Text Mining for Online Mental Health State and Personality Assessment

Doctoral Dissertation submitted to the
Faculty of Informatics of the Università della Svizzera Italiana
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by
Esteban A. Ríssola

under the supervision of
Fabio Crestani

September 2021

Dissertation Committee

Cesare Alippi	Università della Svizzera italiana, Switzerland
Silvia Santini	Università della Svizzera italiana, Switzerland
Christina Lioma	University of Copenhagen, Denmark
Jacques Savoy	University of Neuchâtel, Switzerland

Dissertation accepted on 09 September 2021

Research Advisor	PhD Program Director
Fabio Crestani	**Walter Binder**

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Esteban A. Ríssola
Lugano, 09 September 2021

We can only see a short distance ahead, but we can see plenty there that needs to be done.

Alan Turing

Abstract

Advances on psycho-linguistics have evidenced that the ways in which people use words could act as a reliable source to assess a wide array of behaviours. Language use acts as an indicator of the individuals' current mental state, personality and even personal values. In this thesis, we focus on language analysis to study two closely related processes which encompass integral components of persons' psychological profiles: mental health state and personality.

Mental health state assessment by analysing online user-generated content is a field that has recently attracted considerable attention. We start this dissertation by analysing the online digital traces left by individuals in order to ascertain their mental state condition at a particular point in time. To this aim, we exploit the latent semantic structure of social media users posts to spot early traces of depression. Next, present a weak-supervision framework to derive large quantities of data for the study of depression on online settings. Moreover, we conduct a series of analytical studies aimed at gaining insights and extending the current knowledge on how mental disorders are manifested through language and online behaviour in order to be able to detect the early onset of such disorders.

While the mental state condition of individuals may fluctuate over their lives, there is a core set of patterns concerning thought, affect and behaviour which is consistent across time and context, constituting the basis of what is commonly referred to as personality. In the second part of this dissertation, we focus on the computational assessment of personality from language cues. We present a novel approach to personality recognition in conversations based on capsule neural networks and exploit its inherent interpretability potential to gain insights from its inner functioning. Moreover, we propose a novel open-vocabulary approach based on multiword expressions which aims at discovering distinctive linguistic patterns of a personality trait. Such technologies will open new avenues to building more empathetic and naturalistic conversational systems.

Acknowledgements

First and foremost, I would like to express my deepest gratitude and appreciation to my parents for their unlimited support throughout my Ph.D and for motivating and encouraging me to pursue my goals at every stage of my life.

I am sincerely grateful to my advisor Prof. Fabio Crestani who gave me the opportunity to join his research group and, from the very beginning of my Ph.D., provided me with his constant support. Fabio taught me how to be an independent researcher and gave me many great opportunities to grow as a researcher and as a person. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the my dissertation committee: Prof. Christina Lioma, Prof. Silvia Santini, Prof. Jacques Savoy, and Prof. Cesare Alippi for their insightful comments and encouragement.

I would also like to express my gratitude to my collaborators (former and current), who helped me throughout my Ph.D. via countless meetings and stimulating discussions that we had. In particular, I am grateful to Prof. David E. Losada, Prof. Gabriel H. Tolosa, Dr. Ana Freire, Diana Ramírez-Cifuentes and David Solans. I am equally grateful to my group colleagues (and friends) for the constructive discussions, for their constant support, for the sleepless nights we were working together before deadlines, and for all the fun we shared over the last few years. In particular, I would like to thank Dr. Ali Bahrainian, Dr. Mohammad Aliannejadi, Manajit Chakraborty, Dr. Maram Barifah, Dr. Monica Landoni and Dr. Anastasia Giachanou.

I gratefully acknowledge the financial support of the Federal Commission for Scholarships for Foreign Students (FCS) of Switzerland and the Hasler Foundation.

Last but not least, I thank my friends, specially the closest ones, who never stopped believing in me. I appreciate how much they supported and encouraged me, specially during times of elevated emotional turmoil.

Contents

Figures

Tables

Chapter 1

Introduction

1.1 Motivations

Due to the worldwide accessibility to the Internet along with the continuous advances on mobile technologies, physical and digital worlds have become completely blended and the proliferation of social media platforms has taken a leading role over this evolution [148]. As a result, during the last decade there has been increasing research interest in the identification of mental state alterations through the analysis of online digital traces.

According to the World Health Organisation (WHO), *mental health disorders* (or simply, mental disorders) encompass a wide array of issues with diverse symptoms. In general, they are characterised by the presence of unusual behaviour, emotions and thoughts[1]. Depression, schizophrenia, disorders due to drug abuse and eating disorders are just some examples. In the past year, around 83 million people (between 18-65) have been subjected to one or more mental disorders in the European Economic Area (EEA) countries[2]. Furthermore, the lack of an appropriate treatment can lead to disability, psychotic episodes, thoughts of self-harm and, at its worst, suicide. For this reason, it is important to identify the onset of these kind of mental disruptions at the early stages in order to avoid any undesirable consequences.

Constraints dictated in real-life settings, such as cost and time, hinder the possibility to achieve a timely and effective personal diagnosis. Initiatives such as the Strategic Workshop on Information Retrieval in Lorne [48] (SWIRL) have already proposed the application of principles of core Information Retrieval for the development of decision-making systems applied to fields that years back

[1] See `http://www.who.int/mental_health/management/en/`

[2] See `https://bit.ly/3xq0an3`

were not easy to conceive or imagine. In particular, they highlight the potential for cross-disciplinary collaboration with a number of scientific fields, including psychology. For example, to develop novel and emerging applications related to psychological aspects on information access [103; 146].

Traditionally, mental health practitioners have collected and integrated information from various instruments to characterise the mental state of individuals. These include direct observation, focused questions on the current symptoms and formalised psychological tests. Such instruments have been used to assess several mental-related variables, such as the appearance, mood and attitudes of the subjects to determine the presence of any irregularity. The proliferation of online social media platforms is changing the dynamics in which mental health state assessment is performed [216; 33; 224]. Individuals are using these platforms on a daily basis to share their interests [246] and personal life events [104] as well as to disclose their feelings and moods [180; 43]. These platforms have become promising means to detect different mental disorders since the language employed as well as the emotions expressed in the text (e.g., social media *posts*) and shared with *followers* or *friends* on a daily basis may pinpoint feelings like worthlessness, guilt, or helplessness. This can provide a characterisation of symptoms of psychological disorders, like depression. As argued by Coppersmith et al. [44], there are many factors which motivate users to disclose on social media this sort of personal concerns, including: (a) to pursue or provide encouragement; (b) to counteract the stigma or *taboo* that mental disorders carry on the society (e.g., “mental disorders cannot be cured”) or; (c) to provide an explanation for certain behaviours.

By leveraging user-generated content, risk-assessment and decision-making technologies have the potential to make a significant difference by offering low-cost unobtrusive mechanisms to assist health practitioners in providing preliminary screening and awareness on mental health outcomes at a large scale [186; 243]. For instance, novel solutions involving the use of language technologies [35] have started to be considered in real-life settings. In this respect, research on language and psychology has shown that several useful cues about individuals' mental state (including social and emotional conditions) can be discovered by examining the patterns of their language use [39]. As a matter of fact, language attributes could act as indicators of a person's current mental state [147; 177], personality [157] and even personal values [22]. The main reason, as argued by Pennebaker et al. [172], is that such *latent mental-related variables are encoded in the words that individuals use to communicate*.

Language analysis usually focuses on one of two applications, *prediction* (typically, involving supervised machine learning algorithms) or *insight*. In prediction, the ultimate goal is to accurately map language into a single or few given outcomes. Insight is conceptually focused on the opposite, finding the language that is most characteristic of an outcome, i.e., identifying the patterns that provide a window into thoughts, attitudes, psychology or health of people. With this in mind, in this thesis we seek to move forward the state of the art on both applications.

The first goal of this thesis is to develop text mining models which are able to effectively extract psychological states. Such models could be then used to monitor states evolution from online user-generated content in order to spot well in advance the different variations of individuals' mental state that could suggest the onset of a disorder. Even though achieving an effective positive detection performance is important, we argue that tracking and visualising the development of the mental disorder is equally relevant. In fact, an accurate detection system can be *more useful if it provides a way of understanding the factors that lead to a certain detection*. Therefore, we aim at providing insights that could be used by a predictive system or a health practitioner in the elaboration of a diagnosis. We consider that research on this area should be steered toward building new metrics that could correlate with mental health concerns before traditional symptoms arise and which doctors could use as leading indicators of traditional later-onset symptoms.

While the mental state condition of individuals may fluctuate over different periods during their lives [221], there is a core set of patterns concerning thought (belief about self and others), affect (positive vs. negative approach) and behaviour (actions, fantasies and intention) which is relatively stable across time and context. Human beings greatly differ from one another in their way of thinking, feeling and hence acting, constituting the basis of what is commonly referred to as *personality*. Understanding human personality and how it can be automatically assessed using the different traces that individuals leave in their daily activities, though challenging, can be useful for a variety of potential applications, such as personalised product recommendation [227; 205], partner matching on dating websites [56], authorship attribution [127], customer service and human resources management.

