level of subjectivity of a certain task.

Because of those limitations, researchers develop other methods for labelled data collection. Instead of relying on manual annotation, labels for emotions can be obtain with the use of method called distant supervision, directly from text data. Here, one can search for specific keywords known to be related to emotions, hashtags mentioning affective states of the authors (emotion-carrying hashtags) or emojis used in text. In turn, sentiment of product reviews via distant supervision is discovered by gathering number of stars accompanying the text of review given by the customer. Researchers have to only cluster data into different categories of sentiment depending on the number of starts given. Usually, given a five stars scale, one and two stars reviews are believed to express negative sentiment, three stars reflect neutral sentiment, and reviews with four and five stars are classified as positive. Although, data gathered via this method is believed to be "noisy", on the other hand, part of labels (i.e., those correctly identified) reflect the "ground truth" to a higher degree compared to crowdsourced annotation because here the author marks a piece of text as expressing a given emotion or intention.

Although, popular approaches to emotion detection in NLP are based on explicit expression of emotion using emotion bearing words, implicit manifestations of emotions in language comprise a considerable number of cases. What is more, studies in psychology show that context is a crucial component in understanding emotions (Oatley et al., 2006). This makes the task of automated emotion detection even more challenging.

There are two dominant approaches for detection of sentiment or emotions in text: the first is based on affective lexicons and the second employs machine learning techniques (Dey et al., 2018). Following lexicon-based approach does not require to develop a custom classifier - those lexicons with emotive words are usually off-the-shelf, freely available and ready to use. On the other hand, it could be a major drawback and one of the limitations of this approach as words are often evaluated based on their general, domain-independent meaning. However, many words change their meaning and the power to induce emotions depending on the context of use. This in turn is a major advantage of machine learning approach — one could develop an algorithm to recognise emotions in a narrow domain with a very specific language. On the one hand, those custom models can account for this uniqueness of domain language, but on the other hand they are often not reusable to analyse emotions in other corpora. In what follows I characterise these two approaches in detail.

1.2.1 Lexicon-Based Approach

Lexicon-based methods leverages word-lists that are annotated with sentiment and/or emotions. Those lists comprise a selected set of terms with assigned categories or intensities of emotion they signal or evoke (Zhou et al., 2016). Affective lexicons are used in rule-based systems in which emotion detection often comes down to searching for and counting those emotion bearing words or polarity scores. It is a simple aggregate-and-average method where individual scores for words in a document are added together and divided by the total

number of words in that document (Taboada et al., 2011). One of the most popular tool that implements the lexicon-based approach is VADER (Hutto & Gilbert, 2014).

Usually, affective dictionaries are created manually. Researchers invite a group of participants to evaluate selected phrases in terms of emotions they express or induce (as in Mohammad & Turney (2013), for example). However, there are some limitations of this method — manual annotation of large sets of words are time-consuming and expensive. Therefore, they are often limited in size and contain only a few thousand words. Instead of employing annotators from the general population, sometimes ratings are given by a small group of experts, i.e. psychologists and/or linguists. Notwithstanding that this method is more efficient in terms of time and money, a small number of participants is not representative for a whole community. This point is particularly important for such a subjective task as emotion recognition where individual differences play a major role. In order to get larger dictionaries with emotion ratings researchers sometimes employ a semi-manual method in which they start with a smaller, manually annotated set of terms and automatically obtain (approximate) emotion values for other semantically related or similar words (Mohammad et al., 2009; Taboada et al., 2011). Nonetheless, regarding non-English languages there is a scarcity of resources as numbers of available affective lexicons as well as their sizes are often much smaller than the ones for English language. For example, the Nencki Affective Word List contains 2,902 Polish words (Wierzba et al., 2015), French affective norms corpus (Monnier & Syssau, 2014) have 1,034 words and Berlin Affective Word List Reloaded (Vo et al., 2009) consists of 2,900 German words. Summary of affective dictionaries available for English is presented in Table 2.

Lexicon-based approach follows several assumptions. First, the overall sentiment of a text depends on the words that explicitly express (subjective) opinions. Second, sentiment of a document depends on the majority category (developers rarely allow their models to assign more than one label to a text). Third, sometimes referred to as a *prior polarity* assumption, states that polarity of a word is independent of context; a given term is either positive or negative in itself (Taboada et al., 2011). Unsurprisingly, best results for sentiment analysis of natural language are obtained with lexicons adapted to a specific domain. At the same time, however, it is a major shortcoming of the lexicon-based approach — domain and context specificity. As a result, the range of applications of those types of dictionaries is more narrow than the ones that cover multiple topics (Almatarneh & Gamallo, 2017; Dey et al., 2018).

First affective lexicons comprised almost exclusively adjectives as indicators of sentiments in a document. Nowadays, those dictionaries contain adjectives, adverbs, nouns, verbs as well as common idioms or phrases (bi-grams or tri-grams). The most popular are sentiment dictionaries encompassing annotations only on one scale — polarity — with binary (positive, negative) or ternary (positive, neutral, negative) labels (Taboada et al., 2011; Liu & Zhang, 2012).

Table 2 Affective lexicons for English language

Lexicon	Method	Authors	Description
SO-CAL	Semi-manual	Taboada et al. (2011)	over 5,000 words with polarity categories
SentiWords	Semi-manual	Gatti et al. (2015)	over 16,000 words with polarity scores (between -1 and 1)
General Inquirer	Manual	Stone et al. (1966)	11,00 words classified into over 180 categories; including 1,915 positive and 2,291 negative words
WordNetAffect	Semi-manual	Strapparava et al. (2004)	4,787 words classified as positive, negative, neutral or ambiguous, and Ekman's 6 basic emotion categories (fear, anger, sadness, disgust, joy, surprise)
ANEW extended	Manual	Warriner et al. (2013)	13,915 English lemmas annotated on 1-9 point scale on valence, arousal and dominance
LIWC	Manual	Pennebaker et al. (2001)	ca. 4,500 words assigned to 76 categories; 905 words classified as positive or negative
SentiWordNet 3.0	Semi-manual	Baccianella et al. (2010)	currently 147,306 synsets with assigned scores of positivity, negativity and neutrality
SenticNet	Semi-manual	Cambria et al. (2010)	14,244 concepts with polarity values
VADER	Manual	Hutto & Gilbert (2014)	7,500 words with polarity scores
Opinion lexicon	Semi-manual	Hu & Liu (2004)	6,789 words with binary polarity classes
NRC Emotion Lexicon (EmoLex)	Manual	Mohammad & Turney (2013, 2010)	14,182 words with assigned polarity (negative, positive) and Plutchik's 8 primary emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust)
NRC-VAD	Manual	Mohammad (2018a)	more than 20,000 words with valence, arousal and dominance scores
NRC Emotion Intensity Lexicon (NRC-EIL)	Manual	Mohammad (2018b)	ca. 10,000 words with real-valued scores for Plutchik's 8 primary emotions

In order to detect emotions with this method, researchers need to employ a number of text normalisation methods including conversion to lower case and stemming or lemmatisation. This is because, one needs to match the character of words in a document with those from a chosen affective dictionary in order to analyse them. Here, a researcher usually relies on a manually crafted sequence of rules for word spotting and information retrieval. Nevertheless, tools that employ lexicon-based methods are characterised by low accuracy of results, which is additionally limited by the quality of affective lexicons (for review see Cambria et al. (2013). Here, lexicon-based methods rely on word list comprised of word forms instead of word-meanings, therefore term ambiguity is not always resolved in these approaches. The Pollyanna effect (positivity bias) observed in affective evaluation of language (Boucher & Osgood, 1969) might further decrease the accuracy of lexicon-based sentiment analysis (Taboada et al., 2011).

Lexicon-based approach was adopted in Facebook's Gross National Happiness index to measure the "well-being" or "quality of life" based on the status updates of millions of Facebook users. Kramer (2010) operationalised "gross national happiness" as a standardised difference between the use of positive and negative words from the LIWC corpus. The author found that peaks of happiness indeed co-occurs with national and cultural holidays such as Christmas. That is, people use more positive words in their status updates on those days. Moreover, positivity index correlates significantly with life satisfaction in that study. Nonetheless, the study conducted by Wang et al. (2014) did not validate those findings. Cody et al. (2015) studied sentiment of climate change discussions on Twitter based on affective lexicons. Time span following Hurricane Sandy landfall turned out to be more negative than other days of the studied period.

Although this approach possesses various challenges, such as negation handling and context-dependent meaning, on the other hand it is also characterised by a simplicity of use and explainability of results. Numerous techniques are currently being developed towards improving the lexicon-based method for sentiment analysis. For example, Abdulla et al. (2014) compared two approaches to sentiment analysis, corpus-based and lexicon-based, for Arabic language and concluded that, in the case of the latter method, size of a collected sentiment dictionary has a significant impact on accuracy of polarity labels prediction.

1.2.2 Machine Learning Approach

On the other hand, there are machine learning approaches that rely on (annotated) pieces of text that comprise training data for machine learning algorithms. The aim of training these models is to automatically extract emotions from a collection of texts. Emotion detection from text is defined here as the classification problem. Usually, supervised machine learning methods are employed for this task, although some researchers made use of unsupervised techniques as well. The former requires to collect a text corpus where each example is labelled with some category. For the latter method, annotation is unnecessary as algorithms learn to detect distinct patterns in data on its own and cluster observations into a set of classes. Recently, there is a growth of interest in semi-supervised methods which utilise both unlabelled and labelled examples to learn a classification function. They are employed when

very few labelled and a large amount of unlabelled data instances are available in particular (Sintsova et al., 2014). Subset of algorithms from this category are known as self-training — first, a classifier is trained on annotated examples, then it is employed to assign pseudo-labels to unannotated data. Data points labelled with high confidence by the classifier are leveraged for further training.

Best performance is systematically achieved by algorithms comprising convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in particular bidirectional long short-term memory (BiLSTM) architectures (Kim, 2014; Zhou et al., 2015; Hameed & Garcia-Zapirain, 2020; Tam et al., 2021). However, recently introduced Transformer-based language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are currently considered state-of-the-art (Cortiz, 2021). Nevertheless, support-vector machines as well as logistic regression are still eagerly used classifiers (Dini & Bittar, 2016; Bharti et al., 2022).

The process of ML model development comprises of several interdependent steps. First stage involves data collection. The end domain of emotion classifier determines the choice of source, type and volume of text samples. Also, certain characteristics of language such as its formality should match the end purpose of the task. Many studies report selection of data based on this criterion as communication style on social networking sites differs from the one used in political speeches or financial talks. Other factors such as context availability (e.g., previous turns in dialogue) and text length (short texts vs. long documents) are taken into account in design of emotion recognition pipeline as well. Wankhade et al. (2022) enumerates four essential sources of data:

1. social networking sites: individuals can interact, share and exchange opinions with other users, brands, or friends;

2. forums: place of discussion about various topics; users can express their stance towards an issue and argue with others in support of or against their arguments

3. blog: personal websites, where authors convey their viewpoint or share knowledge in the domain of expertise, might also function as a personal diary;

4. commerce websites: online shopping websites with dedicated section of customer reviews about business products

Preprocessing procedure is a common practice in NLP applied to a collected corpora before model training. Common preprocessing steps include conversion to lowercase, removal of so-called stop-words (function words) such as conjunctions, determiners and prepositions, stemming or lemmatisation, as well as hyperlinks or hashtags normalisation (Jain et al., 2017). However, in machine learning approach to sentiment and emotion analysis preprocessing of textual input could influence the quality of developed models and classification results (Bao et al., 2014).

Next stage focuses on feature selection and feature extraction. It involves extraction of valuable features directly from the collected samples as well as data transformation – feature engineering based on the domain knowledge and previous findings of related works. Regarding textual data the most straightforward features include linguistic features – word uni-grams,

word bi-grams and/or character n-grams. Special attention is sometimes dedicated to features related to the use of punctuation marks, emojis, hashtags and external links which are treated as pragmatic features that represent communicative and social context of language use and language understanding (Wankhade et al., 2022).

Before feeding into a model, textual content needs to be transformed into a numeric representation. There are several methods divided into more traditional ones based on occurrences and/or frequencies of words in a given document and the more advanced ones which employ word embeddings (vector representations in low dimensional space). The former encompasses TF-IDF (Term Frequency-Inverse Document Frequency) as well as bag-of-words methods. The latter includes Word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2016) models. "Embeddings are distributed representations of words in a vector space, capturing syntactic and semantic regularities among the words" (Hovy, 2015, p. 754). The simplest bag-of-words method involves creating a vector with binary values — based on the dictionary created from our collection of documents the occurrence of a word from this dictionary in a document is marked by 1, and 0 otherwise. However, obtained matrix is sparse and characterised by high dimensionality, which are the major shortcomings of this method. Moreover, information about word order in a given document is lost, thus making it unable to capture complex linguistic features. Therefore, bag-of-words representations of text are being replaced by (fixed) word embeddings or more recently by contextualised embeddings such as the Embeddings from Language Models (ELMo) (Peters et al., 2018) or Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) that represent rich semantic/syntactic information. Therefore, this word representation method could overcome shortcomings of bag-of-words approaches. Recently, some researchers proposed so-called sentiment-aware word embeddings that incorporate additional information about emotions tied to lexical units into vector spaces (Tang et al., 2015; Khosla et al., 2018; Mao et al., 2019). This method combines traditional word embeddings (that encompass statistical information about words) with prior (sentiment) knowledge from external sources. Authors of those works report additional improvements of emotion classification results over traditional text representation techniques.

Results of the SemEval-2019 competition in task 3: contextual emotion detection in text indicate that methods based on transfer-learning are particularly advantageous for this kind of task (Chatterjee et al., 2019). Teams that achieved superior performance in the task make use of pre-trained language models such as BERT. For example, ANA team make use of an ensemble model of fine-tuned BERT and hierarchical LSTMs and three types of embeddings for context representation: GloVe, ELMo and DeepMoji, next to dialogue embeddings encoded by LSTM. BERT-based architectures were eagerly used also by the participants of EmotionX 2019 Challenge (Shmueli & Ku, 2019). Unsurprisingly, top-performing teams made use of BERT in their model configuration. For example, KU team first post-trained BERT model on a corpus of data selected for the task – dialogues of *Friends* via Masked Language Model and Next Sentence Prediction. And second, the team employed a max pooling technique on generated token embeddings, which was then fed into a dense classification network. In

turn, two other teams made use of additional CNN and RNN networks on top of the BERT-generated word embeddings.

Superiority of this Transformer-based architecture comes from its ability to create contextualised sequence embeddings, that is account for the context dependent meaning of words to a certain degree. The extraction of context showed substantial improvements in many NLP tasks so far, including textual emotion recognition. This could not be achieved with simple methods such as bag-of-words representations. Therefore, Transformer-based models set the new state-of-the-art for emotion detection in natural language since its introduction in 2018 (Acheampong et al., 2021).

Positional encoders transform textual inputs into contextualised embeddings by making use of 3 types of information about the input sequence: position of words in a sequence, specific token embeddings learned during pre-training, and type of a segment pair. The building block of Transformer architecture is a series of scaled dot-product attention units. This mechanism allows to assign different weights (and therefore importance/relevance) to different parts of input sequences. It thus overcomes the so-called vanishing gradient problem of long-range dependencies in input sequences. Moreover, each layer in this architecture has multiple attention heads and therefore could represent different kinds of importance between a word and every other word in a sequence. Because BERT was pre-trained on two tasks – masked language modelling and next sentence prediction – it possesses some statistical knowledge about language. This knowledge is utilised in downstream tasks such as text classification and question answering via transfer learning. Usually, this BERT fine-tuning process comes down to adding just one classification layer on top of the pre-trained model.

Nonetheless, emotion recognition based on language models such as BERT also possesses several limitations. Although, the design of those models allow to take into account a sequence of text (a pair of sequence separated by a so-called special token [SEP]), it is limited by a number of tokens in the input — 512 tokens for BERT-based models, for example. Accounting for longer spans of text could be useful for modelling context of an utterance and therefore improve detection of emotions in text. Recently, there were several successful attempts to follow such a methodology with the improved performance of classification results (Lee & Lee, 2022).

However, emotion recognition in very short texts such as social media posts (e.g., tweets) is challenging even for Transformer-based architectures as they do not contain much contextual information and assume some implicit knowledge (Maynard et al., 2012; Kateb & Kalita, 2015). Therefore, researchers proposed to make use of additional features here, such as text statistics (period use, text length, hashtag position, etc.) in order to feed machine learning models with more information about the data and obtain more accurate results.