As previously stated, a significant body of work on psycho-linguistics has provided evidence that the ways in which people use words could act as a reliable source to assess a wide array of behaviours [21]. Despite its complexity and ambivalence, language can be highly informative for the study of personality [157]. The main reason, as argued by Boyd et al. [20], is that language use is relatively reliable over time, consistent, and varies considerably among individuals. This inspired and motivated us to complement the assessment of individuals' mental health state through language analysis with the study of personality. Both elements are closely related, and encompass equally important and integral components of individuals' psychological profiles.

With the improvements in speech recognition [87] and voice generation technologies over the last years, several companies have sought to develop conversation understanding systems that run on mobile phones or smart home devices through natural language interfaces. Conversational agents, such as Amazon Alexa[3] or Google Assistant[4], are capable of understanding human speech and interact with users through synthesised voices. By incorporating an understanding of various aspects of a dialogue between a conversational agent and a human, including personality identification, more personalised and effective conversations could be achieved. The Web is a platform where conversational agents (i.e., chat-bots) are increasingly communicating with humans to provide automatic services. Such agents may be able to provide more effective services by improved online dialogues, if they take into account user profile information [247] and the personality traits of people who they are interacting with [30; 128].

Given the emergence of such powerful technological platforms supporting *individuation*[5] and its expression, we envision more empathetic and naturalistic conversational systems which, apart from interpreting and interacting with people, are able to adapt to their personality, akin to humans [64; 139] toward more engaging and effective dialogues. Research has shown that individuals are more engaged in interacting with others who have similar personality profiles as it requires less information processing and cognitive load [245]. In addition, adding personality traits to virtual agents leads to significantly better perceived emotional intelligence of such systems [128]. However, as argued by Stajner et al. [219] up to now conversational agents developers have mostly focused on methods for adapting such systems to users' emotions instead of deeper personality traits, probably due to insufficiently good performances of automatic personality detection systems on short utterances.

Driven by this, we take the first steps in this avenue and decide to advance the state of the art on personality assessment from conversations as the second goal of this thesis. Furthermore, we aim at gaining new insights on personality manifestation through language (in particular on conversations) which could be useful to build conversational agents that can incorporate such linguistics patterns into their dialogues. This is particularly beneficial on the Web, where con-

[3] See `https://amzn.to/3jUxZKT`

[4] See `https://assistant.google.com/`

[5] Individuation refers to the process through which a person achieves a sense of individuality separate from the identities of others and begins to consciously exist as a human in the world.

versational agents are increasingly communicating with humans to assist them through every day activities, like shopping, hotel reservations, and various other services.

1.2 Online Mental Health State Assessment

1.2.1 Mental Health State Assessment

According to the *Diagnostic and Statistical Manual of Mental Disorders* [8] (DSM)[6], a psychiatric taxonomy developed by the American Psychiatric Association (APA)[7], a mental disorder is formally defined as a:

> Syndrome characterised by a clinically significant disturbance in an individual's cognition, emotion regulation, or behaviour that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning.

Among the most common mental disorders there are depression, anxiety disorders, eating disorders (e.g., bulimia, anorexia), self-harm, bipolar affective disorder, post-traumatic stress disorder (PTSD), or schizophrenia. Figure 1.1 depicts a hierarchy of symptomatology of mental disorders. The goal of this pyramidal organisation of disorders is twofold: (a) first, the various disorders are ranked in terms of the severity of their symptoms and; (b) second, it explains that a disorder may show at some time any of the signs of those at lower levels, although these are not characteristic of that disorder. For instance, an individual suffering from depression might manifest at some point self-harming behaviour, which is a symptom typically associated to borderline personality disorder.

There is a great deal of discussion among health professionals on whether psychiatric disorders should be considered medical diseases. The rationale behind this claim is that no brain scans (e.g., MRI), laboratory tests or X-rays are able to confirm any mental disorder as a physical condition. However, recent advances in neuroscience have found that brain circuit functioning and connectivity of specific regions of the brain can be associated with mental disorders. Such discovery has been achieved through the use of brain imaging techniques

[6] The DSM determines a common vocabulary and standard criteria to group and characterise the different mental disorders. Its three main components are: the diagnostic classification, the diagnostic criteria sets and the descriptive text.

[7] See `https://www.psychiatry.org/`

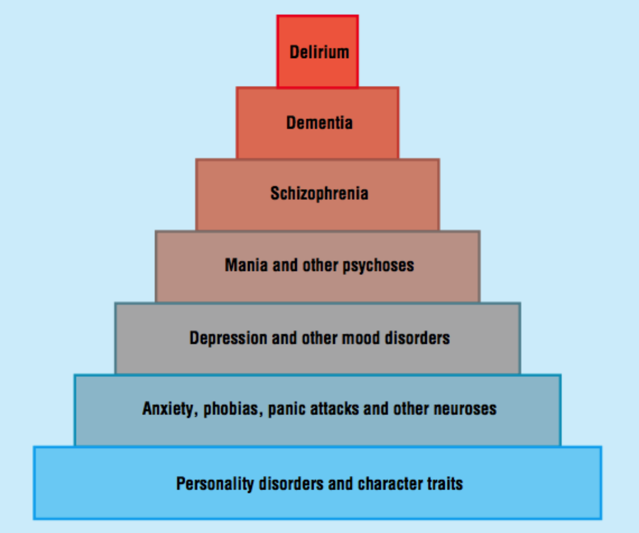

Delirium
Dementia
Schizophrenia
Mania and other psychoses
Depression and other mood disorders
Anxiety, phobias, panic attacks and other neuroses
Personality disorders and character traits

Figure 1.1. Hierarchy of symptomatology of mental disorders. Figure extracted from [49].

complemented with mobile sensing [160]. The *International Classification of Diseases* [163] (ICD-10) states that the term *disorder* is employed to avoid conflicts related to the use of terms such as *disease* or *illness*. Although "disorder" is not a precise term, it is used to denote the presence of a series of symptoms and/or behaviours that are clinically identifiable and usually correlated with distress and disruption in the personal functions of the individual. For this reason, we hereinafter avoid the terms disease or illness to refer to alterations in the psychological state of an individual and use instead the term "disorder".

Mental Health State Assessment (MSA), also known as Mental State Examination, bears direct analogy with the physical examination, allowing practitioners to objectively derive the markers of mental disorders. It refers to a structured procedure to ascertain an individual's behavioural and cognitive functioning at a given point in time. According to Martin [134], MSA encompasses the following descriptions of subject: (i) General behaviour and appearance; (ii) Mood and affect; (iii) Attitude and insight; (iv) Level of consciousness and attentiveness; (v) Thought and perception; (vi) Motor and speech activity; (vii) Reaction stimulated in the examiner; (viii) Higher cognitive abilities.

MSA aims to acquire a comprehensive cross-sectional characterisation of the person's mental condition. In layman's terms, this can be thought as analysing different *snapshots* taken at specific points in time, allowing to compare several different variables at each point. In a subsequent stage, the individual's psychiatric record, including historical and biographical data, is integrated with this characterisation allowing health practitioners to perform an accurate diagnosis, required to define a coherent treatment planning. To achieve a more accurate assessment, a variety of associated symptoms and subject's mental state signals should be taken into account.

The data required to conduct the MSA of an individual can be gathered through a combination of direct and indirect assessment instruments. These include unstructured observation (through extraction of social and biographical information), specific questions regarding the manifesting symptoms, and formalised psychological tests (also known as surveys, questionnaires or inventories). In particular, the latter have been extensively used as a reliable way to collect high-quality data from online sources. As argued by Urbina et al. [231] such instruments are objective and standardised measures of a sample of behaviour, and many of them are amenable to administration over the Internet. Some examples are the CES-D [183] (*Center for Epidemiologic Studies Depression Scale*), PHQ-9 [109] (*Patient Health Questionnaire*) and BDI [12] (*Beck Depression Inventory*) to measure the severity of depression or the TSQ [23] (*Trauma Screening Questionnaire*) used to screen for PTSD, just to mention a few.

1.2.2 Online Social Media Platforms

Although it is hard to provide a precise definition of the term *social media*, in a broad and general sense it comprises a collection of online communication channels fostering content-sharing and interaction between peers while enhancing collaboration and community-based exchange. Wikis, social networking, bookmarking and curation sites, microblogs, and forums constitute the various kinds of social media. People use these platforms to build social relations with other individuals who have common backgrounds, interests or, in many cases, real-life connections. On a daily basis, an uncountable number of users employ these sites to produce an extensive variety of content, such as text messages, photos, videos and links. Among the most prominent examples are Facebook, Reddit, Twitter, Instagram, and Tumblr. Moreover, the continuous improvements in mobile technologies enable users to post updates without any space and time constraints [6]. This creates a unique opportunity for researchers to leverage this plentiful content for a variety of studies [5], especially those concerning human well-being, such as public health analysis [166], disease surveillance [24; 112] and modelling [209].

Ever since the advent of online social media, various types of social platforms have emerged. While many platforms, such as Facebook and Twitter, support various modalities of data (e.g., text, images, and videos) some other platforms focus on a more limited range of content. Moreover, the structure of each platform differs, leading to different user behaviour and use cases. While people follow a topic using related hashtags on Twitter, “subreddits” are the rooms where a topic is discussed on Reddit. More specific topics are in the form of posts in which people can leave comments and continue a discussion. Therefore, users might behave differently on different social media platforms [90; 132; 162]. For this reason, in this thesis we consider two mainstream social media platforms, namely, Reddit[8] and Twitter[9]. Our choice is motivated by the fact that both platforms have very different environments and features, leading to a very different range of behaviours by their users [203; 73].

1.2.3 Mental Health State Assessment on Online Social Media

Bearing in mind the WHO statistics outlined at the beginning of this chapter, the capabilities of public health systems to cope with the plethora of cases that emerge on a daily basis are certainly limited. To date, the judgement and experience of a mental health expert has no technological substitute. However, the constraints dictated in reality make the efficient process of personal diagnosis unfeasible. Population-level analyses via traditional methods, like the Behavioural Risk Factor Surveillance System (BRFSS)[10], are costly, take a significant amount of time, and usually come with a considerable delay. New solutions involving the use of language technologies should be considered at least for a preliminary screening process and to raise awareness. The integration of techniques and approaches from areas such as Information Retrieval (IR), Natural Language Processing (NLP) and Machine Learning (ML), can leverage social media data to assist health professionals in screening and monitoring individuals potentially at risk. This explains the growing interest in exploring the application of computational methods for this problem.

Different modalities, such as images [244; 188; 34], Internet usage data [154; 102], phone usage data [210; 15; 1; 195], or GPS data [81; 80; 29; 236], have been considered to characterise the mental state of individuals through computational methods. In this thesis we focus on the computational analysis of textual records, and therefore the language use, to conduct MSA on a subject. In particular, the emergence of social media platforms has reinforced the interest in this matter and, as a consequence, research on mental state assessment has moved to-

[8] See `https://www.reddit.com`

[9] See `https://twitter.com`

[10] See `https://www.cdc.gov/brfss/index.html`

ward these online streams. Bearing this in mind, we define Online Mental Health State Assessment (OMSA) as the task of analysing the online digital traces left by individuals in order to ascertain their mental state condition at a particular point in time.

1.3 Computational Personality Assessment

1.3.1 Personality Assessment

An individual's personality has a significant impact on the person's live; for example, on job performance [223], interpersonal relations [237], products purchasing [114], and even on health and well-being [84]. According to psychology experts, the term "personality" has no single universally accepted definition, although in a broad and general sense personality comprises a set of consistent patterns of *thought*, *affect* and *behaviour* stemming from individuals' values, attitudes, past memories, social relationships, and habits [149]. These three patterns can be described as follows:

- Thoughts: Concern the cognitive aspect of personality. In essence, it refers to the schemes or patterns through which human beings represent their inner and outer world (beliefs about self and others).

- Affects: Involving three dimensions, valence, arousal and social aspect. In this respect, Plutchik [175] states that persistent situations involving emotions produce persistent traits or personality. For instance, if one is angry most of the time, then anger or a related phenomenon such as aggressiveness become part of the personality. Emotions are considered to be more *transient* phenomena whereas personality is more constant.

- Behaviours: Concern the actions taken by individuals, not only actual actions, but also fantasies and intentions.

The term *consistent* means that such patterns are relatively stable across time and contexts [142; 141].

Traditionally, personality assessment has relied on self-reports (gathered through questionnaires), such as the BFI [97] (Big Five Inventory), the S5 [107] (*Short Five*) or the TIPI [77] (Ten Item Personality Measure), and laboratory studies conducted and analysed by trained psychologists. As the instruments developed via these methods go through many validation processes, such approach for the assessment of personality lies on solid empirical evidences. Nonetheless, self-reports are likely to present several shortcomings [219], such as: (a) They require human assessment as well as a certain level of training of the assessors; (b) They are affected by *response biases*, usually manifested through social desirability bias (completing the questionnaire in a way that makes people look more "favourable" to others [110]); (c) They are not free from the *reference-group effect* [85] (e.g., dispositional introverted individuals might perceive themselves as extroverted if they are surrounded by a peer-group of even more introvert colleagues or friends [245]). To overcome these drawbacks and find implicit measurements rather than explicit self-reports, personality researchers have sought to complement traditional assessment approaches with novel solutions based on the application of computational methods [218]. This rising trend has given birth to the field of *computational personality assessment* [234; 157] (CPA).

1.3.2 The Lexical Hypothesis

The *lexical hypothesis* states that the *individuals' differences that are most prominent and socially relevant in people's lives would eventually become encoded into their language* [96]. Therefore, the stronger such differences are, the more likely it is that they are expressed in word usage. In this way, by sampling language, it would be possible to determine a comprehensive taxonomy of human personality traits.

Individuals exhibit certain patterns when they talk or write about other people. The way those individuals describe others can help to characterise and identify the individuals themselves. For example, a person could be identified as histrionic or friendly. The various adjectives that those individuals use in describing others can be collected and used to ask people how much each adjective associates with them. The result of this assessment will be a set of self-reported measurements, which clustered together using factor analysis reveal five basic "factors" or "dimensions" of personality, which are:

- "Openness to experience" (AGR): unconventional, insightful, imaginative;
- "Conscientiousness" (CON): organised, self-disciplined, ordered;
- "Extroversion" (EXT): cheerful, sociable, assertive;
- "Agreeableness" (AGR): cooperative, friendly, empathetic;
- "Neuroticism" (NEU): anxious, sad, insecure.

This way of studying human personality is known as the Five-Factor Model (FFM) [97; 137], also called the Big Five, and constitutes the most popular methodology used in automatic personality research [157]. The Big Five can be thought as continuous dimensions (i.e., traits) or as discrete categories (i.e., personality types). Furthermore, each of the five factors presents a *positive* and a complementary *negative* dimension. For instance, the complementary aspect to neuroticism is defined as "emotional stability". Each individual can have a combination of these dimensions at a time.

There exist several linguistic and behavioural implications of the Big Five model, in particular the various traits yield distinctive and defining characteristics. For instance, literature has shown that individuals who scored high on introversion and neuroticism preferred written communication channels over face-to-face contexts [86] and that agreeableness is a trait characterised by an affiliative social orientation and tendency to avoid conflict with others [78]. From a psycho-linguistic perspective, it has been shown that each of the five dimensions is characterised by different styles in language usage. For instance, extroverts are found to talk more, louder, and more repetitively, have fewer pauses and hesitations, a lower type/token ratio, use more positive emotion words and less formal language than introverts [130; 171; 68]. Other linguistic cues of conscientiousness, such as the use of filler words and second-person pronouns, have been observed to vary across gender [140], introducing thus additional confounds[11] in automatic personality detection.

An alternative model to measure personality, although less accepted in the scientific community, is the Myers Briggs Type Indicator (MBTI) [155]. MTBI bases its definition on the conceptual theory stating that human personality stems from the different ways in which individuals experience the world around them and take decisions. Such differences are grouped into four opposite pairs resulting in 16 possible personality types: namely, Extraversion/Introversion, Sensing/Intuition, Perception/Judging, and Feeling/Thinking. This model is widely used in the industry.

1.4 Thesis Outline

This thesis is organised into two parts and seven chapters. Prior to the two parts, **Chapter 2** presents the related work on the main themes covered throughout the dissertation. First, we review computational methods for mental health state

[11] In an psychology experiment, a *confound* is an independent variable that is conceptually distinct but empirically inseparable from one or more other independent variables.

assessment. In particular, we focus on the analysis and processing of naturally occurring *digital traces* that individuals leave on their online environments (such as social media platforms). Second, we outline the computational assessment of personality expression from language cues. **Chapter 3** describes the various external datasets used in the thesis. In essence, we summarise how each dataset was collected, we outline their main characteristics through a series of statistical measures and provide examples of the textual records which comprise them.

Part I focuses on OMSA from textual records. Language analysis is performed considering both prediction and insight applications. Part I includes the following two chapters:

Chapter 4 shows how the latent semantic structure of textual posts can be exploited to identify evidence that could suggest the onset of depression. To this aim, we leveraged a methodology built on the concept of semantic proximity previously employed to assess how mental state changes can be detected after drug induction. We evaluate the effectiveness of the features and study the sensitivity of various parameters of the resulting assessment algorithm. Next, we present in this chapter a weak-supervision framework for automatically deriving large samples of data for the study of depression on online social media settings. We empirically validate our methodology and show it can be effectively used for automatically collecting posts samples.