Recent review shows that even state-of-the-art techniques for text processing fall short for classification tasks such as emotion detection (Kocoń et al., 2023). The Chat Generative Pre-trained Transformer (ChatGPT) model performed substantially worse than the systems currently considered state-of-the-art – 56 vs 76% correct predictions on average, respectively. Moreover, the model underperformed on emotion detection compared to other tasks (such as spam and aggression detection), including the personalised version of the task designed by

Kocoń et al. (2023) that in theory adjust the outcomes to the interlocutor Summary of the machine learning approach and its techniques is presented in Figure 3.

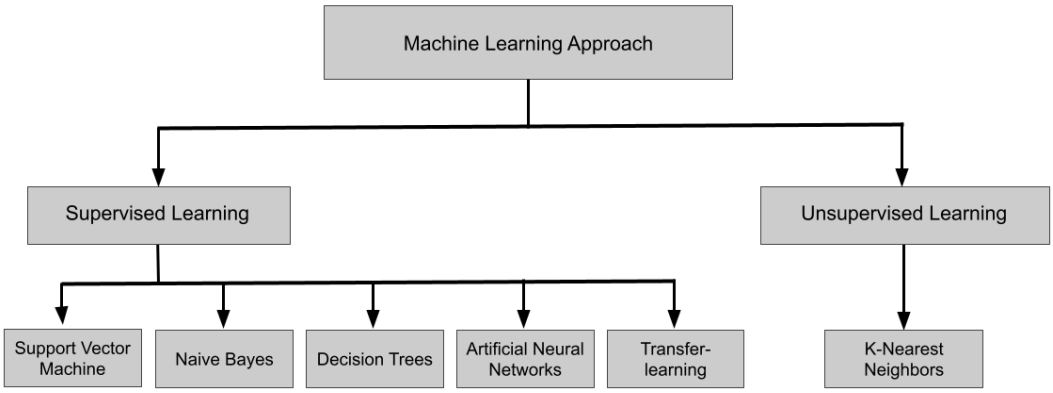

Fig. 3 Summary of the machine learning approach to sentiment analysis and popular algorithms employed for the task. Based on the review conducted by Wankhade et al. (2022)

Supervised machine learning Supervised learning is useful in cases when one seeks to determine which category from a predefined set a given object belongs to. In the case of emotion recognition it is the name of emotion expressed in a given piece of text. This method assumes that each document is assigned with at least one category and each category has a set of documents that belong to it. Collection of data points needs to be split into training and test sets. Sometimes a development set (also called a validation set) for adjusting parameters of an algorithm is partitioned as well. In most cases, a training set comprises the highest number of examples, and a development set the lowest. Prediction of categories on a new data involves measuring the similarity between the document and the profiles of categories the model was trained on and choosing the category (or categories in the multi-category multi-label classification) with the highest similarity. The process of supervised learning can be reduced to the following steps: data collection, data annotation, data preprocessing, model training, model testing, and prediction on new data. The standard approach to text classification is summarised in Figure 4.

Although the use of simple algorithms such as Naive Bayes or Support Vector Machines methods is still widespread in sentiment classification, architectures based on deep neural networks systematically show superior performance. In particular, models comprising convolutional neural networks perform better than other architectures in many text classification tasks (Ouyang et al., 2015; Nedjah et al., 2019). In addition, other works report considerable improvements in performance by including the mechanism of attention in the model structure (Kardakis et al., 2021). However, architectures such as BERT currently achieve state-of-the-art results (Zygadło et al., 2021). One of the advantages of deep learning models comes from the ability to handle high dimensional feature spaces. Although, the overall performance in this case depends on the number of samples used for training.

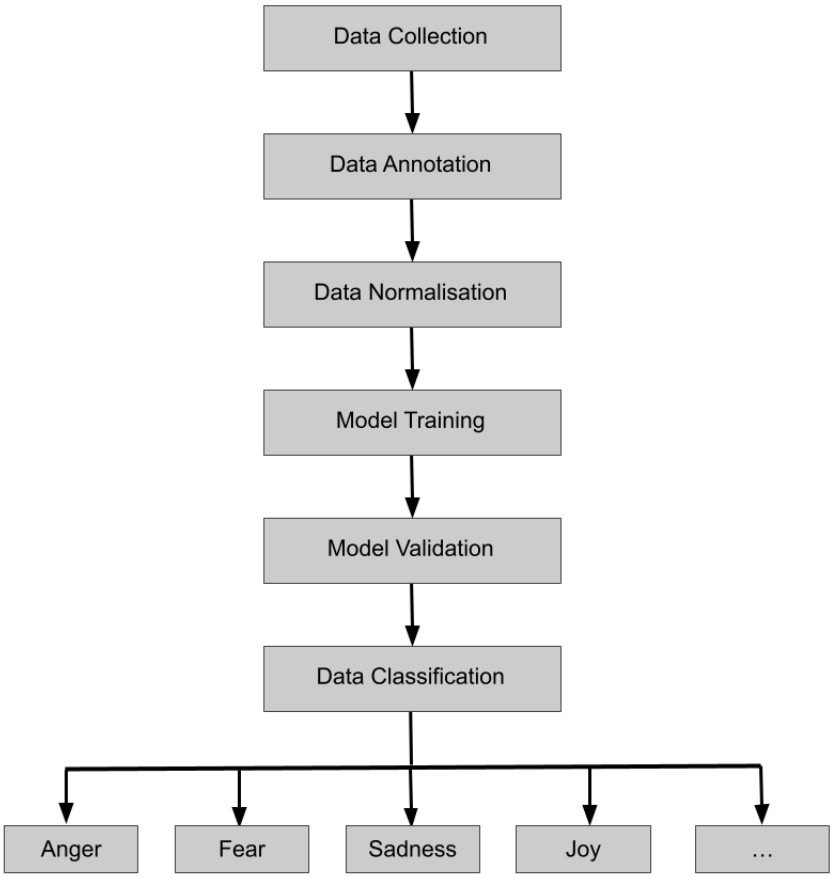

Fig. 4 The process of automated emotion recognition and similar text classification tasks

BERT model used by Fornaciari et al. (2021) achieved third place in WASSA-2021 (the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis) shared task on emotion classification. Combination of convolutional (CNN) and recurrent neural network such as the long short-term memory (LSTM) and gated recurrent unit (GRU) proved useful for sentiment detection in short texts (Wang et al., 2016). Such an architecture is viewed as the most effective in many cases as CNN networks can extract local features from natural language (such as bi-gram phrases) while time-recursive recurrent neural networks (RNNs) are able to learn long-term dependencies in sequential data. Models based on the combination of those architectures achieve better performance than the models build only on one type of these networks (Wang et al., 2016). Nevertheless, traditional machine learning algorithms based on TF-IDF or BoW text representation features are still effective in some cases, especially for two or three level sentiment classification. Rustam et al. (2021) compared performance of popular supervised models for mining sentiment expressed in tweets related to the topic of Covid-19. The authors report superior performance of ETC and the lowest performance for LSTM-based deep learning model. However, authors note that surprisingly poor results of LSTM might come from the small size of training data, insufficient for successful learning of sentiment patterns. In addition, the method proposed by the authors – concatenation of TF-IDF and BoW outperformed either of this two techniques.

Knowledge enhancement is sometimes employed as auxiliary features for machine learning models in order to augment the representation of emotions in text data (Deng & Ren, 2021). It includes emotion-related information about words from emotion lexicons, TF-IDF weighted vectors or hand-crafted rules. This method can also account for the problem with out-of-vocabulary words in text embeddings. Hand-crafted rules are used for negation handling and polarity shifters detection based on the prior knowledge about their linguistic patterns.

Unsupervised machine learning Although most of the current approaches to emotion recognition are based on supervised learning methods, some researchers choose unsupervised learning techniques for emotion mining, in particular by relying on clustering algorithms or the concept of semantic relatedness. The latter technique was used in (Agrawal & An, 2012) in order to compute the "semantic similarity" between affect bearing words in texts and emotion concepts. Here, the highest score would indicate the emotion expressed in a text. Tang et al. (2014) point out that traditional word embeddings model syntactic information about words. For example, words with opposite sentiment such as "good" and "bad" have similar usages and therefore, they are close to each other in a multi-dimensional space of word embeddings. The intuitive approach would place expressions similar in meaning (synonyms) next to each other instead. Thus, traditional word embeddings could be insufficient to effectively detect complex phenomena such as emotions in text. Therefore, Tang et al. (2014) proposed sentiment specific word embedding (SSWE) that encode additional sentiment information into the representation of words and could be more effective than traditional methods in an unsupervised text classification. The basic idea in the unsupervised approach is to divide data into separate groups, categories or clusters, based on the different patters in the input data. Samples within a single cluster are characterised by a common pattern of input features that is at the same time distinguishable from the patterns observed within other clusters. It could be for example, a number of negative vs. positive words in documents that separates data into two categories.

Other works make use of zero-shot learning where in an unsupervised fashion a model (sentence transformer such as BERT) is fed with text input and a set of probable categories (such as the names of emotions), and the model returns the probabilities for the selected set of labels. Then, it could be complemented by a semi-supervised approach. Here, one can treat those automatically obtain results for emotion detection as labels for the main model (Tesfagergish et al., 2022). Unsupervised learning is utilised in cases with little or no annotated data available for the training purposes. It is used also to discover patterns of language use in positive vs. negative sentiment tone. However, it is less popular that the supervised machine learning with clearly defined categories and examples of text that fall into these categories. However, the combination of large language models such are BERT and unsupervised or semi-supervised learning becomes a popular alternative because of the knowledge such models already contain in the weights of their architectures, therefore it seems to relatively easier to employ this method for classification purposes.

1.3 Human-Centred NLP

Methods such as averaging values over the whole group of individual annotations or majority voting are common approaches for the creation of labelled corpora. However, they reflect only a general perspective of a population (rough approximation) on a given topic and at the same time remove a wide range of diversity, for example, in emotion recognition or hate speech detection tasks. Only recently researchers started to take into account such different perspectives of annotators in developing machine learning models for automatic detection of highly subjective phenomena. Offensive content detection has gained much interest in this field and several authors have already published their results (Akhtar et al., 2020; Kocoń et al., 2021). Nonetheless, all authors report further studies are needed to account for the role of individual, sociodemographic and cultural differences in improving performance of machine learning algorithms. Numerous attempts has been made in this strand of work to infer sociodemographic traits of individuals automatically based on their activity and the content they post and/or share on social media platforms, for example (Wong et al., 2016).

Hovy & Yang (2021) compare assumptions about the use of language in NLP to assumptions about decision-making (choices made) by people in economics before 1973 (before the publication of the seminal work by Tversky & Kahneman (1973) and behavioural economics research field). Economists until 1970s viewed individuals as rational decision-makers, completely ignoring the human factor in their equations. Hovy & Yang (2021) claim that the same mistake is made by scientists in NLP field, who vastly simplify language usage by assuming that culturally and demographically diverse groups of people, in general, communicate in the same manner. "[T]he common misconception [is] that language use has primarily to do with words and what they mean. It doesn't. It has primarily to do with people and what they mean" (Schober, 1992, cited in Hovy & Yang 2021). Inclusion of social aspects in NLP could therefore open up new avenues of research in the field to become more human oriented.

Notwithstanding the undeniable success of many NLP applications over the last decades, there is still a room for improvement. Hovy & Yang (2021) go even further and state that this solely data-centred approach heavily limits the NLP field. Thus, they call for modelling seven factors in future NLP systems in order to overcome today's limitations. Those are: speaker's characteristics, characteristics of the audience, type of social relation between the speaker and receiver, context, social norms, culture (ideology), and communicative goals.

In (Hovy & Yang, 2021) the authors provide an example of inaccurate automatic translation where "a message of a 20-year-old German female seems like it was from a 75-year-old American male" (Hovy et al., 2020, cited in Hovy & Yang 2021). Researchers attribute this misunderstanding to the lack of information about personality of the speaker. Researchers cite also several papers whose authors demonstrated superior performance of machine learning models by accounting for speakers' sociodemographic attributes. Moreover, ignoring these factors may give rise to many biases, for instance, gender-bias in translations (Stanovsky et al., 2019). Similarly, the presence of social stereotypes can be found in many word embeddings (Bolukbasi et al., 2016; Sheng et al., 2019) and there are ongoing research initiatives with the aim of investigating and (potentially) mitigating them (Garg et al., 2018).

Hovy & Yang (2021) stress the importance of including context information in develop-

ment of NLP tools as well. That is because discussion on social media sites, such as Twitter, differs from the language used, for example, in scientific papers. In that paper authors point also to another crucial fact — that including extralinguistic information in development of language models, and especially in tasks such as text generation, could make them follow more strictly Gricean maxims, and therefore improve machine generated text so as to resemble to a greater degree human language.

Another interesting finding comes from the work by Welch et al. (2020). The authors demonstrate that the same word could have different representation in vector spaces of word embeddings depending on the sample of data used for training. For example, "health" was related to words such as "treatment", "preventative", "mental", "benefits" and "medical" in a group of younger people on the one hand, and "care", "reform", "coverage", "insurance" and "socialised" in a group of older people, on the other hand. Therefore, the meanings of words and their associations vary across different sociodemographic groups. Conover et al. (2011) show that machine learning models could accurately predict political alignment of users on Twitter based solely on words (textual content) or hashtags individuals use. Therefore, it seems reasonable to assume that left-wing and right-wing groups of people employ different language to communicate on social media.

Annotation of texts in terms of offensiveness or emotions is very subjective, not like, for example, part of speech tagging. Indeed, researchers in the field point out moderate or low levels of agreement between annotators on this tasks (Waseem, 2016; Luo et al., 2020; Kocoń et al., 2021; Prabhakaran et al., 2021). They also point to the problem with "ground truth labels" in text annotation, in particular in the case of highly subjective tasks. They stress the fact that although label aggregation techniques such as majority voting and averaging are commonly employed for the acquisition of single "gold labels", annotators' disagreements may reflect important nuances in language which are lost with those aggregation methods. Alm (2011) argues that achieving a single real "ground truth" is not possible, nor essential, in subjective tasks, and calls for finding ways to model subjective interpretations of annotators, rather than seeking to reduce the variability in annotations. In their own study Prabhakaran et al. (2021) compared inter-annotator agreement coefficients between each individual rater and the majority in three tasks: detection of sentiment, emotions and hate speech. Results show that one-third of annotators had very low agreement scores in terms of emotion and sentiment annotation tasks ($\kappa < 0.4$). In view of those findings authors call for increased transparency in the data annotation process by releasing (raw) annotator-level labels as well as additional information (metadata) about annotators such as their demographic statistics and documentation about recruitment, selection, and assignment of annotators.

Waseem (2016) shows systematic differences in the annotation of offensiveness in text from the perspective of amateurs (crowd sourcing) and experts (feminist and anti-racism activists). Here, the annotator agreement coefficient between experts' annotations and majority voted labels is low (κ=0.3). What is more, Waseem (2016) report that extra-linguistic features tuned out to be the most important in the prediction of models developed on majority aggregated amateur annotations. On the other hand, the content of the tweets was the most important feature for models trained on experts' data.

The idea of including subjectivity of interpretation in the data labelling process fits into the framework of two paradigms of annotation proposed recently in (Röttger et al., 2021). There are at least two factors that contribute to the quality of the annotated data. Annotation guidelines provide a definition of the annotated concept, an available set of categories to annotate from and specification of desired behaviours from human raters. However, personal traits of the raters (e.g., cognitive capacity, motivation) also have an impact on the quality of annotated corpora. The prescriptive approach, although, acknowledges annotator subjectivity, pursues to obtain a consistent sample annotated according to one belief specified in the curated annotation guidelines. Descriptive annotation paradigm on the other hand, encourages individual interpretation of the task at hand. Corpora creators should adopt one or the other approach depending on the intended downstream use of data as the intended use of the corpora is different in the two paradigms. The former framework is able to implement company policy about moderation of harmful content on their website, for example (Röttger et al., 2021). The latter allows modelling of different beliefs in regard to subjective task such as offensive content detection and therefore could be an area of special interest for researchers working in (computational) social sciences in particular. Therefore, the current work follows the prescriptive paradigm to data annotation and model training.

A prescriptive paradigm puts into question the idea of "gold standard" for data annotation. There is no single ground truth label. Neither a majority vote, nor a mere average or centroid reproduces prototypical user beliefs in a downstream task. The idea of a hypothetical "average person" – the "average Jane" is flawed in this approach (Rizos & Schuller, 2020). Klenner et al. (2020) argue that regarding sentiment inference, the annotation process reflects a distribution of annotations representing a diversity of opinions and thus, corpora creators should leave a room for those individual perspectives and avoid (at least simple) harmonisation. Therefore, alternative strategies of acquiring a single target from several labelers need to be implemented. Rizos & Schuller (2020) propose to model trustability of individual annotators, exploit sample informativeness and weighing samples by annotator certainty. In section 2.1 I propose a custom method – Personalisation Metric – for modelling annotator uncertainty and disagreement. This approach fits into the concept of recommender systems and the framework of Subjective AI (Chen et al., 2022a).

Despite many studies reporting low inter-annotator agreement rates on highly subjective tasks such as sentiment detection, corpora with relatively high inter-annotator agreement coefficients could be found as well. For instance, Mohammad & Turney (2010) noticed that in most cases annotators from Mechanical Turk agreed on emotion intensity (intensity as binary labels — 1: weak or no emotion, 2: moderate or strong emotion) elicited by a given word — in over 50% of a sample all 5 annotators agreed and 80% of time at least 4 out of 5 agreed. However, in the **??** (Results) section I show that even for corpora with relatively high agreement scores it is beneficial to adopt personalised approach to text analysis.

Kocoń et al. (2021) emphasise that part of the problem on might also lie in the fact that there is no formal definition of offensive or hate speech content. In regard to those findings Kocoń et al. (2021) propose their own solution — changing perspective from *data-centric* to *human-centred*. They hypothesise that taking into account annotators' personal

information and their previous behaviours could improve the quality of machine learning models for offensive content identification. In the proposed human-centred approach the authors recommend to shift attention to readers' perceptions rather than focusing solely on authors' intentions. Personal differences in opinions and judgements on highly subjective phenomena such as offensive content and emotion detection constitute the core of this novel human-centred approach.

In the proposed framework Kocoń et al. (2021) confront three different perspectives (scales) for the development of automated text classification systems in NLP:

- macroscopic (general, whole society)
- mesoscopic (group-based)
- and microscopic (individuals)

With the proposed human-centred approach the authors were able to answer in the affirmative the following research questions:

1. Does the agreement level of the group of annotators influence the quality of recognition methods trained on the data related to the individual group?
2. Does additional information about annotators and content improve the quality of offensive content recognition?
3. Do personalised models significantly improve the quality of offensive content recognition?

The researchers also proposed two methods for splitting data into separate groups at the mesoscopic level. First, they were able to differentiate 11 groups based on the patterns of individuals' annotation behaviour, i.e. inter-annotator agreement coefficients for the classification of offensive and non-offensive comments. Here, the authors employed a custom WAVE (Wroclaw Annotators Variability Estimator) kappa statistic and the Krippendorff's alpha measure. Second, the data could be split into separate datasets based on annotators' demographic features. However, the authors reported not significant differences between groups differentiated in such a manner in the used corpus.

The authors of human-centred approach adopted in the current thesis also propose several methods of building personalised text classifiers. First, at the mesoscopic level they extend training input for machine learning models with a set of auxiliary features related to content metadata such as the publication year of text, as well as demographic information about annotators such as gender, education level, and age group. Second, for the microscopic perspective they design two scenarios of model training. In the first user-based scenario the authors make use of annotator identifiers (id numbers) encoded as one-hot vectors. In the second scenario that adopts the microscopic perspective they rely on opinion-based features. That is, additional text embeddings are created separately for texts annotated by a given annotator as offensive and non-offensive. It is achieved by concatenation of all comments classified as offensive and non-offensive by a given rater and transformation of the content by

the FastText model (Bojanowski et al., 2016). Those personalised approaches are compared to the baseline model that makes use of text embeddings only and majority aggregated labels. To summarise this part, the feature set used in the experimental settings include the following data: text of a comment, text metadata (including publication year, whether author was logged in), annotator metadata (gender, age, education, whether a person is a native English speaker), annotator id number, text emmbedding for the neutral class (not offensive comment) and text emmbedding for the offensive class. In addition, Kocoń et al. (2021) tested four different architectures for offensive content recognition: XLM-RoBERTa, LSTM, fastText+LogisticRegression, and BERT (Devlin et al., 2019). Experiments were conducted on the *Wikipedia Talk Labels: Aggression* dataset (Wulczyn et al., 2017) composed of 100,000 comments posted on English Wikipedia.

Results demonstrate superiority of human-centred approach both at the mesoscopic and the microscopic level compared to the baseline. Additional demographic features increased performance in 74% of cases. Implementation of other auxiliary variables (annotator id numbers) improved the results even more – in over 95% of cases. The authors summarise their study in a following manner "results show that personalised models significantly improve the quality of offensive content recognition and help to determine the potential quality limit of systems based on available data, regardless of the text representation model used" (Kocoń et al., 2021, p. 2). Later the authors conclude that "there are other factors that differentiate annotators' perception (human opinions) rather than obvious demographic features" (Kocoń et al., 2021, p. 17). Nonetheless, later they notice that "additional information related to personalisation increases the quality of prediction, regardless of the model and quality measure used". In addition, the authors noticed almost perfect linear correlation between the intra-group agreement and accuracy of machine learning classifiers. It proves that agreement levels can be "artificially" enforced in order to achieve high performance results for machine learning models which, however, is a bad practice in particular for highly subjective tasks.

The framework proposed in (Kocoń et al., 2021) was inspired by another work (Akhtar et al., 2020) in which researchers in each of 3 cases split annotators into two groups based on their annotation behaviour. Overall, they trained 9 models on 3 separate datasets with binary labels [0, 1]. The first dataset in English language consists of annotations for presence or absence of sexism in tweets, the second dataset is also in English and contains labels for identification of racism in tweets content, and the third one is annotated dataset of tweets which brings up the topic of hate speech detection against the LGBT community in Italian language. Dividing annotators into groups and building models based on them showed improvement in hate speech detection.

In a similar manner Hovy (2015) demonstrates that extra-linguistic factors influence language use and therefore should be accounted for in automatic systems for text analysis. Information about age or gender was found to consistently improve performance over demographic-agnostic models across different text-classification tasks (sentiment analysis, topic classification). Therefore, the author argues against the traditional approach in NLP that treats language as uniform.

Another strand of work that demonstrates the importance of a personalised approach to

AI comes from the studies on annotation bias. Binns et al. (2017) analysed gender bias in the annotation and machine learning classifiers related to "toxicity" of language. The authors made use of a Wikipedia dataset of 100k comments labelled with levels of offensiveness by male and female raters. Results show that on average women annotated comments as less toxic than men. Content of comments the group of female annotators did not labelled as offensive was more likely to be misclassified as such by both "male" and "female" classifiers. Similar methodology was employed in (Al Kuwatly et al., 2020). The authors used personal attack corpora from the Wikipedia's Detox project to investigate whether demographic features of annotators influence their judgements of hateful content. They trained models on the data from one group of annotators and tested on the samples from another group based on factors such as gender, age, education level and first language. Besides gender, Al Kuwatly et al. (2020) found significant differences in the performance of models trained on the data from different groups of raters. Therefore, some demographic factors play a notable role in perception of hateful content and reveal annotator bias. Thus, algorithms developed on data annotated by a homogeneous group of individuals could be of low quality and lead to unfair judgement of some subgroups.

Wich et al. (2020) also investigate the impact of political bias on the automated identification of hate speech in text. Based on the data collected from left-wing and right-wing Twitter communities the authors reveal differences in the prediction quality of models trained with a politically biased data. In turn Larimore et al. (2021) demonstrate that racial identity as well as textual features of tweets influence annotator perceptions of racism. The author discovered that racial identity (White vs. non-White) have a small but significant impact on the annotation of racial sentiment regarding Black people. Those differences were even more influential when controlling for the specific content of a tweet. Results indicate that the higher the proportion of the police brutality topic being present in a tweet, the higher the rating of sentiment given by White annotators. Racial bias was found in the annotation of toxic language also by Sap et al. (2022). For example, more conservative raters were more likely to rate African American English dialect as toxic, however less likely to rate anti-Black language as such.

Automated prediction of persuasion and belief change is another area of study where adoption of a personalised approach proved useful (Lukin et al., 2017; Durmus & Cardie, 2019; Wang et al., 2019; Al Khatib et al., 2020). In this strand of work personality traits of debaters are the most often exploited feature. Here, researchers profile the Big Five personality traits (so-called OCEAN traits: O: openness to experience, C: conscientiousness, E: extraversion, A: agreeableness, and N: neuroticism) collected with the use of traditional questionnaires or via automated prediction with language models. Personality of a persuadee or a voter is reported as an important factor in prediction of persuasion with machine learning. Lukin et al. (2017) found that conscientious people are more convinced by emotional arguments, whereas people scoring high on the agreeableness scale are more persuaded by factual arguments. Prior beliefs about the topics of discussion or other controversial issues is another feature utilised in many studies. Results of (Longpre et al., 2019) indicate that the combination of linguistic features of an argument itself and audience features related to individuals' prior beliefs, online

behaviour and demographic information are the most successful in prediction of persuasion of an individual voter. The authors also note that the more controversial the issue of discussion (political and religious debate categories) the more informative are the factors related to the audience members in this kind of prediction task. Scholars working in this area propose to change the focus of NLP research from uncovering linguistic features that define persuasive arguments to more human-centred approach accounting for factors related to the persuadees as well. Therefore, the argue for incorporation of findings on persuasion in social sciences. This area of research suggest that there are important differences between the persuasion of *a priori* decided and undecided. Findings in psychology show that prior experiences influence the framing of a message perceived by an individual in order to maintain consistency between prior beliefs and current attitudes towards a topic of discussion. Undecided voters in turn hold high potential for persuasion (Longpre et al., 2019).

Miłkowski et al. (2021) emphasise that popular methods of labels aggregation — averaging and majority voting — could accurately identify only the general views of some population. However, the utility of those methods for prediction of emotions elicited in a particular person is limited. The authors also note low inter-annotator agreement in most corpora annotated with emotional ratings because of highly subjective nature of the task. Therefore, Miłkowski et al. (2021) propose personalised to detection of emotional ratings. However, instead of relying on demographic features or personality traits employed in previous works the authors use custom *Personal Emotional Bias* (PEB) metric for measuring individual subjectivity of emotion annotation. PEB formula is based on the calculation of Z-scores between the average and the individual user rating and is suitable for real-valued ratings (numeric variables) of emotional intensity. It reflects the previous annotation behaviour of an individual and the degree to which this behaviour differs from the average one. PEB metric was tested on a custom corpus of 7,000 online reviews developed by the authors and annotated in terms of 10 affective categories — valence, arousal and 8 primary emotions proposed by Plutchik (Plutchik, 1982). Each text sample was rated on average by 50 participants from the study group of almost 9,000 annotators. In (Miłkowski et al., 2021) the researchers tested the usefulness of PEB metric against other features used in supervised text analysis such as text embeddings, demographic information, and their combination. The authors also tested different model architectures – HerBERT, XLM-RoBERTa, fastText+LSTM and RoBERTa. Best performance was systematically achieved by models employing PEB and text features. Nonetheless, experimental settings that included all features (average ratings, text embeddings, demographic data, PEB) achieved similar results. Results of this study indicate that personal information about individuals is an important feature in automated detection of emotional ratings. The authors conclude that in regard to highly subjective tasks performance of the reasoning based on text only "is greatest for the highest agreement (valence) and lowest for low agreements (surprise, arousal and anticipation). It means that the more users disagree, to the greater extent we should rely on personal biases rather than solely on the textual content" (Miłkowski et al., 2021, p. 255).

Similarly, Davani et al. (2022) argue for several problems associated with label aggregation methods, especially in the case of annotation tasks that are highly subjective. Those include

internal inconsistency in labels and limited representation of minority perspectives in data. In their work, they present a multi-annotator (multi-task) model which makes use of individual annotations and was tested in two tasks — abusive content detection and emotion detection. Authors tested 3 different types of multi-annotator models: ensemble, multi-label and multi-task. In the first approach N individual models (N being a number of annotators) are trained on data and labels provided by n-th annotator. The final label is predicted by aggregating individual predictions, therefore being selected by the majority vote of all N models. In the second approach a model's task can be seen as a prediction of N labels, that is, predicting whether the n-th annotator would assign label 0 or 1 to a given text instance. The outcome layer consists of one fully-connected layer. In this case the final label is produced by the majority voting as well. The third architecture comprises N binary classification tasks, trained on the shared representation layer which is fine-tuned based on the predictions of all tasks. Similarly as in the previous cases, here the final label is chosen with the use of the majority voting method, however there are N final fully-connected layers which are fine-tuned separately for each annotator. For evaluation authors firstly, compare how well predicted aggregated labels match the majority vote of annotations, and secondly, how well individual predicted labels match labels given by individual annotators. In terms of F_1 scores, in the case of hate speech detection the third (multi-task) model performed better than other models, both in predicting majority vote labels and individual labels. Based on results authors formulate an interesting hypothesis: "modelling each annotator, and their presumable internal consistency, could lead to more stable prediction results" [p. 7]. Additionally, authors measure model uncertainty, which reflects disagreements between annotators present in datasets. They conclude that the multi-task model not only performs best at hate speech detection but also at model uncertainty prediction — reflecting annotation disagreements. Because multi-task gave the best results for hate speech detection, authors used this architecture to compare traditional vs. multi-annotator models for emotion recognition. Multi-annotator architecture performed better than baseline in predicting the majority label for four emotions given the adjusted subset of data samples (taking into account only those annotators that labelled at least 1000 instances). Following Alm (2011), authors state that in highly subjective tasks such as emotion recognition and hate speech detection the aim could be to find the most acceptable answer (taking into account many perspectives from different annotators at the same time) rather than the most accurate one.

In turn Chou & Lee (2019) demonstrate the superiority of a simple blended model for emotion recognition from speech by including both hard and soft labels prediction layers in the model. Moreover, they proposed a novel architecture that comprises not only layers for hard and soft label prediction but also annotators' individual perspectives.

Personalised approach was also implemented in the domain of persuasion by Al Khatib et al. (2020). Authors explore personal characteristics of argument sources and the argument audience in order to develop a classifier for automatic prediction of persuasiveness of messages. Instead of relying on traditional questionnaires or explicitly available information, Al Khatib et al. (2020) utilise the history of debaters activity to model their prior beliefs, personality traits and interests that are hypothesised to be the main factors for establishing persuasion.

Prior beliefs were operationalised as stance (sentiment) towards a particular topic. In regard to personality traits mining, researchers made use of categories in the Linguistic Inquiry and Word Count (LIWC) dictionary. Each post generated by a debater was fed through this tool in order to obtain a debater's score on the Big Five personality traits. Finally, interests were captured by the level of the debater's activity (number of posts generated) in selected subreddits. In addition, the authors compared characteristics of the persuaders and persuadees using cosine similarity and the features described above. Results indicate that the features related to personality traits of debaters were the most effective in prediction of the persuasiveness of arguments.

In a similar manner, Durmus & Cardie (2019) study the role of voters' prior beliefs for argument persuasion. It follows the results of research in psychology that has shown the influence of prior beliefs on argument interpretation and therefore their role in belief change and resistance to persuasion. With the use of feature engineering techniques the authors account for various linguistic as well as user-related factors. Regarding the former, researchers explore several content-based features: argument length; TF-IDF weighted unigrams, bigrams and trigrams, referring to the opponent phrases, politeness cues, citation of sources, text polarity, use of personal pronouns and modal verbs, as well as the number of swear words, spelling errors, links, exclamation marks, questions and phrases corresponding to different argumentation styles. In regard to the latter, user-based features, the authors measure opinion similarity between debaters (the cosine similarity on controversial issues voting behaviour) and check whether two users match on political and religious ideology. The best accuracy is achieved with a combination of all user features and length and TF-IDF features when controlling for debaters political ideology. Similar results were obtained in the study by Longpre et al. (2019), which is an extension of (Durmus & Cardie, 2019). Here, also various features related to the voters and language features proved useful in prediction of persuasion of undecided and decided voters.