Chapter 5 elaborates on a series of analytical studies which aimed at gaining a better understanding on the language and behaviour that characterise individuals affected by mental disorders on online settings. We investigated the writing style, how people express their emotions, and their online behavioural patterns on social media via visualising certain probabilistic attributes. More precisely, we analysed and visualised the activity, vocabulary, psychometric attributes, and emotional indicators on people's posts on two very different social media platforms. In addition, we present a detailed analysis toward improving the understanding of how depression assessment inventories, in particular the BDI, could be automatically estimated based on the evidence available on social media. To this aim, we investigated the relative incidence of the 21 BDI items on users' feeds, as well as how incidence and other features influence the effectiveness of the automatic tools that extract depression symptoms from social media posts. The ultimate goal of these studies was to provide insights that could be used to *help a predictive system* or a health practitioner in the elaboration of a diagnosis. We describe how we designed each experiment to study social media posts and outlined the main takeaways after conducting each of them.

Part II elaborates on the work we have done on personality assessment from conversations. Again, we focus on both prediction and insight applications. It includes the following chapter:

Chapter 6 introduces a novel approach to personality recognition in conversations based on capsule neural networks. A capsule hosts a small group of neurons whose activities represent the various properties of a specific type of entity. The capsule-model learns a hierarchy of feature detectors through a routing-by-agreement algorithm. In a conversation, these feature detectors can represent sets of words whose occurrence co-vary, thus revealing latent underlying personality traits. As research has shown, the study of linguistic personality-related attributes in natural language provides a window into individuals' mind for better understanding of the rational behind their behaviour. Motivated by this, in this chapter we proposed a novel open-vocabulary approach based on multiword expressions which aims at discovering linguistic patterns associated with personality traits. We describe the procedure defined to discover a set of candidate expressions from the text produced by individuals with a particular trait and propose a discrimination scoring function taking into account their co-occurrence patterns producing a meaningful set.

Finally, **Chapter 7** concludes the dissertation and describes insights into future directions or research that stem from the developed work. Moreover, important ethical concerns and implications of this work are also discussed.

1.5 Main Contributions

The main contributions of this thesis are as follows:

- We present a novel early risk-assessment system that measures the semantic proximity between users' textual posts and a set of words with topical relevance to depression and use such information to identify the onset of this mental disorder. We perform different experiments to assess the effectiveness of the system. In particular, we show the effect of the decision threshold on its effectiveness and the temporal spread of the cues that indicate the onset of depression.

- We present a weak-supervision framework to derive large samples of data for the study of depression on online social media settings. Alongside, we release the dataset created following the proposed methodology and develop a range of models to serve as a benchmark to foster the research on data-driven approaches for automatic identification of depression.

- We present a series of techniques based on statistical and visual analyses which are able to identify significant differences in the language and behaviour of high-risk individuals on social media platforms. For the first time, two social media platforms with diverse characteristics are compared and analysed in this matter.

- We present a novel approach based on capsule neural networks for automatic recognition of personality in conversations. To this end, we tackled this task as a frequency co-variance between different sets of words modelled as capsules. We assess the effectiveness of the model and compared it with state-of-the-art approaches on personality recognition from text. Moreover, we provide a quantitative analysis showing its inherent interpretability potential.

- We present a novel open-vocabulary approach based on multiword expressions to gain insights on personality and its manifestation through language. In particular, we discover a set of candidates and define a discrimination scoring function to select meaningful expressions that emerge from both spoken and written sources and that are intended to capture the most distinctive linguistic patterns of a personality trait. We show the potential of our method through a quantitative and comparative analysis.

1.6 Publications Overview

1.6.1 Publications in Thesis

Most of the material presented in this thesis was published in conferences and journals as listed below:

1. **E. A. Ríssola**, S. A. Bahrainian, F. Crestani. Anticipating Depression Based on Online Social Media Behaviour. *In proceedings of the 13th Flexible Query Answering Systems International Conference (FQAS)*, pages 278-290, 2019.

2. **E. A. Ríssola**, S. A. Bahrainian, F. Crestani. Personality Recognition in Conversations using Capsule Neural Networks. *In Proceedings of the 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019*, pages 180-187, 2019.

3. **E. A. Ríssola**, D. E. Losada, F. Crestani. Discovering Latent Depression Patterns in Online Social Media. *In Proceedings of the 10th Italian Information Retrieval Workshop (IIR)*, pages 13-16, 2019.

4. **E. A. Ríssola**, M. Aliannejadi, F. Crestani. Beyond Modelling: Understanding Mental Disorders in Online Social Media. *In Proceedings of the 42th European Conference on Advances in Information Retrieval*, ECIR '20, pages 296–310, Lisbon, Portugal, April 14–17, 2020.

5. **E. A. Ríssola**, S. A. Bahrainian, F. Crestani. A Dataset for Research on Depression in Social Media. *In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP'20, page 338-342, Genoa, Italy, July 14-17, 2020.

6. **E. A. Ríssola**, D. E. Losada, F. Crestani. A Survey of Computational Methods for Online Mental State Assessment on Social Media. *ACM Transactions on Computing for Healthcare*, 2(2):17:1–17:31, 2021.

7. **E. A. Ríssola**, J. Parapar, D. E. Losada, and F. Crestani. A Survey of the Last 5 Years of eRisk: Findings and Conclusions. In F. Crestani, D. E. Losada, and J. Parapar, editors, *Early Risk Detection of Psychological Conditions: the eRisk Experience*, Studies in Computational Intelligence, chapter 2. Springer, 2021 (In Press).

1.6.2 Additional Publications

These additional papers were published in conferences, workshops, and evaluation forums during this thesis, but were not included in it to maintain the coherency of the thesis.

1. **E. A. Ríssola**, A. Giachanou, F. Crestani. USI-IR at IEST 2018: Sequence Modeling and Pseudo-Relevance Feedback for Implicit Emotion Detection. *In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2018*, pages 231-234, 2018.

2. **E. A. Ríssola**, D. Ramírez-Cifuentes, A. Freire, F. Crestani. Suicide Risk Assessment on Social Media: USI-UPF at the CLPsych 2019 Shared Task. *In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 167-171, 2019.

3. **E. A. Ríssola**, M. Chakraborty, F. Crestani, and M. Aliannejadi. Predicting Relevant Conversation Turns for Improved Retrieval in Multi-turn Conversational Search. *In Proceedings of the Twenty-Eighth Text Retrieval Conference*, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019.

4. A. Giachanou, **E. A. Ríssola**, B. Ghanem, F. Crestani, and P. Rosso. The Role of Personality and Linguistic Patterns in Discriminating between Fake News Spreaders and Fact Checkers. *In Proceedings of the 25th International Conference on Applications of Natural Language to Information Systems*, NLDB 2020, pages 181–192, Saarbrücken, Germany, June 24-26, 2020.

5. M. Aliannejadi, M. Chakraborty, **E. A. Ríssola**, F. Crestani. Harnessing Evolution of Multi-Turn Conversations for Effective Answer Retrieval. *In Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, CHIIR' 20*, pages 33–42, Vancouver, Canada, March 14–18, 2020.

Chapter 2

Related Work

In this chapter we review the most relevant related work with a focus on two main themes. These themes are, first *online mental health state assessment* a notion closely connected with most of the research in this thesis and the focus of the first part the thesis. Second, and equally important, *computational personality assessment* focus of the second part of the thesis.

2.1 Online Mental Health State Assessment

The majority of the works in the area have been mostly focused on the automatic identification of mental disorders on social media. Here, we first outline those where some effort have been devoted to better understanding how the language of individuals suffering from mental disorders and their online behaviour differ from that of healthy individuals (Section 2.1.1). The second series of works here described comprises studies whose focus is on the development of assessment technologies devoted to automatically identifying the onset of a mental disorder. With this in mind, we organise the presentation of these studies according to the feature extraction process conducted and the decision-making technology employed (Sections 2.1.2 and 2.1.3).

2.1.1 Insights on Language and Behaviour

As defined in the Introduction (Section 1.2), OMSA deals with the unobtrusive identification of mental state alterations through the analysis of online digital traces, in particular examining the patterns of language use from textual records. Gaining insights on the language and behaviour of individuals affected by mental disorders could lead to identify new predictive markers up to now not considered in the medical literature, motivating new inquiries into behavioural traits of mental health disorders as observed on social media. For instance, Reece et al. [189] discovered that increases in average document length were positively associated with depression and PTSD. However, studies linking verbosity to psychological conditions are scarce in traditional mental health literature. Motivated by this, in this section we outline a series of analytical studies which aimed at gaining a better understanding on the language and online behaviour that characterise individuals affected by mental disorders.