2 Methodology

By analogy to works on offensive language detection, for example (Sap et al., 2022), that show annotator identity has considerable influence on her/his perception of hate speech, the same might hold true for emotion recognition. Thus, contextualising such tasks with social or personal variables, i.e. incorporating some type of *persona* modelling could result in the development of more accurate solutions.

My work follows the novel human-centred perspective presented in (Kocoń et al., 2021) and employs similar methods for building machine learning models. My study is also inspired by research in the related field of detection of annotator bias — including gender, racial and political bias (Binns et al., 2017; Wich et al., 2020; Davidson et al., 2019). In the domain of personalised approach to emotion detection work by Miłkowski et al. (2021) is especially informative for my own research. Nevertheless, what I want to present, similarly to several previous works of others, is that instead of looking at differences in annotation data as annotator biases or undesired noise, we can look at them as individual differences that are natural and well documented in psychology's research.

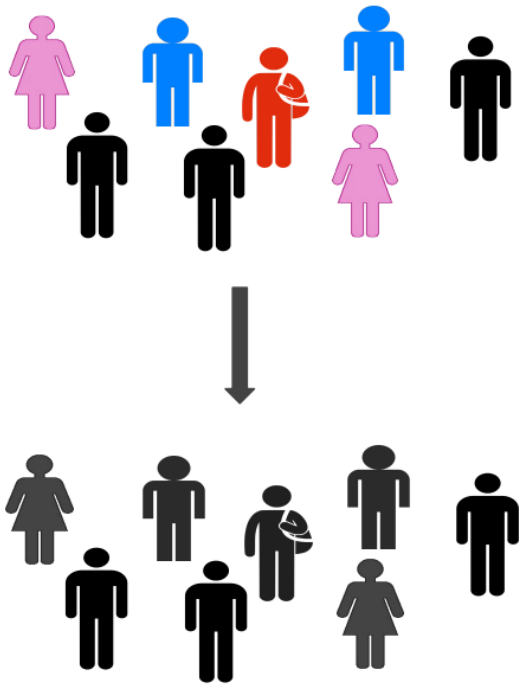

Fig. 5 Schematic presentation of the common approach to corpora development. Although individuals possess different opinions on the issue such as emotion recognition (represented on the top of the figure by different colours of silhouettes), in the end they are treated uniformly (represented on the bottom of the figure by greyish colour). As a result, any information on differences among individuals is lost on behalf of the search for the objective "ground truth"

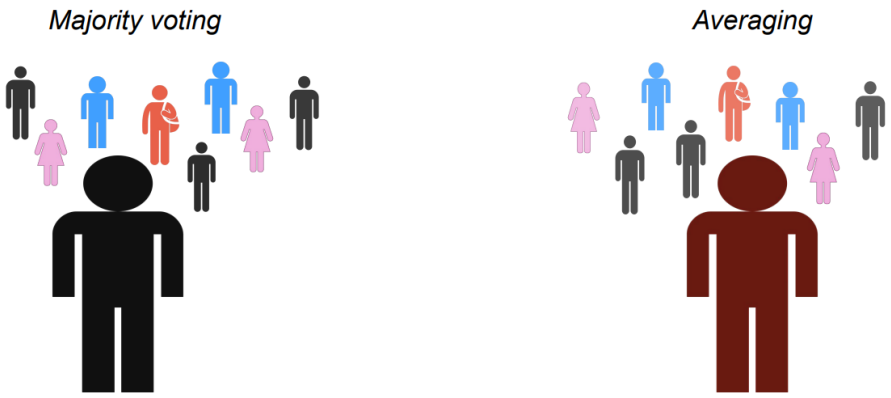

Fig. 6 Results obtained with the use of label aggregation techniques in standard approaches to text classification. On the left hand-side, schematic presentation of the majority-voting technique is drawn. On the right hand-side, another label aggregating technique is displayed – averaging of values. Individual perspectives are not represented in either of these approaches – regarding the majority-voting, these opinions are simply discarded, only the dominant view is hold (in this case represented by a black silhouette); regarding the averaging, all opinions are blended into the final one, which, however, does not represent any individual opinion from the sample (represented by a brownish silhouette in the figure)

Similarly to Chou & Lee (2019), my aim in this work is to account for the variability and the subjective nature of emotion recognition. Thus, first, following the methodology presented in (Al Kuwatly et al., 2020), I demonstrate systematic (and noteworthy) differences in emotion recognition between individuals. Second, by providing machine learning models with additional information about each annotator I show its positive impact on the predictive ability of emotion recognition classifiers. In addition, by using noisy (individual) labels in my experiments, I show their value in training deep learning algorithms.

The solution proposed in the current thesis fits into the recently introduced human-centred perspective in NLP (as an alternative to data-centric accounts), which accounts for social factors in tasks such as offensive language detection, sentiment analysis and sarcasm recognition (Kocoń et al., 2021). Several techniques has been proposed so far in terms of human-centred text classification, however, they are either limited to the specific types of data (real-valued labels in the case of Miłkowski et al. (2021)) or require potentially sensitive information about individuals such as demographic factors and personality types Lukin et al. (2017); Hovy (2015). The Personalisation Metric proposed in the current thesis does not require collection of any private information about individuals and could be calculated based solely on their annotation behaviour (i.e., labels assigned by individuals on a sample of data). Moreover, the Personalisation Metric is designed to work on categorical labels, which are predominant type of data available in annotated corpora as well as text processing tasks in NLP. Thus, it is a useful extension of the PEB feature designed for real-valued labels (Miłkowski et al., 2021)

2.1 Personalisation Metric

Personalisation Metric (PM) proposed in the current work is based on Miłkowski et al. (2021)'s *Personal Emotion Bias* (PEB) statistic. PEB represents the subjectivity of emotion recognition in terms of an individual user deviation from the average. For annotator a and emotional category e it calculates the Z-score of a's ratings from the average rating for the category e. In other words, it measures the distance from the mean. Thus, it is designed for numeric labels such intensity of elicited emotion measured on a continuous scale. However, sentiment analysis and emotion recognition usually comprise a classification task, therefore I develop my own version of PEB, suitable for categorical labels. I propose to use widely known agreement measures as a method of encoding the subjectivity of emotion recognition (quantified similarity with the majority opinion, in fact) into deep learning models. It comprises two different types of values — Cohen's kappa coefficient and percentage agreement (accuracy score). Therefore, first, it calculates the kappa statistic between annotator a_l and the majority, and second, it computes the percentage agreement between annotator a_l and the majority. Both coefficients could be calculated for binary as well as multi-category labels.

Cohen's kappa is a measure of agreement of a qualitative variable between two instances – two individuals, two groups or one individual and one group. Therefore, the more one's annotations deviate from the comparison ones, the lower the value for Cohen's kappa. It is one of the simplest and most commonly employed measures of agreement. However, because of the correction for chance agreement it employs, sometimes low scores are obtained despite fair agreement between two individuals (Zec et al., 2017). That is the reason why my PM metric comprises two measures, in fact – Cohen's kappa and accuracy. The formula for calculating accuracy score is similar to that for Cohen's kappa, although it is not corrected for agreement by chance between judges. Thus, it returns so-called observed agreement. As a result, there could be cases when Cohen's kappa is moderately low but the value of accuracy is moderately high. Cohen's kappa is given in the formula 1.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o is relative observed agreement between two raters (in other words, it is simple percentage agreement), and p_e is a hypothetical probability of chance agreement calculated as follows:

$$p_{e1} = \frac{TN + FP}{TN + FN + FP + TP} \times \frac{TN + FN}{TN + FN + FP + TP} \tag{2}$$

$$p_{e2} = \frac{TP + FN}{TN + FN + FP + TP} \times \frac{TP + FP}{TN + FN + FP + TP} \tag{3}$$

$$p_e = p_{e1} + p_{e2} \tag{4}$$

where TP is a number of true positives, TN is a number of true negatives, FP is a number of false positives, and FN is a number of false negatives.

Accuracy score (percentage agreement) is calculated according to the equation 5.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Given annotator a_l's ratings r_l, i.e. a vector composed of labels s_n, and the majority's vector m comprised of majority-aggregated labels s_n, PM metrics are calculated according to equations 6 and 7.

$$PM\_kappa(a_l) = \kappa(r_l, m) \tag{6}$$

$$PM\_accuracy(a_l) = accuracy(r_l, m) \tag{7}$$

In both measures, the values close to 1 signal perfect agreement, whereas scores around 0 indicate no agreement in the case of accuracy, and random agreement in the case of Cohen's kappa. Negative values of kappa reflect less than random agreement between individuals. Thus, in my approach I treat disagreements as an additional source of information instead of noise, which is the main assumption underling majority-based accounts.

2.2 Model Architecture

All experiments were performed with the use of *Keras* and *TensorFlow* libraries in Python programming language. As a text representation method I employed *GloVe* 100-dimensional pre-trained word embeddings (Pennington et al., 2014). Architecture of the proposed model is fixed in each setting. Development of the model structure follows recent findings in the sentiment analysis literature (Hameed & Garcia-Zapirain, 2020; Tam et al., 2021). That is, the proposed model is based on the combination of convolutional neural networks (CNN) and recurrent neural networks – specifically bidirectional long short term memory networks (BiLSTM). Although, state-of-the-art performance is achieved mostly by Transformer-based architectures such as BERT (Devlin et al., 2019), research shows that CNN and LSTM networks still perform well in many text classification tasks. The proposed model architecture comprises 10 layers. The first one, an embedding layer, transforms the textual input into 100-dimensional word embeddings. It is followed by one convolutional layer with 128 units and a max pooling layer. The output of the max pooling layer is then fed to a bidirectional LSTM layer with 64 units, which is followed by a global max pooling layer. Then, I use a sequence of fully-connected layers with 246 and 128 units, that are separated by a dropout layer and a concatenation layer. Here, the output of the first Dense layer is concatenated with the additional features (one-hot vectors comprised of annotator ids or PM metric values) in the two experimental conditions that implement personalised approach (PM, and PM and vector-ids in the Pilot and in the large-scale study, respectively).

The output layer is designed differently in two studies. In the small-scale study it consists of 4 units and the *softmax* function. Also, the loss function is measured by the categorical cross-entropy here. In the large-scale study, the last layer consists of 7 nodes and the *sigmoid* activation function as the task comprises multi-class multi-label classification. Thus, for each text example it could detect more than one emotion. And loss is measured with the binary

cross-entropy function. All hidden layers (CNN, LSTM, Dense) make use of the ReLu activation. The proposed CNN-BiLSTM architecture is presented in Figure 7 and describe in detail in Table 3. The models were trained for 10 and 30 epochs in the Pilot and in the large-scale study, respectively. Finally, *Adam* was chosen as an optimiser in both cases. All experiments were repeated 5 times, i.e. there were 5 different random splits of data. Each time 20% of data was used as a test set and the remaining 80% comprised a train set. PM metrics were computed *de novo* each time, separately for training and test sets.

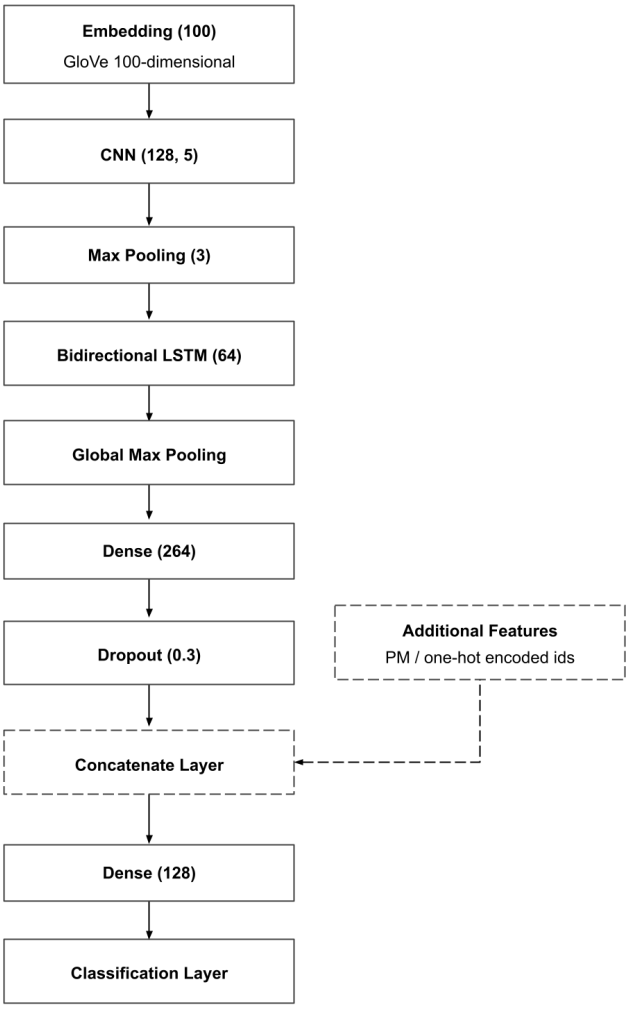

Fig. 7 Graphical representation of the CNN-BiLSTM architecture proposed in the current work

Table 3 Architecture of the CNN-BiLSTM model proposed in the study

Layer	Parameters
Embedding	Type: GloVe; Size: 100
CNN	Filters: 128; Size: 5
MaxPooling	Size: 3
BiLSTM	Units: 64
GlobalMaxPooling	—
Dense	Units: 264
Dropout	Rate: 0.3
Concatenate	—
Dense	Units: 128

3 Small-scale Study

3.1 Material

Multi-View Sentiment Corpus (MVSC) (Nozza et al., 2017) comprises 3,000 Twitter posts related to two movies — *Deadpool* and *Suicide Squad*. Each text was manually labelled by three annotators on five different aspects: irony, subjectivity/objectivity, sentiment, emotion, and implicitness/explicitness of opinion. For the purpose of my study, annotation only on the level of sentiment was chosen. In regard to this aspect, annotators were asked to classify each tweet as expressing one out of the four following categories of sentiment: *none*, *negative*, *neutral* and *positive*. *None* label was assigned to all texts classified earlier by a particular annotator as objective, and the rest, i.e. all subjective tweets, needed to be labelled with either of the other three categories of sentiment.

Regarding the first step in the annotation of sentiment, i.e., decision whether a tweet is objective or subjective, Nozza et al. (2017) provide the following definition "an *objective* post p_o presents some factual information about the world, while a *subjective* post p_s expresses some personal feelings, views, or beliefs" (Nozza et al., 2017, p. 275). Then, raters must determine sentiment (polarity) of all subjective posts given positive, negative and neutral classes. Raters are instructed to annotate neutral when sentiment of a tweet lies between positive and negative category. No further instructions are given regarding the understanding of positivity and negativity. However, annotation guidelines put emphasis on identification of the target of the subjective tweets as it might influence the judgement and interpretation of sentiment expressed in text (Nozza et al., 2017).

There are 9,000 individual annotations in total — 3,000 texts labelled by each of three raters. The most frequent sentiment category is positive (4,235; 47.1%), none is annotated 2,772 times (30.8%), negative category is assigned to 1,047 data points (11.6%), and neutral label is present in 946 samples (10.5%). Majority voting technique was employed for the estimation of aggregate labels. The distribution of majority voted sentiment categories is as follows: positive category is assigned to 1,431 tweets (47.8%), none occurs 923 times (30.8%), negative sentiment is expressed in 363 texts (1.,1%), and neutral labels is annotated in 279 (9.3%) cases. Although there are small differences in terms of the number of tweets annotated with each label by different raters, positive labels is annotated the most frequently by everyone (from 42.6% to 50.9% of tweets are annotated with this label by individual raters), followed by the none category (25.2%-34.3%), neutral (12.3%-14.6%), and negative (3.8%-12.6%).

Pair-wise agreement between annotators measured by the kappa coefficient suggest moderate agreement (κ=0.6). In addition, I calculated Fleiss' kappa coefficient, which allows for assessing the agreement of multiple (more than two) raters. The obtained value of 0.53 also indicates moderate agreement. Perfect agreement (i.e., everyone choose the same label) is observed in the case of 2428 tweets. In 64% of cases two raters disagree, and in 36% of the time all 3 raters assign different sentiment labels to tweets. Interestingly, pair-wise comparison reveals that rater-1 and rater2 as well as rater-2 and rater-3 never agree on the label in those 572 cases of disagreement, and rater-1 and rater-3 agree on the label in 64% of cases (and simultaneously disagree in these instances with rater-2). Most frequently raters

disagree on the following labels: none-positive (113 cases), none-neutral-positive (113 cases), and neutral-positive (106 cases). In Table 4 I present examples of tweets that: all raters agree on the sentiment label, two of three raters agree on the label, and all three raters disagree on the label.