Ramirez-Esparza et al. [187] has been some of the first to examine online forums posts identifying noticeable linguistic differences between depressed and non-depressed people. In accordance with the medical literature, they found that depressed users exhibited a significantly higher use of first personal singular pronouns as well as negative emotion words. Furthermore, they also studied how the expression of depression varied across cultures. They analysed forums written by English and Spanish speakers and discovered that while the first were more concerned with medicinal questions, the latter tended to share and disclose information about social matters.

A significant body of work has been done considering Twitter as the online source. Starting with De Choudhury et al. [38] who presented early work on automatic depression detection by using crowd-sourcing to collect assessments from several Twitter users who reported being diagnosed with depression. They built a depression lexicon containing words that are associated with depression and its symptoms. Their data analysis revealed that depression sufferers exhibited a noticeable decrease in social activity (lower posting volume), greater negative emotion and reduced arousal, in addition to higher usage of first-person pronouns when compared with non-depressed users.

Coppersmith et al. [43] studied four different mental disorders (depression, bipolar disorder, PTSD, and seasonal affective disorder) using Twitter data. To derive insights into quantifiable and relevant mental health signals, they searched for deviations in language usage and pattern-of-life factors (social engagement, insomnia indicators, exercise-related terms and sentiment words) from a control group of healthy users. Their analysis yielded differences that reach statistical significance from the control group. In the case of depression, they observed a significantly higher use of first-person pronouns (*I* and *we*), as well as swear, anger, and negative emotions words. Conversely, for the PTSD group both the usage of second- and third-person pronouns (*he/she* and *they*) was higher. A significantly higher mention of anxiety words, such as anguish, fear, and overwhelm, was observed for the four mental disorders under study. Finally, the analyses also indicated that depressed users' patterns-of-life factors, such as social engagement and exercise mentions, were more similar to control users than with the rest of the disorders.

Park et al. [165] provided a preliminary study towards verifying whether online social media data were truly reflective of users' clinical depressive symptoms. To this end, they analysed the expression of depression among the general Twitter population. Over a period of two months, they collected tweets that contained the word "depression". A subsequent analysis showed that depression was most frequently mentioned to describe one's depressed status and, to a lesser extent, to share general information about depression. Also, they studied whether self-report inventories could be reliably explained from language usage on social media. The conducted analyses revealed that there was a significant difference between the depressed and non-depressed groups regarding the usage of words expressing anger and negative emotion. In addition, depressed users posted more monologue-like tweets.

Hwang et al. [92] studied a set of common terms on social media that were the focus of various anti-stigma campaigns and which could be employed in a derogatory or pejorative sense, thus creating a stigmatising language. In particular, they assessed whether awareness of mental disorders promoted a more restricted use of these terms, either avoiding mentioning the terms or diminishing their use. Considering a wide range of mental disorders, the analysis showed that a difference in the frequency of stigmatising senses as well as a change in the target of pejorative senses existed between the two groups of mental disorder *aware* and *unaware* users. The authors concluded that individuals' awareness about mental disorders influenced the way in which they expressed themselves on social media suggesting the existence of a degree of sensitivity towards stigmatisation of those affected by a disorder.

Reece et al. [189] conducted a state-space temporal analysis whose goal was to track the evolution of depression and PTSD. Utilising self-assessment questionnaires, they collected several Twitter users who reported being diagnosed with either of these disorders. Based only on the tweets' textual content, they derived time series models that were able to reconstruct from unlabelled data the division between the affected individuals and the control group and describe a timeline of the development (onset and recovery) of the disorder.

To a lesser extent other online sources, such as Reddit or Facebook, have been also investigated. Based on several items used to measure the neuroticism personality trait, whose scales overlap with self-reported items that screen for depression, Schwartz et al. [212] studied the association between the degree of depression (*severity*) and the language used by a set of Facebook users. For each user, the degree of depression was obtained based on the average response to seven depression-related items. Their language analysis revealed that there are various sets of words highly correlated with depression severity and, as previous findings confirm, the level of depression often increase from summer to winter.

Moreno et al. [153] sought for traces of depression from publicly available Facebook profiles. They recruited manual coders to review the history of profile updates according to established clinical criteria. They analysed the relationship between depression on profile (prevalence of displayed depression symptoms on status updates) and a series of variables, such as age, graduation year, and gender. Overall, the most common type of depression symptom reference corresponded to depressed mood, followed by feelings of guilt or worthlessness, indecisiveness and loss of energy. References to sleep difficulties were also common among participants, specially between those who exhibited depression symptoms. Interestingly, individuals who updated more recently their profiles were more inclined to exhibit a reference to depression.

De Choudhury et al. [52] studied several mental health and suicide support communities on Reddit (better known as *subreddits*) whose members mostly participate looking for help and support. In particular, they developed a methodology, based on causal inference, to characterise and infer which users could experience shifts from mental health disclosure to an expression of suicidal ideation (i.e., likely to contemplate committing suicide in the future). The authors identified that such transitions were associated with various markers, such as heightened self-attentional focus, greater detachment from the social dimension, poor linguistic coherence, reduced social engagement, increased self-disclosure, and expression of negative affects such as anxiety, loneliness and hopelessness.

Gkotsis et al. [70] analysed various mental health communities on Reddit to discover discriminating language features between the users in the different communities. They found that, overall, the subreddits that were topically unrelated had condition-specific vocabularies as well as discriminating lexical and syntactic characteristics. Similarly, Gaur et al. [66] presented an unsupervised approach to map the content of various mental health-related subreddits to the best matching DSM-5[1] categories. By leveraging the DSM-5 manual and other curated medical knowledge bases, they developed a domain-specific lexicon containing n-grams associated with each mental health disorder in the DSM-5 categories as well as an enriched drug abuse ontology with mental health-related terminology and slang terms from Reddit. Subsequently, they utilised these lexicons to quantify the relationship between subreddits' content and DSM-5 categories automatically assigning the corresponding labels.

[1] DSM-5 stands for *Diagnostic and Statistical Manual of Mental Disorders - 5th Edition.*

Wolf et al. [238] worked with pro-eating disorder blogs, recovery blogs and control blogs. The aim of pro-eating disorders blogs is to promote an anorexic lifestyle, often encouraging unhealthy habits and in particular, they can promote eating disorders. Recovery blogs are online communities dedicated to the prevention of eating disorders. Conversely, control blogs do not contain any content related to eating disorders or any other clinically relevant contents. The authors studied which language patterns reflect the psychological conditions of the blog authors and provided valuable insights into the various stages of coping. In particular, they analysed cognitive, affective, social and disorder-related behavioral dimensions. Compared with recovery blogs, pro-eating disorder ones exhibited a lower usage of cognitive mechanisms, insight and abstraction but higher closed-mindedness as well as fewer negative emotion words and fewer social and communication words. Compared to control blogs, individuals participating in pro-eating communities, used more exclamation marks and first-person singular pronouns, but less social words. Also, they used more words related to food and eating as well as body and symptoms.

2.1.2 Feature Extraction

In the following sections we focus on summarising predictive screening technology used in OMSA and that is are relevant to the thesis. We start by reviewing several methods used to derive various attributes from users' textual records or *documents* which are later used to develop assessment technologies. Recall that research on language and psychology has shown that several latent mental health related variables can be discovered by examining the patterns of language use. There are several ways to encode the information contained in users' documents. For instance, one could take into account quantitative elements of the text, such as the frequency of specific words, or consider other aspects related to the level of engagement that users exhibit in the social media platform, such as the average post length. These different attributes, know as *features* and their combination play a very important role for conducting mental health state assessment. In most of the cases, the feature extraction stage involves strategies that were previously shown to be effective in other domains or text classification tasks, such as sentiment analysis [164] or authorship attribution [211]. Despite being related, such domains are not focused on assessing the mental health state condition of the users but still provide useful methods to derive text-based features. Subsequently, the generated features are used to train standard machine learning algorithms, given hand-labelled data, or integrated with less sophisticated ruled-based approaches to conduct the assessment. Feature extraction is not a trivial task, and thus developers of predictive technology need to find the most effective and efficient way to represent and encode the information present in users' documents, taking into account *what* users write about, *how* users write it and, for certain specific applications, keep track of the changes that these variables exhibit over time to spot even subtle and/or sudden variations.