Table 4 Examples of annotation from the MVSC corpus that show raters agreement and disagreement on the assigned sentiment labels

Tweet	Rater-1	Rater-2	Rater-3
RT @username: Joker is the definition of a ride or die bruh #SuicideSquad	neutral	neutral	neutral
RT @username: #SuicideSquad continued the trend of a blue beam shooting into the sky in a CBM.	negative	neutral	negative
RT @username: Yes, I am that one annoying person that will sing Heathens during #SuicideSquad #thecliqueis-goingtojailparty	neutral	negative	none

3.2 Experimental Setup

I hypothesise that feeding algorithms with the proposed PM metric as a set of additional features will improve the classification performance compared to the standard majority based approach. To test this hypothesis I design three experimental conditions in the small-scale study. In the first condition (*PM*), I make use of the proposed Personalisation Metric as additional features fed to a deep learning model. PM is calculated jointly for all four categories of sentiment in the corpus. In the second condition (*cross*) I train a model on majority aggregated labels and evaluate its performance at the level of individual annotations. Here, I randomly sample 20% of text from the aggregated data and retrieve all the individual annotations from the full dataset assigned to texts in this set in order to give deep learning models samples unseen before in the evaluation process. Finally, in the third condition I implement a traditional general view approach where a model is both train and tested on majority aggregated labels. The experimental conditions in the small-scale study could be summarised as follows:

1. **PM**:
    - a training feature set X comprises text embeddings and PM features,
    - a training label set y comprises labels given by individual annotators,
    - evaluation is performed on individual labels
2. **cross**:
    - a training feature set X comprises only text embeddings,
    - a training label set y comprises majority estimated labels,

- evaluation is performed on individual labels

3. **majority**:

   - a training feature set X comprises only text embeddings,
   - a training label set y comprises majority estimated labels,
   - evaluation is performed on individual labels

Source code for the small-scale study is publicly available in Google Colab[4]. Distribution of the calculated Personalisation Metric (PM) for raters in the small-scale study is displayed in Figure 8, and summary of the PM feature is presented in Table 5. Calculation was conducted on the full dataset, i.e., without division into train and test sets, thus actual values of PM fed to the CNN-BiLSTM model in each iteration might slightly deviate from the values presented here.

Table 5 Personalisation Metric (PM) for raters in the small-scale study

Rater-id	PM	Value
1	PM-kappa	0.714
2	PM-kappa	0.706
3	PM-kappa	0.812
1	PM-accuracy	0.812
2	PM-accuracy	0.808
3	PM-accuracy	0.883

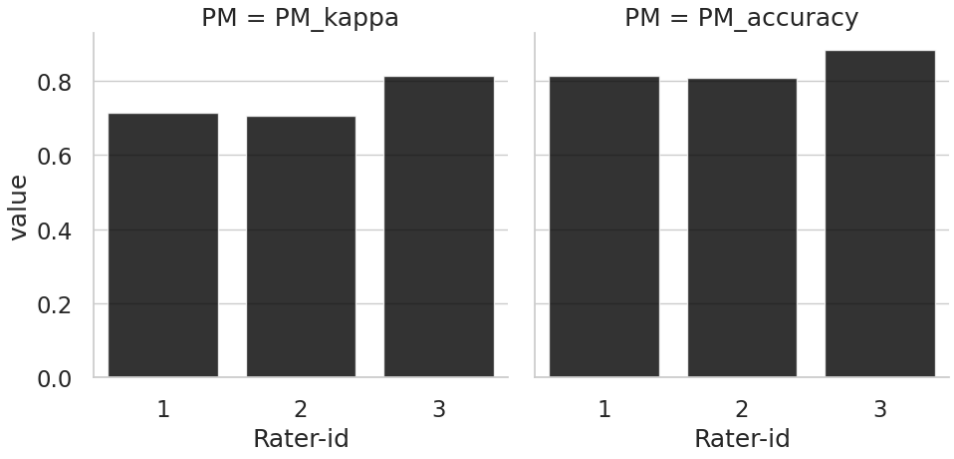

Fig. 8 Distribution of the calculated Personalisation Metric (PM) for raters in the small-scale study

[4] https://colab.research.google.com/drive/14_DR0LrFXfRfsliW_2O1M_04u8S_vnWc?usp=sharing

3.3 Results

In order to measure the performance of machine learning models I employed two standard evaluation metrics in the small-scale study, that is micro-averaged (formula 8) and macro-averaged F_1 (formula 9), calculated as follows:

- **micro-F_1**:

$$Micro - F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

- **macro-F_1**:

$$Macro - F_1 = \frac{1}{|C|} \sum_{c=1}^{|C|} 2 \times \frac{Precision_c \times Recall_c}{Precision_c + Recall_c} \tag{9}$$

Summary results are presented in Figure 9. In order to test differences in performance between conditions, I conducted one-sided t-tests with alpha level (*p-value*) equal to 0.05 and the alternative hypothesis "greater" for the PM model. In addition, the Bonferroni correction for applied to account for multiple comparisons. Equality of variances was tested using F-test ($p > 0.05$). T-test were performed separately for the two metrics (macro-F_1 and micro-F_1).

All tests turned out to be significant ($p < 0.001$). That is, the proposed CNN-BiLSTM in the PM condition achieved significantly better results than the model in the cross condition and in the majority condition. Evaluation results are presented in Table 6 and in Figure 9.

Table 6 Results averaged over 5 random splits of data in the small-scale study. Depicted mean ± standard deviations in parentheses

Model	F_1-micro	F_1-macro
PM	**0.689** (± 0.013)	**0.582** (± 0.021)
cross	0.569 (± 0.018)	0.450 (± 0.023)
majority	0.612 (± 0.009)	0.469 (± 0.024)

Note. Best performance is depicted in bold font.

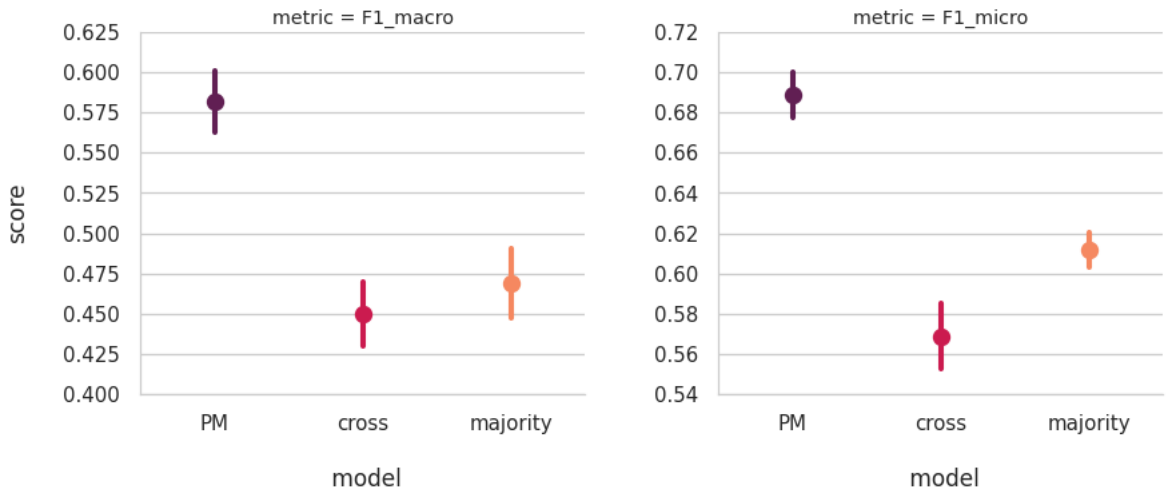

Fig. 9 Results from the small-scale study. Evaluation was performed on 5 different runs of experimental conditions. Vertical bars indicate standard deviations

4 Large-scale Study

4.1 Material

Regarding experiments in the large-scale study I use the GoEmotions dataset (Demszky et al., 2020). It comprises 58,000 Reddit comments from popular subreddits manually annotated with a granular taxonomy of 27 categories of emotion. Thus, it is the largest available collection of documents with the human annotation of emotions. In addition, it encompasses a wide range of emotional categories, which in most cases is limited only to a few categories – 5 or 6 basic emotions (joy, anger, fear, sadness, disgust, and surprise) proposed by Ekman. In regard to the emotional taxonomy, the authors followed works by Cowen et al. (2019) who distinguished 27 distinct varieties of affective experience conveyed by a diverse set of stimuli (short videos, facial expressions, speech prosody).

The authors note that content on Reddit is biased towards toxic and offensive language. Therefore, before the annotation process Reddit comments were filtered for offensive language. Those cases were subsequently removed from the dataset. The authors also applied length filtering on data samples and selected comments that are 3-30 tokens long. Lastly, in order to achieve a balanced sample of negative, positive and neutral categories subreddits were automatically labelled with a machine learning classifier for sentiment and emotions. Then, subreddit discussions (i.e., individual discussions inside a topic-centred groups on Reddit) consisting of more than 30% neutral comments or less than 20% of negative, positive, or ambiguous comments were excluded from the data pool. Emotion bias was reduced by downsampling weakly-labelled posts.

The final set of 58,000 comments was randomly sampled from the data pool after the whole preprocessing procedure. Then, each comment in the GoEmotions dataset was labelled by 3-5 raters (82 unique annotators in total) whose task was to identify emotions expressed by the writer in text. Annotators could assign multiple categories to each example, however, they were asked to annotate only those emotions they feel reasonably certain about. Raters were provided with a pre-defined definitions of emotions and examples of text that, according to the authors of GoEmotions, express a given emotion. Six basic emotions from Ekman's taxonomy are defined as follows (Demszky et al., 2020, p. 4051):

- anger: a strong feeling of displeasure or antagonism;
- disgust: revulsion or strong disapproval aroused by something unpleasant or offensive;
- fear: being afraid or worried;
- joy: feeling of pleasure and happiness;
- sadness: emotional pain, sorrow;
- surprise: feeling astonished, startled by something unexpected.

If no emotion was expressed in a text, raters could choose the neutral category. Estimated inter-rater agreement via the Spearman correlation varies for different categories of emotions.

The highest coefficient (r=0.6) was achieved for ratings on gratitude and admiration, and the lowest for grief (r=0.2).

The authors of GoEmotions released both a dataset with aggregated labels as well as all data samples from the annotation process. In total, there are 211,225 data points manually annotated by human subjects of which 131,369 are labelled with at least one of the emotion of interest, i.e anger, annoyance, disgust, disappointment, grief, remorse, sadness, fear, nervousness, amusement, excitement, joy, love, surprise, or neutral. There are 76,071 (58%) comments labelled with at least one categorical emotion and 55,298 (42%) neutral comments. There are 82 annotators, who labelled 1,602 data units on average (max=6,249, min=1). Training data for experiments in the large-scale study comprises annotations only from raters that labelled at least 334 data points (25th percentile), following remarks given by Davani et al. (2022). It was done in order to achieve more stable and reliable PM scores for individual annotators between train and test sets for experiments in the large-scale study of this thesis. As a result, the remaining 128,950 training data samples were annotated by 61 individuals. After the whole preprocessing procedure I was left with 50,064 unique pieces of content.

Nevertheless, for less frequent categories there is still scarce sample of cases annotated with those emotion for some annotators. For the anger category, on average, each annotator assigned it to 327 texts, with a minimum of 48 cases, and maximum 1,203 text samples. In the case of disgust, each annotator recognised this emotion in 86 texts on average, with a minimum equal to 3 cases for one annotator, and maximum equal to 393 texts for one individual. Sadness category is recognised by each annotator in 274 texts on average, with a minimum number of texts equal to 49 for one rater, and with a maximum number of cases equal to 669 with expressed sadness marked by one annotator. The emotion of fear is annotated in 73 texts on average by each rater, minimum 7 texts in the case of one annotator, maximum 298 text samples for one rater. Joy is recognised by each annotator in 465 texts on average, with minimum number of cases marked with joy equal to 91 for one rater, and a maximum number of 1,608 cases. Surprise is assigned by each annotator to 88 texts on average, with minimum 12 and maximum 256. Finally, neutral category is marked in 889 comments by each rater on average, minimum 22 times by one rater, and maximum 3,404 times by one individual. As a result, the classification task with the use of GoEmotions dataset is challenging even for a standard approach because of the substantial class imbalance on the aggregated level as well as the individual level. Imbalanced distribution generates the prevalence problem in regard to the calculation of kappa coefficient, therefore the proposed combination of Cohen's kappa and percentage agreement in the PM metric seems optimal for this scenario.

Authors of the GoEmotions conducted correlation analysis between ratings for each emotion. Based on results from this analysis I grouped emotions into Ekman's classification of 6 basic emotions — anger, fear, sadness, disgust, joy, and surprise. I considered emotions with correlation coefficient equal to 0.15 or higher. Thus, ratings on anger and annoyance are mapped to the anger group; disgust maps to the disgust group; sadness, disappointment, grief, and remorse are mapped to the sadness group; nervousness and fear map to the fear

group; amusement, excitement, joy, and love map to the joy group; and ratings on surprise become the surprise group. In the training data I include comments classified as neutral as well. Summary of emotion categories in the GoEmotions dataset after preprocessing is presented in Table 7. In turn, Table 8 displays examples of data in the GoEmotions dataset together with annotations of emotions at the individual level, which shows (dis)agreement in the recognition of emotions by individual raters. Qualitative content analysis reveals that when it comes to the annotation of sadness, 'sorry', 'bad', 'sad', 'miss', 'lose', 'hard' are in the top 20 most frequently occurring words in texts, however, one annotator has words such as 'die' and 'death' in this top 20 ranking, and for another annotator these words include 'sh*t' and 'hate'. Then, in regard to the annotation of disgust emotion, overall top 20 ranking comprises of words such as 'bad', 'disgusting', 'weird', 'awful', 'hate', and 'terrible', however, for one of the raters it comprises of words such as 'f**k', 'fight', 'shower', 'body', and 'rape'. This type of analysis shows that individuals follow different linguistic cues when it comes to the annotation of particular emotions, and as a result automatic classifier trained on the majority perspective could underperform at the individual level of emotion detection.

Majority voted labels were estimated separately for each emotional category because of the GoEmotions design that allow multi-label annotations. As a result, 45917 (91.72%) comments are assign with one majority voted label, 1138 (2.27%) express two emotions, 19 (0.04%) are annotated with three categories of emotion, and in 2990 (5.97%) cases majority vote technique could not assign any emotion. When it comes to the distribution of majority voted labels, anger is present in 12.4% of comments, disgust in 2.3%, sadness is expressed in 10.3% data instances, fear is assigned to 2.4% comments, joy occurs in 21.5% of data, surprise in 3.4% and neutral label is assigned 44% of a times by the majority.

Similarly, most of the training data, i.e. all 128,950 data points, is assigned with one category (123,791; 96.0%), 4,940 (3.83%) times raters assigned two emotions, 195 (0.16%) comments are annotated with three labels, 21 (0.02%) times raters assigned four categories of emotion, and 3 (0.002%) comments express five emotions.

Table 7 Summary of the training data for the large-scale study after preprocessing

Emotion Group	Emotion	Percentage	Percentage Group	No.	No. Group
anger	anger	6.2%	15.5%	7,949	19,977
	annoyance	10.4%		13,396	
disgust	disgust	4.1%	4.1%	5,229	5,229
sadness	sadness	5.1%	13.0%	6,627	16,735
	disappointment	6.4%		8,303	
	grief	0.5%		667	
	remorse	1.9%		2,484	
fear	fear	2.4%	3.5%	3,155	4,452
	nervousness	1.4%		1,787	
joy	joy	6.1%	22.0%	7,839	28,348
	amusement	7.0%		9,069	
	excitement	4.3%		5,540	
	love	6.2%		8,058	
surprise	surprise	4.2%	4.2%	5,385	5,385
neutral	neutral	42.1%	42.1%	54,229	54,229

Note. **Emotion Group** – name of the emotion category after mapping selected categories into 6 basic emotions (+ neutral category). **Emotion** – original category of emotion in the GoEmotions dataset. **Percentage** – percentage of the original data labelled with the category from the Emotion column. **Percentage Group** – percentage of training data that is labelled with the emotion from the Emotion Group column. **No.** – number of the original data instances labelled with the category from the Emotion column. **No. Group** – number of training data points that are labelled with the emotion from the Emotion Group column.