Open-Vocabulary: A simple yet effective way to encode the textual content of a document as a vector is known as *Bag-of-Words* (BoW). The basic idea is to focus only on the occurrence and frequency of words in the document, without considering other syntactic attributes, such as their part-of-speech. Thus, any given document can be encoded as a fixed-length vector whose length is equal to the number of terms in the vocabulary. The vocabulary contains the unique set of words found in a collection of documents or as it happens in this case, social media posts. Each position in that vector provides information about the occurrence of a given term in the documents. Such value could simply denote the presence of the term (binary), the number of times that the term is repeated (frequency) or it can represent a more sophisticated term-weighting scheme, such as TF-IDF [47]. Several studies in OMSA have applied BoW to transform documents published by users into vectorial representations [45; 161; 192; 228; 28]. The majority of them extracted single terms (*unigrams*) and sequences of contiguous and overlapping sequences of lexical or sub-lexical units, such as words or parts of them, known as *n*-grams to encode textual contents. Usually, the size of the vocabulary can easily reach the order of thousands of terms [47]. For this reason, feature selection techniques are first applied to discard potentially irrelevant and noisy terms, and therefore reduce the feature space favouring the learning process as only the top-*k* most discriminating terms are kept.

An alternative way to generate a fixed-length feature representation from users posts is to consider the semantic relationships between the words in them. Meaning can be understood as emerging from mutual dependencies of words within the language. Semantically related words co-occur in texts with coherent topics at a higher frequency than unrelated words. Using this property, the relation between two words can be quantitatively measured by the frequency of the co-occurrence patterns each of them present. Latent Semantic Indexing [53] (LSA) , word2vec [145], GloVe [173], fastText [18; 99] and Random Indexing [101] are associative models that capture the meaning of words by means of linear representations in a high-dimensional semantic space, known as *word embeddings* (WE). The semantic content of a word is encoded as a vector and this vectorial representation can be used to estimate how *semantically close* other words are. Such natural language representation are known to be *context-free* since they generate a single WE representation for each word in the vocabulary. Because words mean different things in different contexts, such methods require that word representations capture *all* of the possibilities (i.e., all the possibilities in the semantic universe of a language). It should be noted that, although convenient, this makes some assumptions about the language that often do not fit with reality. Various studies explored the used of word embeddings in the context of OMSA [228; 28; 184; 168].

Lexicons: A common method for linking language with psychological variables involves counting words belonging to manually-created categories of language [38; 45; 39]. This method is also known as "closed-vocabulary" analysis [213; 57]. In this case, a document is represented by the frequency, possibly normalised, of the words that comprise each particular lexicon. Such dictionaries are usually manually constructed by psychologists covering various psychologically meaningful categories which are useful to analyse the linguistic style patterns of an individual's way of writing. Probably one of the most extensively used lexica is the *Linguistic Inquiry and Word Count* [222; 170] (LIWC)[2], which provides mental health practitioners with a tool for gathering quantitative data regarding the mental state of patients from their writing style. In essence, LIWC is equipped with a set of dictionaries manually constructed by psychologists which cover various psychologically meaningful categories which are useful to analyse the linguistic style patterns of an individual's way of writing. Examples of categories are *affect* (hopeless, cheerful, trust), *perceptual processes* (beauty, bright, spicy), *cognitive processes* (idea, obedience, normally), *biological processes* (hormone, thirst, hungry), *drive* (children, hug, emptiness), *social* (divorce, bachelor, paternity), among others.

In the context of OMSA, several dictionaries have been used to generate lexicon-based features, LIWC [190; 228; 185], ANEW[3] [70; 189], AFINN [4] [95], LabMT [70; 189], the Unified Medical Language System (UMLS) Metathesaurus[5] [208; 168], Non-Suicidal Self-Injury (NSSI) [3], Valence Aware Dictionary and sEntiment Reasoner [91] (VADER) [2] and NRC Word-Emotion Association Lexicon [151] (EmoLex) [230]. Moreover, these lexica are usually extended with other ad-hoc lexica, such as [38; 159; 129; 4], built for specific tasks, like depression detection or suicide risk assessment. In essence, they consist of words closely associated with texts written by individuals sharing their experience on

[2] See http://liwc.wpengine.com/

[3] See https://csea.phhp.ufl.edu/media/anewmessage.html

[4] See https://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

[5] See https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

mental disorders or its symptoms in online settings and also include words related to names of medications.

Topic Models: Topic models are hierarchical Bayesian models of discrete data where each topic is a set of words drawn from a fixed vocabulary representing a high-level concept. These models have been also exploited for OMSA. To overcome potential issues related to the short length of the text, all the documents authored by an individual are usually concatenated into a single piece of text. Topic model posterior distributions can be used as features since they encode latent relationships between the words in a document. Resnik et al. [192] used a supervised variation of Latent Dirichlet Allocation [17] (LDA) to discover groups of words with high discriminative power between positive instances of mental disorders and control individuals. Based on previous work [193], they argued that topic models provide a way of gaining both clinical insight and predictive ability. Preoţiuc et al. [179] proposed the application of different word clustering methods to automatically derive groups of related terms (i.e., topics), thus, obtaining various textual features. The authors argued that the use of unrelated corpora confers certain level of generality on the methods as the clusters were computed as task agnostic as possible, aiming not to use medical condition specific keywords or data. Therefore, it is expected similar methods to perform well in identifying different types of disorders.

Social Engagement and Posting Behaviour: These set of features intend to profile certain attributes of users' documents, such as the length or the time they were created, as well as their level of participation on the social media platform. As stated by Coppersmith et al. [46] such attributes provide a proxy, although imperfect, for significant findings in the mental health literature which may manifest and be measured on social media. This category of features includes the computation of the following parameters (but not limited to):

- Textual spreading [131; 228; 28; 3]: amount of textual information provided by the users in their posts, such as the length.
- Time gap [28; 2]: time between two consecutive posts
- Time span [59; 28]: Recording of the time and date when users published their posts.
- Posting Frequency [7].
- Average/median number of words per post [185; 184; 131].

- Follower/friend ratio, number of followers, favourite counts, replies and mentions [179; 95].

Writing Style: Many words such as pronouns, articles and prepositions somehow reveal people's emotional state, personality, thinking style and connection with others individuals [39]. Much of these attributes are used to quantify the importance of personal phrases (i.e., sentences with a first person pronoun) to model the profiles of users suffering from a mental disorder. More specifically, they assumed that the words used in self-reference statements might disclose the individual's social and psychological states and, thus, they might be useful proxies to identify alterations in behaviours and mental states. In particular, these attributes attempt to capture the writing style of the users and are used to quantify specific stylistic patterns that could differentiate positive instances of a mental disorder from control individuals. This set includes part-of-speech (PoS), verbs tense analysis, function words[6], use of negation words, punctuation and measures of text readability and complexity [70; 7; 3; 230].

2.1.3 Assessment Technologies

The successive stage after deriving different features involves their integration with some kind of predictive technology to develop the final assessment system. Recall that the ultimate goal of risk-assessment and decision-making applications is to aid and facilitate health practitioners daily labour. Therefore, such a system should be able to effectively determine at early stages whether an individual is likely to develop (or has already developed) a mental disorder. The various types of features previously outlined provide a way to encode the latent information associated with the mental state of the individuals such that a decision-making algorithm could *learn* or *leverage* the patterns present in the data to conduct the assessment. In this case, the outcome of such assessment is a binary variable stating, usually with a certain level confidence, whether the individual is suffering from a specific mental disorder or not. Most of the assessment technologies are based on standard supervised learning algorithms without any modification and, generally, with their parameters optimised for the task. On the other hand, a minority of studies proposed less sophisticated yet effective ruled-based approaches to conduct the assessment [169; 45; 65].

The majority of the approaches proposed in the literature that deal with OMSA use a standard prediction algorithm, from the field of machine learn-

[6] A *function word* is a word whose purpose is to contribute to the syntax rather than to the meaning of the sentence.

ing, which is usually trained on various combinations of features by using hand-labelled examples to learn meaningful patterns and optimise their parameters. The great success of such algorithms lies in their ability to learn latent nuances in the input data. Subsequently, the outcome and generalisation ability of the prediction algorithm is evaluated using a set of unseen instances. Although there is not a general consensus on the best prediction algorithm (or combination of them) for conducting OMSA, a variety of statistical learning algorithms have been considered. The most applied ones are Naïve Bayes (NB) [208; 59], Support Vector Machines (SVM) [192; 208; 168], Logistic/Linear Regression (LR) [179; 228; 184; 168], Random Forest (RF) [131; 7; 28; 185], Ada Boost (AB) [168] and *K*-nearest Neighbours (KNN) [233]. Furthermore, individual prediction algorithms can be combined into *ensembles* to produce gains in performance. In such cases, the final output of the ensemble can be, for instance, calculated as an unweighted mean of each individual member.

More recently, the emergence of *deep learning*, one of the fastest-growing fields of machine learning, has drawn the attention of researchers working on OMSA. The main difference with the previously outlined statistical learning methods is that deep learning approaches use neural networks, usually with a *deep* or large number of hidden layers, to learn several levels of abstraction. While some studies apply simple network architectures such as Multilayer Perceptron (MLP) [136] others adopt more complex architectures tailored for sequential data [89; 36], such as Long Short-Term Memory (LSTM) [168], Gated Recurrent Unit (GRU) [208], or other successful methods in the field image recognition [76], such as Convolutional Neural Network (CNN) [228]. It should be noted that the learning algorithms of these deep learning architectures usually rely on large sets of hand-labelled training data, which in some occasions might prevent their application as not always it is possible to count on datasets of such magnitude. However, there has been recent advances on the development of approaches that work very well with very small datasets, for instance one-shot or zero-shot learning architectures. Moreover, it is worth mentioning that in some occasions labels do not have to be manually decided. They can be approximated or weakly learned (e.g. the output of unsupervised clustering). Finally, the complex design and functioning of deep learning architectures make usually hard to understand and interpret the rational behind their outcomes. In spite of their competitive performance, for the clinical practice this is a very sensitive issue that might hinder their application in real-life settings. Nonetheless, the field of interpretability and understandability of deep learning architectures has today exploded with very relevant advances on the matter [9].

2.2 Computational Personality Assessment

The computational analysis and assessment of personality comprises a broad spectrum of cues, including written texts, verbal interactions, nonverbal behaviour, data collected via mobile or wearable devices and online games [234]. Here, we only review the literature on the assessment of personality from textual records (in some cases transcribed conversations) which is relevant to the thesis.

2.2.1 Linguistic Factors and Personality Traits

Pennebaker et al. [171] conducted a study on a set of stream-of-consciousness essays to gain insights on how various psychometric language variables, such as those included in hand-tuned dictionary-based tools as LIWC, correlate with the five-factor measures of personality, the various demographics measures and several health-related behaviour markers. The aim of their study was twofold. On the hand, to examine how language use can reflect individuals differences, in particular personality styles. On the other hand, to determine whether such differences in language are stable across writing contexts. Their findings revealed that agreeableness is characterised by more positive emotion words and fewer articles; neurotism is featured by more negative emotion words and more first-person pronouns; openness-to-experience is correlated with longer words and avoidance of first-person pronouns; and conscientiousness with fewer negations and negative words.

Building on previous psychological findings about correlations between measurable linguistic factors and personality traits Mairesse et al. [130] presented an early work on automatic personality assessment on conversations of all Big Five personality traits using various closed-vocabulary psychometric variables (such anger words, social processes and past tense verbs) and prosodic features. To this end, they used a collection of daily-life conversation extracts using self and observer ratings of personality [140]. In this way, they studied both how personality was perceived on the basis of language and how personality is expressed in language. Additionally, they also analysed the corpus of stream-of-consciousness essays along with self-report ratings. Their analysis and results showed that extraversion, emotional stability and conscientiousness traits were the most evident in spoken language. While, regarding written language, openness to experience was the most evident trait.

Neuman et al. [158] proposed the construction of a set of vectors using a small group of adjectives, which according to theoretical and empirical knowledge, encode the essence of personality traits and personality disorders. Using a context-free word embeddings they measured the semantic similarity between these vectors and the text written by different individuals. The similarity scores allowed to quantify the degree in which a particular personality trait or disorder was evident in the text. The scores obtained were used as features to train a classifier and automatically recognise the personality from a set of stream-of-consciousness essays [171]. In a following work, Neuman et al. [157] proposed a method for measuring psychological dimensions of the five factors of personality. To this end, they manually derived a set of propositions and lexico-syntactic patterns from validated psychological questionnaires and measured their semantic similarity with essays written by several individuals. In this way, they estimated the degree to which a target trait was evident in the text, given the pre-defined patterns obtained from the questionnaire.

2.2.2 Cues from Online Records

The study of personality quickly permeated into online sources. Yarkoni [242] conducted an analysis of personality and word use using a corpus of 694 blog entries spanning a mean period of 23.9 months. In this way, associations between personality and language could be studied over a longer period of time than previous attempts in the literature (which usually spanned only several hours or days). The author conducted a category-based analysis (close-vocabulary) using the various categories provided in LIWC. Categories that were non-semantic (e.g., proportion of long words) or relevant primarily to speech (e.g., non-fluencies and fillers) were not considered. His results pinpointed robust correlations between the Big Five traits and the frequency with which bloggers used different word categories, many of which strongly agreed with prior findings. For instance, neuroticism correlated positively with usage of several different negative emotion word categories (such as anxiety/fear, sadness and anger) and extraversion was associated with increased use of categories related to positive emotions and interpersonal interaction (such friends, sexuality and second person references). This suggests that personality presents similar influences on offline and online forms of self-expression. Moreover, new findings were revealed, such as a significant positive correlation between agreeableness and the use of sexual words.

In a similar vein, Iacobelli et al. [93] collected a corpus of personal blogs to investigate the relationship between language and personality in this domain. To this end, they explored the used of n-grams in comparison with psycholinguistic word categories to conduct personality assessment over the dimensions defined by the Big Five. They found that open-vocabulary features, such as bigrams, provided superior performance than lexicon-based features, showing that language structure is important when assessing personality types. Similarly, functional stopwords are also important in this context.

More recently social media platforms, such as Facebook and Twitter, have also been extensively studied for the assessment of personality. Schwartz et al. [213] introduced the concept of "open-vocabulary" analysis to gain insights on how personality is manifested in a collection of Facebook status updates as well as its relationship with age and gender. Hence, instead of studying predefined and manually-created word lists, they examined two linguistic features automatically derived from the text: words frequency and sets of semantically related words, better known as topics. While many of their findings confirmed previous research, they also discovered new markers of language associated with different traits, such as the mentions of social sports and life activities for emotional stability (opposite dimension of neurotiscm). A considerable amount work has been devoted to analysing and modelling personality expression on social media [74; 181; 108; 178; 69]. These studies have focused on the use of language, user's social engagement attributes (e.g., follower/friend ratio, total number of messages posted, number of likes/favourites), network structure/characteristics and demographics to gain a better a understanding of which are the elements that strongly correlates with personality traits in the context of social platforms.

Mohammad et al. [152] studied the association between emotions and personality on social media. Their experiments showed that fine emotion categories (using emotion-word hashtags association lexicons) such as that of excitement, guilt, yearning, and admiration are useful in automatically detecting personality from text. Celli et al. [32] conducted a comparison between Big Five and MBTI gold-standard personality types for automatic personality recognition on social media. They utilised two multilingual datasets gathered from Twitter, one annotated with Big Five assessments and one with MBTI assessments. Despite the various limitations of the study, such as the fact that users in each dataset were not the same, they concluded that models trained on MBTI could potentially yield better performance when compared to those trained with Big Five assessments. However, using Big Five assessments allows to gain more informative insights regarding the features used to fit the different models.

2.3 Summary

Throughout this section we have shown that the use of computational methods for OMSA is an emerging area drawing the attention of many researchers. The interaction between mental disorders and language on online settings is a challenging task and requires further research and solutions. We identify much room for improvement both on prediction and, specially, on insights applications. Language technologies have a the great potential to identify and exploit latent linguistic nuances encoded in users' textual records that correlate with mental disorders; thus, they can provide practitioners with a holistic and comprehensive way to assess the individual.

Regarding personality, much work has been devoted to studying and predicting it from online sources. Conversely, research on the analysis of personality's expression on conversations has been noticeably scarce. Given the recent advances on conversation understanding system, a desirable feature for such systems would to be able to adapt their language and, mostly, their expression according to the personality of the human partner. To this end, they first need to determine in the most unobtrusive, natural and effective way, avoiding naive questions that simply prolong conversations, the personality of the individual that is chatting with them.

Chapter 3

Data Resources

3.1 Introduction

An important limitation of collecting data for the computational analysis and assessment of individuals' mental-state and personality is related to the way in which *positive instances* are extracted. Given the characteristics and particularities of the domain under study, there must exist some supporting *evidence* that ensures, at least to some extent, that a given person is really suffering from a specific mental disorder or scores positively on a certain personality trait. Irrespective of the social context from which data is retrieved, researchers have followed a variety of methods to collected data based on different probative elements. Most of these methods are based on (a) self-statements [43; 44; 41; 69]; (b) self-assessment inventories [38; 189; 108; 213].

Each technique has its advantages and caveats. Self-assessment questionnaires, such as the BDI or the IPIP, produce high-quality data but are limited in scope (e.g., by mental disorders with a psychological screening test that can be easily distributed over the Internet) and size (by questionnaire respondents). On the other hand, retrieving cases based on self-statements makes it possible to automatically (or semi-automatically) derive labelled samples from large amounts of data. However, this method might only capture a sub-population (i.e., those who publicly talk about such a private matter), and hence might not really represent all the attributes of the population as a whole [43]. Furthermore, the *ground truth* obtained in this way is prone to error because an accurate verification is not feasible since people are not always truthful in self-statements. Nonetheless, given the social stigma normally related with mental disorders, the probability that individuals would publish in their social feed that they are diagnosed with a condition they do not have is rather low [43]. Similarly, they are not strong motivations why individuals would lie about their personality traits as generally they choose to include this information as a way of introducing and describing themselves to the community. Finally, when self-statements are used to gather data, users *passively* participate in the collection process, as their publicly available data is retrieved without their direct involvement in the process[1]. Conversely, individuals *actively* participate when ground truth data is gathered using self-assessment inventories, as they are contacted and requested to fill up one or more questionnaires (additionally, they might be asked to provide consent for a one-time crawl of their social media feed or to be recorded when performing some task).