Table 8 Examples of 10 randomly selected texts from the GoEmotions dataset with annotated categories of emotions from individual raters

Text	Rater-1	Rater-2	Rater-3	Rater-4
NAME 12 min long? That other 1 min video spoiled me.	sadness, surprise	anger	sadness	joy, surprise
NAME pulls publicity drama stunt with mild success. FTFY	anger	anger	fear	neutral
Imagine thinking like this. I feel kinda bad for them.	sadness	sadness	sadness	sadness
NAME hates us man. He's a stone cold killer.	sadness	neutral	disgust	anger
Looks like something you would see in a comedy movie.	joy	joy	neutral	neutral
Didn't even know we were in the running. I thought if he leaves Chelsea he's only going to real ;)	anger	sadness	joy, surprise	–
Thank you! Don't be afraid to stand up for yourself otherwise you'll ended up in your late 30s exhausted and unhappy!	fear	sadness, fear	–	–
And still our government is not responsible enough to pay its bills. It's rather pathetic.	sadness, fear	anger, disgust, sadness	sadness neutral	
Politics in the KRG are some of the dumbest and most petty of the Middle East.	anger	disgust	neutral	anger
Hahahaha. Wow. Pure gold!	joy	joy	joy	–

4.2 Experimental Setup

The aim of the large-scale study is to develop state-of-the-art classifier that is able to detect emotions in text tailored to an individual tendencies (or preferences). Here, the baseline model is trained in a standard fashion — on majority aggregated labels. It is compared with two other models that account for the subjectivity of emotion recognition. The first one makes use of id numbers assigned to raters in the annotation process. Those ids are encoded as one-hot vectors and comprise an additional source of information for a deep learning model. The second model employs my PM metric as a measure of subjectivity of emotion perception.

The preprocessed GoEmotions dataset comprises ratings from 61 different annotators. After applying data cleaning and filtering procedures a training pool comprises 128,950 instances labelled with 7 different categories (6 emotions plus the neutral category) in total. Furthermore, more than one category could be assigned by one annotator to a single instance of text here. Distribution of the calculated Personalisation Metric (PM) for raters in the

large-scale study is displayed in Figure 10, and summary of the PM feature is presented in Table 9. Calculation was conducted on the full dataset, i.e., without division into train and test sets, thus actual values of PM fed to the CNN-BiLSTM model in each iteration might deviate from the values presented here.

Table 9 Personalisation Metric (PM) descriptive statistics for the large-scale study

Statistic	PM-kappa	PM-accuracy
mean	0.682	0.896
std	0.076	0.029
min	0.444	0.792
25%	0.626	0.886
50%	0.703	0.902
75%	0.731	0.914
max	0.847	0.954

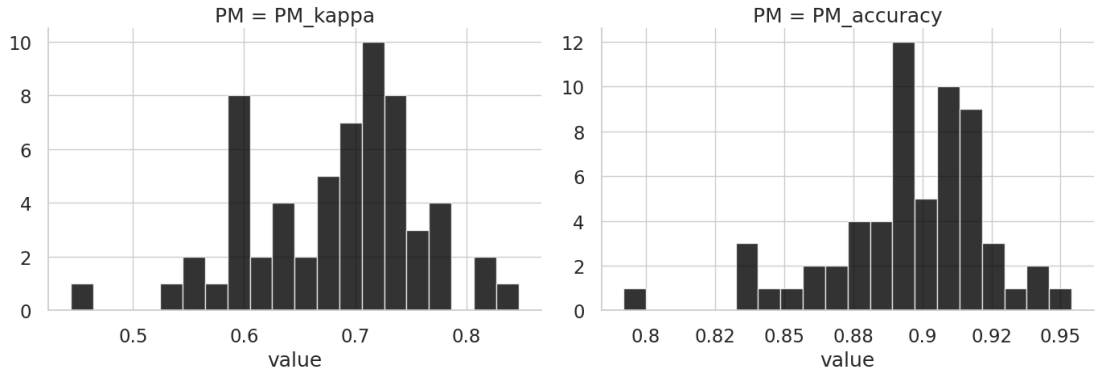

Fig. 10 Distribution of the calculated Personalisation Metric (PM) scores for raters in the large-scale study

Classification task in the large-scale study is more challenging than the one tested in the small-scale study. This is because of several reasons. First, the task changes from a single-label to a multi-label prediction. Second, the model needs to predict 7 different categories of emotion instead of 4. Third, the number of unique annotators is higher — it is 61 in the large-scale study, compared with only 3 raters in the small-scale study. Therefore, the model needs to learn to predict emotion labels for a higher number of individuals. Fourth, the topic of text samples is not uniform, as the GoEmotions dataset comprises of comments extracted from 483 different subreddits (online communities dedicated to discussions on specific topics) – most frequent include r/cringe, r/loveafterlockup and r/socialanxiety. Fifth, the agreement between individual raters is lower, compared to the sentiment annotation in the MVSC corpus utilised in the small-scale study. Reliability of annotation is measured by inter-rater correlation, specifically the Spearman's r coefficient, which varies between 0.16 in the case of

grief annotation and 0.48 for amusement annotation. Therefore, individual tendencies could be assumed to be more diverse in terms of recognition of particular emotions in text compared to individual raters in the MVSC corpus. As a result, automatic prediction of emotion labels for at least some individuals is more challenging in the large-scale study. Lastly, there are not clear and distinct boundaries between some categories of emotion (e.g., disgust is often confused with anger) and other emotions such as surprise are more complex and perceived ambiguous (Bhowmick et al., 2010).

In the first training condition I feed a deep learning model with vectors comprising word embeddings, and *PM_kappa* and *PM_accuracy* measures as the source of subjectivity information. PM measures are calculated separately for each category, that is they make use of binary labels, where 1 indicates the presence of a particular emotion, and 0 its absence. Then these 7 values for *PM_kappa* and *PM_accuracy* are averaged over all categories in order to end up with only one value for *PM_kappa* and *PM_accuracy* in for each text instance. These calculations were dictated by the fact that the task in the large-scale study is a multi-label prediction and kappa statistic accounts only for single labels. As a result, the PM metric is a vector with 2 real-valued scores. The PM model is compared with the second training condition in which input data comprises word embeddings and the annotator id encoded as a one-hot vector. Then, these two models are evaluated against a baseline classifier trained in a traditional way, i.e. with the use of the majority aggregated labels. The large-scale study has one testing condition — prediction of emotion labels given by individual raters. Therefore, three experimental settings are conducted in the large-scale study:

1. **PM**:
    - a training feature set X comprises text embeddings and PM features,
    - a training label set y comprises a vector of labels given by individual annotators,
    - evaluation is performed on individual labels
2. **vector ids**:
    - a training feature set X comprises text embeddings and a one-hot encoded vector of an annotator id,
    - a training label set y comprises a vector of labels given by individual annotators,
    - evaluation is performed on individual labels

Source code for the large-scale study is made publicly available in Google Colab[5].

4.3 Results

Performance of models is evaluated in terms of four standard metrics commonly employed in machine learning research, that is precision (formula 10), recall (formula 11), micro-F_1 (formula 12), and macro-F_1 (formula 13), calculated as follows:

[5] `https://colab.research.google.com/drive/1dHDuhEa_-Ezb3_zvFAtjf-oeNNT8tsa-?usp=sharing`

- **precision**:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

where TP and FP is a number of true positives and false positives, respectively.

- **recall**:

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

where TP and FN is a number of true positives and false negatives, respectively.

- **micro-F_1**:

$$Micro - F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

- **macro-F_1**:

$$Macro - F_1 = \frac{1}{|C|} \sum_{c=1}^{|C|} 2 \times \frac{Precision_c \times Recall_c}{Precision_c + Recall_c} \tag{13}$$

Statistical significance of differences in the performance between models was tested using paired t-tests with alpha level equal to 0.05. Results are presented in Figure 11.

Table 10 Results of experiments averaged over 5 runs in the large-scale study. Depicted mean ± standard deviations in parentheses

Model	Precision	Recall	F_1-micro	F_1-macro
PM	**0.758** (± 0.007)	**0.793** (± 0.005)	**0.775** (± 0.005)	**0.677** (± 0.006)
BERT	n/a	n/a	n/a	0.640 (n/a)
vector ids	0.602 (± 0.007)	0.685 (± 0.006)	0.641 (± 0.004)	0.537 (± 0.003)

Note. Best performance is depicted in bold font.

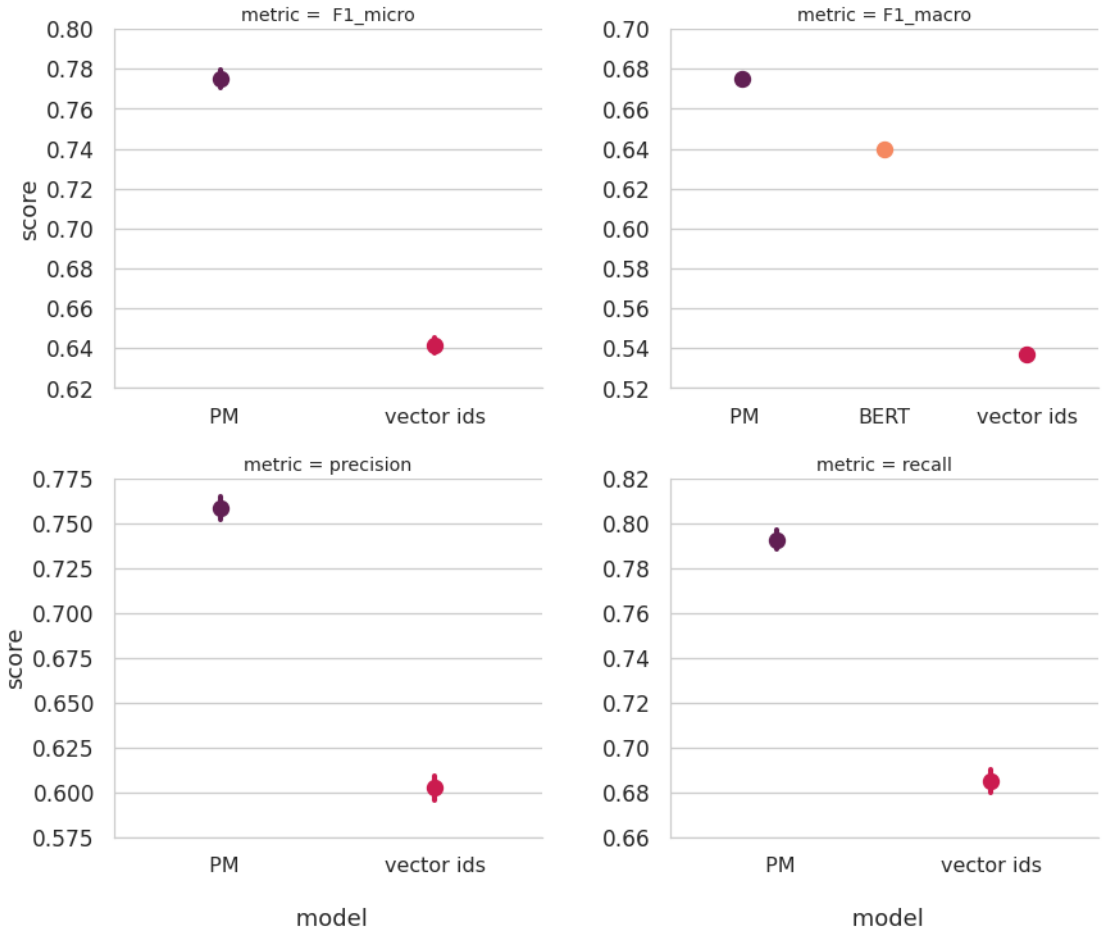

Fig. 11 Comparative analysis of results from the large-scale study. Depicted are the average scores from 5 different random splits of data in the case of PM and vector-ids models. Results for BERT are reported from Demszky et al. (2020). Vertical bars indicate standard deviations

In regard to emotion recognition at the individual level, my PM model performs better in terms of all 4 metrics than the vector-ids model ($p < 0.001$). What is more, my PM model achieves state-of-the-art performance — 0.68 macro-averaged F_1 score, which is 4 percentage points higher than the BERT model fine-tuned by the authors of the GoEmotions dataset (Demszky et al., 2020).

5 Discussion

Results of both studies conducted in the current work indicate that the proposed approach to emotion recognition in text could compete with state-of-the-art systems introduced in other works. My CNN-BiLSTM model that utilises additional information about raters from the PM metric achieves superior performance not only in regard to majority-based approaches represented in this work by Demszky et al. (2020)'s BERT model (see Table 10) and CNN-BiLSTM trained in a standard fashion (see Table 6) but also alternative solutions to personalised classifiers (see Table 10). In addition, results of the small-scale study indicate that models developed on majority-aggregated data performs poorly when employed for detection of emotions on the individual level (*cross* condition in the small-scale study), which supports previous findings (Gordon et al., 2021). In regard to the alternative techniques in personalised emotion recognition, simple addition of a numerical identifier of an annotator as an additional feature fed to the CNN-BiLSTM model is not enough to achieve good performance (*vector ids* the large-scale study), which has been shown by other studies as well (Kocoń et al., 2021). Meta-data about raters seems to boost baseline performance (Kocoń et al., 2021), similarly adding demographic features improves the results of deep learning models (Miłkowski et al., 2021). However, approaches that model rater behaviour such as my PM metric or PEB feature (Miłkowski et al., 2021) seem to be superior compared to the alternatives.

The value of the proposed Personalisation Metric has been proven in two studies that comprise the current work – the small-scale study and the large-scale study. The former comprises of fewer data points a model could learn from compared to the large-scale study (9,000 vs. over 120,000 individual text samples), as well as smaller number of raters the model needs to predict a label for (3 vs. 61, respectively). In the small-scale study however, every piece of text is labelled by each of 3 raters, in the large-scale study individual rate not only different text samples but also different number of text samples (minimum 334, maximum 6249). Then, the level of agreement in the text annotation between raters is higher in the small-scale study compared to the large-scale study (κ=0.5-0.6 vs. Spearman's r=0.2-0.6, respectively), which suggest that there are smaller differences in regard to the emotion perception a model needs to learn (in other words, there is a similar tendency among raters to view particular texts as negative or positive, for example). The number of categories a model needs to predict also differs between the two studies (4 vs. 7 categories, respectively). Although the two task then in both studies is a multi-category prediction, classification in the large-scale study is multi-label, in addition. That is, more than one label could be assigned, and therefore needs to predicted, by a deep learning model. Imbalanced class distribution is observed in the case of both datasets, however it is much more pronounced in the GoEmotions dataset employed for the classification task in the large-scale study. Therefore, the task of emotion recognition in both studies are challenging, because of different reasons though.

The over one hundred-year debate on the nature of emotions shows the complexity of the human behaviour and resulting challenges in studying it. Throughout the years, many psychologist and philosophers introduced different views, classifications and processes underlying emotional experience. The most prominent theories are interchangeably employed in emotion research – in regard to the dimensional view on the nature of emotions, Russell's

circumplex model with two dimensions – valence and arousal – is especially popular in lexicon-based sentiment analysis (Warriner et al., 2013; Mohammad, 2018a). In regard to categorical accounts, Ekman's model of 5 (or 6) basic emotions is commonly utilised in many tasks on emotion recognition in text (Demszky et al., 2020; Davani et al., 2022; Bharti et al., 2022).

Emotion recognition from text caught the attention of scientific researchers as well as private companies for the commercial purposes. In regard to the former, inference regarding emotions states of individuals aids professionals in detection of health problems, especially related to mental disorders such as depression and anxiety (Zeberga et al., 2022). In regard to the latter, automatic classification of emotion could be utilised in therapeutic chatbots to aid human therapists via computer-aided solutions (Zygadło et al., 2021). Forecasting election results (Anstead & O'Loughlin, 2015), analysis of public opinion (on controversial/hot issues such as immigration (Kopacheva & Yantseva, 2022)), and prediction of stock market changes (Pimprikar et al., 2017) are another area of application of automatic emotion recognition tools.

Two main approaches to emotion recognition in natural language processing could be distinguished: lexicon-based and machine learning. The former is a well studied method in psychology – affective lexicons of emotion-laden words are usually created by social scientists that gather a large sample of participants from the general population in order to collect affective ratings of words. This kind of studies follow psychological research standards and methodology, where a word functions as a stimulus that could induce a response in its recipient. On the other hand, machine learning approach to emotion analysis in text is preferred by computer scientists. Here, the basic unit evaluated in terms of expressed emotions is usually a sentence or a post (like tweet or Reddit comment) instead of a word in lexicon-based approach. The annotation process is similar, however a smaller group of raters is asked to annotate instances of text and it is rather conducted online via dedicated crowdsourcing platforms such as Amazon Mechanical Turk. As a result, insight into the process of annotation – even basic statistics about the group of annotators – is often restricted. With the rise of online social media sites and big data research the idea of distant supervision is sometimes employed as an alternative for manual annotation for gathering annotated data for the training purposes of machine learning models. Here, developers of such a dataset utilise the use of hashtags in social media posts which suggest labels for the content written in these posts (for example, hashtags that are the names of emotions). A manual validation on a subsample of collected data is often employ to assess the quality of such datasets.

However, systematically achieved low inter-annotator-agreement coefficients in data annotation indicates there is no objective "ground truth" when it comes to emotion recognition from text by different individuals. Consensus is hardly ever achieved, usually only on a subset of labels from all annotators. Yet, computer scientist attempt at developing state-of-the-art classifiers, which is successful only in term of technical performance and a particular training dataset. However, these models perform rather poorly when applied to predict individuals' preferences in reality (Gordon et al., 2021). Therefore, a gap between the system performance on individual datasets and real-life applications is observed. Two questions posed in the previous sections of the current work – *Is everyone taken into account?* in a majority-voting

scheme, and *Whose perspective is represented in the end?* after label averaging – cannot be answered affirmatively in the light of existing literature on bias observed in aggregated corpora (Sap et al., 2022; Waseem, 2016) as well as results of the current work.

Although the traditional approach to text classification works well for the majority of people, the alternative acknowledges those individuals that do not have enough representation in the majority perspective to work satisfactorily for them. The alternative human-centred (Kocoń et al., 2021) or perspectivist (Abercrombie et al., 2022) approach to text classification introduced in recent years accounts for natural differences between individuals on the issue of interest. Other researchers provide solutions with the use of demographic information (Hovy, 2015) or meta-data about users (Kocoń et al., 2021), in the current thesis I propose the one based on the annotation behaviour of an individual. Nonetheless, the objective of this strand of research is the same – to treat disagreement in content annotation as a source of information instead of noise. The researcher ensures every individual has a representation of their perspective in data annotation process here so as to achieve not only superior technical (quantitative) but also practical (qualitative) performance. The alternative approach is therefore inclusive of different perspectives.

The focus of the current work is implementation of this approach in emotion research, however, studies show it can be successfully adopted in other task as well, in particular the ones that are highly subjective in nature such as offensive language detection (Kocoń et al., 2021; Gajewska, 2023; Gajewska & Konat, 2023). Gajewska & Konat (2023) demonstrate that the CNN-BiLSTM model with PM metric achieved superior performance on 5-categorical imbalanced dataset of abusive language detection, compared to the majority and cross-approach (which directly translates experimental conditions designed for the small-scale study in the current work). Moreover, Gajewska (2023) shows that the advocated personalised approach to text classification and the introduced PM metric could be translated to similar classification tasks as the SemEval-2023 shared task on learning with disagreements (Le-Wi-Di) Leonardelli et al. (2023). Here, the extended version of the PM metric (which included two additional measures – precision and recall – besides Cohen's kappa coefficient and percentage agreement) has proved useful in the case of binary classification in short texts – the proposed system was ranked 6th out of 30 teams participating in the task. In particular, it was utilised for the most challenging sub-task – detection of offensiveness in the MD-Agreement dataset (Leonardelli et al., 2021), which comprises of disaggregated data annotations from over 800 raters. The proposed approach caught the attention of the task organisers as most teams disregarded the information about individual raters in their systems, and the team represented by the author of the current work (team eevvgg) was among the two teams that leveraged this data in the training process.

Results of my study indicate that it is beneficial to adopt the personalised approach to detection of highly subjective phenomena such as emotions in NLP. At the same time, this could be achieved at low cost and with the use of simple methods that does not require to collect any personal data about individuals. My study shows that adding just two features which quantify the similarity of opinions between an individual and the majority in terms of emotion recognition improves the performance of deep learning classifiers and the quality

of predictions. With the use of this method I was able to develop a personalised emotion detection algorithm that achieves state-of-the-art performance and outperforms Transformer-based architectures such as BERT. It proved useful in cases with a small training sample (experiments in the small-scale study), and highly imbalanced distributions of predicted categories (experiments in the large-scale study).

Recent studies on automated classification of persuasive messages (Lukin et al., 2017) makes my Personalisation Metric suitable even for the detection of persuasion of certain individuals. Here, Al Khatib et al. (2020) demonstrate that personal characteristics of debaters have an impact on perceived persuasiveness of a message as well as resistance to persuasive strategies, for example.

A number of different features, algorithms, and datasets has been proposed in emotion recognition research in recent years. Analysis of results achieved by these systems allows me to conclude that the proposed CNN-BiLSTM model trained in a personalised fashion could be regarded as state-of-the-art solution in emotion detection from text. Its performance is in line or superior to other deep learning and Transformers-based systems, which summary is presented below:

- the model developed by Bhowmick et al. (2010) scored 0.60 on micro-averaged F_1 for the classification of 4 emotions (disgust-anger, fear, happiness, sadness);
- Dini & Bittar (2016)'s machine learning classifier achieved 0.58 in terms of F_1 given 6 emotional categories;
- Dini & Bittar (2016)'s aggregated multi-label classifier (comprised of 7 independent binary SVM models, one for each of the 7 emotions) working at a tweet-level scored macro-averaged F_1 of 0.67;
- BLSTM-MC model (Wang et al., 2018) for the multi-emotion classification of code-switching texts scored 0.47 macro-averaged F_1;
- Ibrahiem et al. (2019)'s CNN architecture evaluated separately for each of the 4 emotions (anger, fear, joy, sadness) achieved between 0.15 and 0.75 in terms of F_1 metric;
- Ahmed et al. (2020) propose attention-based CNN-BiLSTM architecture that achieved 0.80 and 0.78 in terms of accuracy and F_1, respectively, in classification of 4 emotions (anger, fear, sadness, joy)
- the use of BERT base model and transfer-learning technique allows Kumar et al. (2021a) to achieve 0.93 accuracy in regard to the sentence-level recognition of 6 categories of emotions (anger, fear, happy, love, sadness, and surprise)
- Bharti et al. (2022) propose a hybrid model that combines deep learning and machine learning algorithms, specifically CNN-BiGRU and SVM models - the former are used to encode text, and the latter (SVM) is fed with the obtained latent vector of features and used as a final classification layer algorithm. It achieves 0.80 accuracy (0.82 precision and 0.80 recall) in the classification of Ekman's 6 basic emotions on a combined ISEAR, WASSA, and the Emotion-Stimulus dataset.

- BERT and RoBERTa models trained by Zanwar et al. (2022) on the 6-class (anger, sadness, fear, disgust, joy, surprise) filtered GoEmotions dataset achieve 0.68 and 0.69 F_1 scores, respectively. Feeding these models with an additional set of 256 psycholinguistic features extracted from text boosts the performance to 0.70 and 0.71 F_1, respectively.

In regard to the first research question of the current thesis – *How could subjective factors of emotion recognition be modelled in algorithms?* – I propose the Personalisation Metric that models raters' annotation behaviour (to be precise, quantifies the individual's agreement of emotion recognition with the view of the majority on this issue) and which could be successfully incorporated into a vector of features fed to a machine learning model, which as a result leads to the improved performance of the automatic classifier compared to both the majority-based approach as well as other solutions that fit into the perspectivist agenda. Therefore, this answers the second research question – *Does inclusion of those factors in deep learning models lead to the improvement of their performance?* The proposed metric is simple – comprises of two well known coefficients employed for reliability assessment, i.e., Cohen's kappa and accuracy – yet effective – CNN-BiLSTM model that incorporates it achieves results in terms of macro-averaged F_1 that could be regarded as state-of-the-art, outperforming novel architectures and techniques in NLP such as BERT. Further comparison with other emotion detection systems allows for similar conclusions that the combination of the proposed PM metric and CNN-BiLSTM model achieves competitive technical performance in terms of accuracy and/or F_1 scores. In addition, the proposed solution yields improved qualitative performance as the labels returned from the proposed CNN-BiLSTM-PM model are tailored to the opinions of individuals, instead of prediction of only one "average" value.

References

Abdulla, N. A., Ahmed, N. A., Shehab, M. A., Al-Ayyoub, M., Al-Kabi, M. N. & Al-rifai, S. (2014). Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and Web Engineering (IJITWE)*, *9(3)*, 55–71.

Abercrombie, G., Basile, V., Tonelli, S., Rieser, V. & Uma, A. (2022). Proceedings of the 1st workshop on perspectivist approaches to NLP@ LREC2022. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*.

Acheampong, F. A., Nunoo-Mensah, H. & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, *54(8)*, 5789–5829.

Addi, H. A., Ezzahir, R. & Mahmoudi, A. (2020). Three-level binary tree structure for sentiment classification in Arabic text. In *Proceedings of the 3rd International Conference on Networking, Information Systems & Security* (pp. 1–8).

Agrawal, A. & An, A. (2012). Unsupervised emotion detection from text using semantic and syntactic relations. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Volume 1 (pp. 346–353).

Ahmed, S., Reyadh, A. S., Sithil, F. T., Shah, F. M. & Shaafi, A. I. (2020). An attention-based approach to detect emotion from tweets. In *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)* (pp. 182–187).

Akhtar, S., Basile, V. & Patti, V. (2020). Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Volume 8 (pp. 151–154).

Al Khatib, K., Völske, M., Syed, S., Kolyada, N. & Stein, B. (2020). Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7067–7072).

Al Kuwatly, H., Wich, M. & Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 184–190).

Al-Saqqa, S., Abdel-Nabi, H. & Awajan, A. (2018). A survey of textual emotion detection. In *2018 8th International Conference on Computer Science and Information Technology (CSIT)* (pp. 136–142).

Aljedaani, W., Abuhaimed, I., Rustam, F., Mkaouer, M. W., Ouni, A. & Jenhani, I. (2022). Automatically detecting and understanding the perception of COVID-19 vaccination: a middle east case study. *Social Network Analysis and Mining*, *12(1)*, 1–26.

Alm, C. O. (2011). Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 107–112).

Almatarneh, S. & Gamallo, P. (2017). Automatic construction of domain-specific sentiment lexicons for polarity classification. In *International Conference on Practical Applications of Agents and Multi-Agent Systems* (pp. 175–182).

Alswaidan, N. & Menai, M. E. B. (2020). A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, *62(8)*, 2937–2987.

Anstead, N. & O'Loughlin, B. (2015). Social media analysis and public opinion: The 2010 UK general election. *Journal of computer-mediated communication*, *20(2)*, 204–220.

Baccianella, S., Esuli, A. & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Volume 10 (pp. 2200–2204).

Bao, Y., Quan, C., Wang, L. & Ren, F. (2014). The role of pre-processing in Twitter sentiment analysis. In *Intelligent Computing Methodologies: 10th International Conference* (pp. 615–624).

Barrett, L. F. (2006a). Are emotions natural kinds? *Perspectives on psychological science*, *1(1)*, 28–58.

Barrett, L. F. (2006b). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, *10(1)*, 20–46.

Bassignana, E., Basile, V., Patti, V. et al. (2018). Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop Proceedings*, Volume 2253 (pp. 1–6).

Bharti, S. K., Varadhaganapathy, S., Gupta, R. K., Shukla, P. K., Bouye, M., Hingaa, S. K. & Mahmoud, A. (2022). Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience*.

Bhowmick, P. K., Basu, A. & Mitra, P. (2008). An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics* (pp. 58–65).

Bhowmick, P. K., Basu, A. & Mitra, P. (2010). Classifying emotion in news sentences: When machine classification meets human classification. *International Journal on Computer Science and Engineering*, *2(1)*, 98–108.

Binns, R., Veale, M., Van Kleek, M. & Shadbolt, N. (2017). Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics* (pp. 405–415).

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*, 4349–4357.

Bouazizi, M. & Ohtsuki, T. (2019). Multi-class sentiment analysis on Twitter: Classification performance and challenges. *Big Data Mining and Analytics*, *2(3)*, 181–194.

Boucher, J. & Osgood, C. E. (1969). The pollyanna hypothesis. *Journal of verbal learning and verbal behavior*, *8(1)*, 1–8.

Brosch, T., Pourtois, G. & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition and emotion*, *24(3)*, 377–400.

Calvo, R. A. & Mac Kim, S. (2013). Emotions in text: dimensional and categorical models. *Computational Intelligence*, *29(3)*, 527–543.

Cambria, E., Schuller, B., Xia, Y. & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, *28(2)*, 15–21.

Cambria, E., Speer, R., Havasi, C. & Hussain, A. (2010). Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI fall symposium series*.

Chachra, A., Mehndiratta, P. & Gupta, M. (2017). Sentiment analysis of text using deep convolution neural networks. In *2017 Tenth international conference on contemporary computing (IC3)* (pp. 1–6).

Chatterjee, A., Narahari, K. N., Joshi, M. & Agrawal, P. (2019). SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 39–48).

Chauhan, P., Sharma, N. & Sikka, G. (2021). The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, *12(2)*, 2601–2627.

Chen, X., Zhang, Y. & Wen, J.-R. (2022a). Measuring "why" in recommender systems: a comprehensive survey on the evaluation of explainable recommendation. *arXiv preprint arXiv:2202.06466*.

Chen, Y., Sack, H. & Alam, M. (2022b). Analyzing social media for measuring public attitudes toward controversies and their driving factors: a case study of migration. *Social Network Analysis and Mining*, *12(1)*, 1–27.

Chou, H.-C. & Lee, C.-C. (2019). Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5886–5890).

Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S. & Danforth, C. M. (2015). Climate change sentiment on Twitter: An unsolicited public opinion poll. *PloS one*, *10(8)*, e0136092.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20(1)*, 37–46.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A. & Menczer, F. (2011). Predicting the political alignment of Twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing* (pp. 192–199).

Cortiz, D. (2021). Exploring transformers in emotion recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA. *arXiv preprint arXiv:2104.02041*.

Cowen, A., Sauter, D., Tracy, J. L. & Keltner, D. (2019). Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, *20(1)*, 69–90.

Davani, A. M., Díaz, M. & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, *10*, 92–110.

Davidson, T., Bhattacharya, D. & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 25–35).

Dawid, A. P. & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28(1)*, 20–28.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G. & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4040–4054).

Deng, J. & Ren, F. (2021). A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT* (pp. 4171–4186). Association for Computational Linguistics.

Dey, A., Jenamani, M. & Thakkar, J. J. (2018). Senti-N-Gram: An n-gram lexicon for sentiment analysis. *Expert Systems with Applications*, *103*, 92–105.

Diaz-Aviles, E., Stewart, A., Velasco, E., Denecke, K. & Nejdl, W. (2012). Epidemic intelligence for the crowd, by the crowd. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 6 (pp. 439–442).

Dini, L. & Bittar, A. (2016). Emotion analysis on Twitter: the hidden challenge. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3953–3958).

Durmus, E. & Cardie, C. (2019). Exploring the role of prior beliefs for argument persuasion. *arXiv preprint arXiv:1906.11301*.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6(3-4)*, 169–200.

Ekman, P. (1999). Basic emotions. *Handbook of Cognition and Emotion* (pp. 45–60).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76(5)*, 378.

Fornaciari, T., Bianchi, F., Nozza, D. & Hovy, D. (2021). MilaNLP@ WASSA: Does BERT feel sad when you cry? In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 269–273).

Frénay, B. & Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, *25(5)*, 845–869.

Frijda, N. H. (2017). *The laws of emotion*. Psychology Press.

Gajewska, E. (2023). eevvgg at SemEval-2023 task 11: Offensive language classification with rater-based information. In *Proceedings of the 17th International Workshop on Semantic Evaluation*. Toronto, Canada. Accepted.

Gajewska, E. & Konat, B. (2023). Text classification for subjective phenomena on disaggregated data and rater behaviour. In Z. Vetulani & P. Paroubek (Eds.), *Human Language Technologies as a Challenge for Computer Science and Linguistics* (pp. 73–77). Wydawnictwo Naukowe UAM.

Gallagher, C., Furey, E. & Curran, K. (2019). The application of sentiment analysis and text analytics to customer experience reviews to understand what customers are really saying. *International Journal of Data Warehousing and Mining (IJDWM)*, *15(4)*, 21–47.

Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115(16)*, E3635–E3644.

Gatti, L., Guerini, M. & Turchi, M. (2015). Sentiwords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, *7(4)*, 409–421.

Gitari, N. D., Zuping, Z., Damien, H. & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10(4)*, 215–230.

Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T. & Bernstein, M. S. (2021). The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).

Grimm, M., Kroschel, K., Mower, E. & Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech communication*, *49(10-11)*, 787–800.

Hameed, Z. & Garcia-Zapirain, B. (2020). Sentiment classification using a single-layered BiLSTM model. *Ieee Access*, *8*, 73992–74001.

Hofmann, J., Troiano, E. & Klinger, R. (2021). Emotion-aware, emotion-agnostic, or automatic: Corpus creation strategies to obtain cognitive event appraisal annotations. *arXiv preprint arXiv:2102.12858*.

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R. & Weber, R. (2021). The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, *53*, 232–246.

Houghton, K. J., Upadhyay, S. S. N. & Klin, C. M. (2018). Punctuation in text messages may convey abruptness. period. *Computers in Human Behavior*, *80*, 112–121.

Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (pp. 752–762).

Hovy, D., Bianchi, F. & Fornaciari, T. (2020). "You sound just like your father" Commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1686–1690).

Hovy, D. & Yang, D. (2021). The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 588–602).

Hu, M. & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168–177).

Hutto, C. & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 8.

Ibrahiem, S., Ismail, S., Bahnasy, K. & Aref, M. (2019). Convolutional neural network multi-emotion classifiers. *Jordanian Journal of Computers and Information Technology*, *5(2)*.

Jain, V. K., Kumar, S. & Fernandes, S. L. (2017). Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. *Journal of computational science*, *21*, 316–326.

Jakobsen, T. S. T., Barrett, M., Søgaard, A. & Lassen, D. (2022). The sensitivity of annotator bias to task definitions in argument mining. In *Proceedings of the 16th Lingusitic Annotation Workshop (LAW-XVI) within LREC2022* (pp. 44–61).

James, W. (1884). What is an emotion? *Mind*, *9(34)*, 188–205.

Jingar, M. & Lindgren, H. (2019). Tangible communication of emotions with a digital companion for managing stress: an exploratory co-design study. In *Proceedings of the 7th International Conference on Human-Agent Interaction* (pp. 28–36).

Kardakis, S., Perikos, I., Grivokostopoulou, F. & Hatzilygeroudis, I. (2021). Examining attention mechanisms in deep learning models for sentiment analysis. *Applied Sciences*, *11(9)*, 3883.

Kateb, F. & Kalita, J. (2015). Classifying short text in social media: Twitter as case study. *International Journal of Computer Applications*, *111(9)*, 1–12.

Khosla, S., Chhaya, N. & Chawla, K. (2018). Aff2Vec: Affect–enriched distributional word representations. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2204–2218).

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751). Doha, Qatar: Association for Computational Linguistics.

Klenner, M., Göhring, A., Amsler, M., Ebling, S., Tuggener, D., Hürlimann, M. & Volk, M. (2020). Harmonization sometimes harms. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) 16th Conference on Natural Language Processing*. SwissText/KONVENS 2020.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K. et al. (2023). Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T. & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, *58(5)*, 102643.

Kocoń, J., Janz, A. & Piasecki, M. (2018). Context-sensitive sentiment propagation in wordnet. In *Proceedings of the 9th Global Wordnet Conference* (pp. 329–334).

Kopacheva, E. & Yantseva, V. (2022). Users' polarisation in dynamic discussion networks: The case of refugee crisis in sweden. *Plos one*, *17(2)*, e0262992.

Kramer, A. D. (2010). An unobtrusive behavioral model of "gross national happiness". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 287–290).

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology.* Sage publications.

Kumar, A., Cambria, E. & Trueman, T. E. (2021a). Transformer-based bidirectional encoder representations for emotion detection from text. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–6).

Kumar, D., Kelley, P. G., Consolvo, S., Mason, J., Bursztein, E., Durumeric, Z., Thomas, K. & Bailey, M. (2021b). Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)* (pp. 299–318).

Lakoff, G. (1990). The invariance hypothesis: is abstract reason based on image-schemas? *Cognitive Linguistics*, *1(1)*.

Larimore, S., Kennedy, I., Haskett, B. & Arseniev-Koehler, A. (2021). Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media* (pp. 81–90).

Lee, J. & Lee, W. (2022). CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5669–5679).

Leonardelli, E., Menini, S., Palmero Aprosio, A., Guerini, M. & Tonelli, S. (2021). Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10528–10539).

Leonardelli, E., Uma, A., Abercrombie, G., Almanea, D., Basile, V., Fornaciari, T., Plank, B., Rieser, V. & Poesio, M. (2023). Semeval-2023 task 11: Learning with disagreements (lewidi). Toronto, Canada.

Lewis, M., Haviland-Jones, J. M. & Barrett, L. F. (2010). *Handbook of emotions.* Guilford Press.

Liapis, C. M., Karanikola, A. & Kotsiantis, S. (2021). A multi-method survey on the use of sentiment analysis in multivariate financial time series forecasting. *Entropy*, *23(12)*, 1603.

Litman, D. & Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)* (pp. 351–358).

Liu, B. & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). Springer.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Longpre, L., Durmus, E. & Cardie, C. (2019). Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*.

López-Chau, A., Valle-Cruz, D. & Sandoval-Almazán, R. (2020). Sentiment analysis of Twitter data through machine learning techniques. In *Software Engineering in the Era of Cloud Computing* (pp. 185–209). Springer.

Lukin, S., Anand, P., Walker, M. & Whittaker, S. (2017). Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 742–753).

Luo, Y., Card, D. & Jurafsky, D. (2020). Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3296–3315).

Mao, X., Chang, S., Shi, J., Li, F. & Shi, R. (2019). Sentiment-aware word embedding for emotion classification. *Applied Sciences*, *9(7)*, 1334.

Marrero-Fernández, P., Montoya-Padrón, A., Jaume-i Capó, A. & Buades Rubio, J. M. (2014). Evaluating the research in automatic emotion recognition. *IETE Technical Review*, *31(3)*, 220–232.

Maynard, D., Bontcheva, K. & Rout, D. (2012). Challenges in developing opinion mining tools for social media. *Proceedings of @NLP Can U Tag User-generated-content?! Workshop at LREC 2012*.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miłkowski, P., Gruza, M., Kanclerz, K., Kazienko, P., Grimling, D. & Kocoń, J. (2021). Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop* (pp. 248–259).

Mohammad, S., Dunne, C. & Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 599–608).

Mohammad, S. & Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 26–34).

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement* (pp. 201–237). Elsevier.

Mohammad, S. M. (2018a). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*. Melbourne, Australia.

Mohammad, S. M. (2018b). Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*. Miyazaki, Japan.

Mohammad, S. M. & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29(3)*, 436–465.

Mohapatra, S., Ahmed, N. & Alencar, P. (2019). KryptoOracle: a real-time cryptocurrency price prediction platform using Twitter sentiments. In *2019 IEEE international conference on big data (Big Data)* (pp. 5544–5551).

Mondal, M., Silva, L. A. & Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media* (pp. 85–94).

Monnier, C. & Syssau, A. (2014). Affective norms for French words (FAN). *Behavior research methods*, *46(4)*, 1128–1137.

Nedjah, N., Santos, I. & de Macedo Mourelle, L. (2019). Sentiment analysis using convolutional neural network via word embeddings. *Evolutionary Intelligence* (pp. 1–25).

Nozza, D., Fersini, E. & Messina, E. (2017). A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 273–280).

Oatley, K., Keltner, D. & Jenkins, J. M. (2006). *Understanding emotions*. Blackwell publishing.

Ortony, A., Clore, G. L. & Collins, A. (1990). *The cognitive structure of emotions*. Cambridge university press.

Ortony, A. & Turner, T. J. (1990). What's basic about basic emotions? *Psychological review*, *97(3)*, 315.

Ouyang, X., Zhou, P., Li, C. H. & Liu, L. (2015). Sentiment analysis using convolutional neural network. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing* (pp. 2359–2364).

Pennebaker, J. W., Francis, M. E. & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71(2001)*, 2001.

Pennebaker, J. W., Mehl, M. R. & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54(1)*, 547–577.

Pennington, J., Socher, R. & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2227–2237).

Pimprikar, R., Ramachadran, S. & Senthilkumar, K. (2017). Use of machine learning algorithms and Twitter sentiment analysis for stock market prediction. *Int J Pure Appl Math*, *115(6)*, 521–526.

Plutchik, R. (1982). A psychoevolutionary theory of emotions. *Social Science Information*, *21(4-5)*.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, *89(4)*, 344–350.

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association.

Prabhakaran, V., Davani, A. M. & Diaz, M. (2021). On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop* (pp. 133–138).

Ren, J. & Nickerson, J. V. (2014). Online review systems: How emotional language drives sales. In *Twentieth Americas Conference on Information Systems*.

Rizos, G. & Schuller, B. W. (2020). Average Jane, where art thou? – recent avenues in efficient machine learning under subjectivity uncertainty. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 42–55).

Rodriguez, A., Argueta, C. & Chen, Y.-L. (2019). Automatic detection of hate speech on Facebook using sentiment and emotion analysis. In *2019 international conference on artificial intelligence in information and communication (ICAIIC)* (pp. 169–174).

Röttger, P., Vidgen, B., Hovy, D. & Pierrehumbert, J. B. (2021). Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39(6)*, 1161.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, *110(1)*, 145.

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A. & Choi, G. S. (2021). A performance comparison of supervised machine learning models for COVID-19 tweets sentiment analysis. *Plos one*, *16(2)*, e0245909.

Saganowski, S., Komoszyńska, J., Behnke, M., Perz, B., Kunc, D., Klich, B., Kaczmarek, Ł. D. & Kazienko, P. (2022). Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific data*, *9(1)*, 1–11.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y. & Smith, N. A. (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5884–5906).

Scherer, K. R. (1987). Toward a dynamic theory of emotion: The component process model of affective states. *Geneva studies in Emotion and Communication*, *1*, 1–98.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, *23(7)*, 1307–1351.

Scherer, K. R. & Ekman, P. (2014). *Approaches to emotion*. Psychology Press.

Schober, M. F. (1992). Asking questions and influencing answers. In Tanur, J. M. (Ed.), *Questions about Questions: Inquiries into the Cognitive Bases of Surveys* (pp. 15–48). Russell Sage Foundation.

Seyeditabari, A., Tabari, N. & Zadrozny, W. (2018). Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Sheng, E., Chang, K.-W., Natarajan, P. & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3407–3412).

Sheng, V. S. (2011). Simple multiple noisy label utilization strategies. In *2011 IEEE 11th International Conference on Data Mining* (pp. 635–644).

Sheng, V. S., Provost, F. & Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 614–622).

Shmueli, B. & Ku, L.-W. (2019). SocialNLP EmotionX 2019 challenge overview: Predicting emotions in spoken dialogues and chats. *arXiv preprint arXiv:1909.07734*.

Shuman, V. & Scherer, K. R. (2014). Concepts and structures of emotions. *International handbook of emotions in education* (pp. 23–45).

Sintsova, V., Musat, C. & Pu, P. (2014). Semi-supervised method for multi-category emotion recognition in tweets. In *2014 IEEE International Conference on Data Mining Workshop* (pp. 393–402).

Smith, R., Killgore, W. D. & Lane, R. D. (2018). The structure of emotional experience and its relation to trait emotional awareness: A theoretical review. *Emotion*, *18(5)*, 670–692.

Stanovsky, G., Smith, N. A. & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1679–1684).

Stone, P., Dunphy, D., Smith, M. & Ogilvie, D. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, Cambridge.

Strapparava, C., Valitutti, A. et al. (2004). Wordnet-affect:: an affective extension of WordNet. In *LREC*, Volume 4 (p.40̃).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37(2)*, 267–307.

Tam, S., Said, R. B. & Tanriöver, Ö. Ö. (2021). A ConvBiLSTM deep learning model-based approach for Twitter sentiment classification. *IEEE Access*, *9*, 41283–41293.

Tang, D., Wei, F., Qin, B., Yang, N., Liu, T. & Zhou, M. (2015). Sentiment embeddings with applications to sentiment analysis. *IEEE transactions on knowledge and data Engineering*, *28(2)*, 496–509.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B. et al. (2014). Learning sentiment-specific word embedding for Twitter sentiment classification. In *ACL (1)* (pp. 1555–1565).

Tesfagergish, S. G., Kapočiūtė-Dzikienė, J. & Damaševičius, R. (2022). Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning. *Applied Sciences*, *12(17)*, 8662.

Tsapatsoulis, N. & Djouvas, C. (2019). Opinion mining from social media short texts: Does collective intelligence beat deep learning? *Frontiers in Robotics and AI* (p. 138).

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5(2)*, 207–232.

Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E. & Poesio, M. (2021a). Semeval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 338–347).

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B. & Poesio, M. (2021b). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, *72*, 1385–1470.

Vo, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J. & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior research methods*, *41(2)*, 534–538.

Wang, N., Kosinski, M., Stillwell, D. & Rust, J. (2014). Can well-being be measured using Facebook status updates? Validation of Facebook's Gross National Happiness Index. *Social Indicators Research*, *115(1)*, 483–491.

Wang, T., Yang, X., Ouyang, C., Guo, A., Liu, Y. & Li, Z. (2018). A multi-emotion classification method based on BLSTM-MC in code-switching text. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 190–199).

Wang, X., Jiang, W. & Luo, Z. (2016). Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 2428–2437).

Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J. & Yu, Z. (2019). Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5635–5649).

Wankhade, M., Rao, A. C. S. & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review* (pp. 1–50).

Warriner, A. B., Kuperman, V. & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, *45(4)*, 1191–1207.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138–142).

Welch, C., Kummerfeld, J. K., Pérez-Rosas, V. & Mihalcea, R. (2020). Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4076–4089).

Wich, M., Bauer, J. & Groh, G. (2020). Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 54–64).

Wierzba, M., Riegel, M., Wypych, M., Jednoróg, K., Turnau, P., Grabowska, A. & Marchewka, A. (2015). Basic emotions in the Nencki Affective Word List (NAWL BE): New method of classifying emotional stimuli. *PLOS ONE*, *10(7)*, e0132305.

Wilson, T., Wiebe, J. & Hwa, R. (2004). Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of the 19th national conference on Artifical intelligence* (pp. 761–767).

Wong, F. M. F., Tan, C. W., Sen, S. & Chiang, M. (2016). Quantifying political leaning from tweets, retweets, and retweeters. *IEEE transactions on knowledge and data engineering*, *28(8)*, 2158–2172.

Wulczyn, E., Thain, N. & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web* (pp. 1391–1399).

Zahiri, S. M. & Choi, J. D. (2017). Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv preprint arXiv:1708.04299*.

Zanwar, S., Wiechmann, D., Qiao, Y. & Kerz, E. (2022). Improving the generalizability of text-based emotion detection by leveraging transformers with psycholinguistic features. *arXiv preprint arXiv:2212.09465*.

Zeberga, K., Attique, M., Shah, B., Ali, F., Jembre, Y. Z. & Chung, T.-S. (2022). A novel text mining approach for mental health prediction using bi-lstm and bert model. *Computational Intelligence and Neuroscience*, *2022*.

Zec, S., Soriani, N., Comoretto, R. & Baldi, I. (2017). High agreement and high prevalence: The paradox of cohen's kappa. *The Open Nursing Journal*, *11*, 211–218.

Zhou, C., Sun, C., Liu, Z. & Lau, F. (2015). A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630*.

Zhou, D., Zhang, X., Zhou, Y., Zhao, Q. & Geng, X. (2016). Emotion distribution learning from texts. In *EMNLP* (pp. 638–647).

Zygadło, A., Kozłowski, M. & Janicki, A. (2021). Text-based emotion recognition in english and polish for therapeutic chatbot. *Applied Sciences*, *11(21)*, 10146.