In this chapter, we describe the various external datasets used to conduct OMSA and CPA. In particular, we summarise how each dataset was collected, we outline their main characteristics through a series of statistical parameters and provide examples of the textual records which comprise them. The remainder of this chapter is organised as follows. Section 3.2 briefly discusses the motivation and challenges of using social media sources to conduct computational studies which involve the analysis of individuals' psychological processes. With this in mind, we describe and characterise the test collections created at two evaluation campaigns that have fostered the research on language technologies and mental health. Similarly, Section 3.3 introduces the notion of self-assessment inventories and describe two datasets collected for the assessment of personality in daily life.

3.2 Self-stated Diagnosis on Social Media

By nature, social media is *social*. This means that social interaction patterns, a crucial element in OMSA, may be readily observable and quantifiable from raw data. The (semi-)anonymous and open nature of social media encourages people to socialise and self-disclose. In fact, many people share content regarding their daily life, and frequently announce major life milestones [51]. As stated in Chapter 1, a considerable number of signs related to individuals' mental health state and emotional conditions can be captured by analysing the way in which people communicate. Therefore, the language used and the emotions conveyed in the content that users share in their social feed may highlight feelings like for example worthlessness, guilt, or helplessness. This provides a profitable mean for identifying and characterising different mental disorders. Social media data, which naturally occurs in a non-reactive way, becomes therefore a valuable com-

[1] Ethical concerns in this respect are discussed in Chapter 7.2

plement to more conventional assessment instruments (Section 1.2) used to determine the potential presence of mental health concerns [57].

Gathering the data to conduct a computational study using social media sources is often a challenging endeavour in itself. Platforms such as Twitter or Reddit provide APIs (Application Programming Interfaces) which facilitate the process of collecting data, by enabling operations such as keyword search or random sampling of their data streams. Nonetheless, not all social media platforms provide an easy way to access their data and, if they do, it is mostly at limited volume and rate. This limitation might hinder the possibility of running any computational study on mental health since identifying measurable signals of mental state alterations through statistical or machine learning methods involves large quantities of data that associate individuals mental health state to their social media feed.

An *Experimental Framework* (EF) is a benchmark exercise that strives to bring together many researchers to tackle the same problem. This type of initiatives often contribute to establishing solid foundations, common standards, and a thorough understanding of the problem and the data required to conduct different tasks. Typically, issues related to evaluation methodologies, performance metrics and creation of test collections are actively discussed. These evaluation exercises commonly run as separated workshops or labs in larger conferences. In this respect, the Early Risk Prediction on the Internet (eRisk) [119; 120; 121; 122], as well as the Computational Linguistics and Clinical Psychology (CLPsych) [45] initiatives have fostered over the last few the years collective efforts worldwide to leverage social media data to develop models for estimating the occurrence of signs of mental disorders.

The collections created by the organisers of these EFs are comprised of sets of documents posted by users of a social media platform, such as Reddit or Twitter. Users are split into two groups: (a) cases of individuals potentially suffering from mental health concerns, such as anorexia or depression (positive group); and (b) control individuals. On both EFs, the organisers followed the extraction methodology proposed by Coppersmith et al. [43], wherein users of the positive group were gathered by retrieving self-statements of diagnosis (e.g., the sentence "I was diagnosed with depression") and manually verifying if they contained a genuine statement of diagnosis. Control users were collected by randomly sampling from a large set of users available on the specific social media platform. It cannot be excluded that the control group might contain some positive case. Similarly, it cannot be ruled out that the positive group might contain some negative cases (because a user's claim about being diagnosed might be deceitful). However, the impact of such cases is expected to be negligible and, in any case, other extraction strategies (e.g., depression screening based on self-assessment inventories) can also incorporate such type of noise in the collections.

Throughout this thesis, we study and analyse various collections released at different editions of eRisk and CLPsych to conduct OMSA of individuals. We decide to conduct our various studies using these collections since they have been carefully curated, validated and used through the various workshops' editions. The fact that they have been extensively tested by several researchers allow us to build our work based on their experience with the collections. Furthermore, they are publicly available for research, which is not usually the case in the domain under study. The collections described below are used in Chapters 4 and 5.

3.2.1 CLPsych

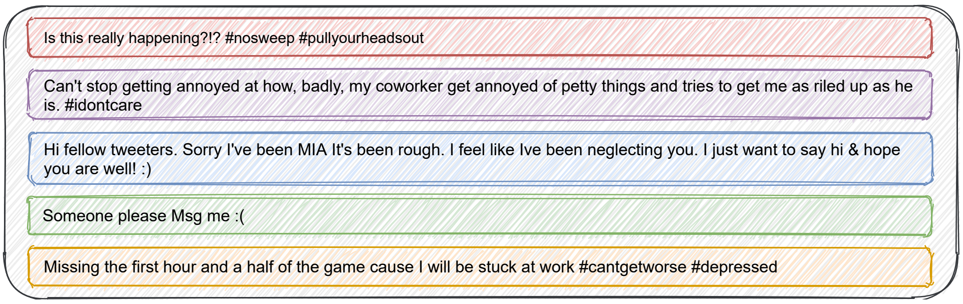
The *Computational Linguistics and Clinical Psychology (CLPsych*[2]*)* workshop series began in 2014 and since then it has sought to bring together language technologies and clinical psychology, fostering the discussion on how such technologies can improve mental health. In particular, it aims at communicating relevant computational methods and results to an interdisciplinary audience, and continuously linking the work back to its clinical relevance. The 2015 edition [45] focused on three user-level binary classification using data retrieved from Twitter. Workshop participants were asked to distinguish users between: (i) Depression and a control group (DvC); (ii) PTSD and a control group (PvC); (iii) Depression and PTSD (DvP).

Twitter is a micro-blogging social media platform in which users post short messages (limited to 280 characters) known as *tweets*. Even though the platform features topic-specific lists where people can follow messages related to certain topics, most users follow topics by the hashtags (i.e., "#") that are used in each tweet. The structure and features of Twitter are very different from typical blog websites. For instance, sharing a post (i.e., *retweet*) is a common means of interaction with a post.

An example of the posts included in the collection created by the workshop organisers is shown in Figure 3.1. It is worth mentioning that this dataset exhibits a balanced distribution between the positive and control classes. The organisers opted for building a dataset maximally relevant to the tasks at the cost of releasing a less *realistic* collection (i.e., not accurately reflecting the user population). In addition, age- and gender-matched community controls were also retrieved. To this end, they approximated the age and gender of each depressed and PTSD user through the analysis of their language and collected a paired control group

[2] See `https://clpsych.org/`

with the same attributes. For each user, a maximum of 3,000 tweets were retrieved and included in the dataset. Those users with less than 25 tweets and/or whose posts were not at least 75% in English were discarded. A summary of the CLPsych collection is shown in Table 3.1.



Is this really happening?!? #nosweep #pullyourheadsout

Can't stop getting annoyed at how, badly, my coworker get annoyed of petty things and tries to get me as riled up as he is. #idontcare

Hi fellow tweeters. Sorry I've been MIA It's been rough. I feel like Ive been neglecting you. I just want to say hi & hope you are well! :)

Someone please Msg me :(

Missing the first hour and a half of the game cause I will be stuck at work #cantgetworse #depressed

Figure 3.1. Example posts from the CLPsych 2015 dataset for Depression and PTSD. Only users' posts are shown as labels were omitted to protect their privacy. The example submissions include positive and control users.

Table 3.1. Summary of CLPsych 2015 collection (Twitter). The activity period represents the number of days passed from the first to last the document collected for each user. On average, a user's corpus spans over a period of roughly one year. The oldest documents in the collection date from the middle of 2008, while the latest ones are from 2014.

	Depression	PTSD	Control
# of Users	477	396	872
# of Documents	1,131,997	919,131	1,978,121
Avg. # of Documents/User	2373.73	2321.64	2268.51
Avg. # Words/Document	13.9	16.5	13.8
Avg. Activity Period (Days)	≈ 379	≈ 479	≈ 460

3.2.2 eRisk

The *Early Risk Prediction on the Internet* (eRisk[3]) [119; 120; 121; 122] lab has been running since 2017 as part of the Conference and Labs of the Evaluation

[3] See http://erisk.irlab.org/

Forum (CLEF). Its main goal is to promote the development of reusable benchmarks for evaluating early risk-detection systems (ERDS). Unlike traditional risk-assessment and decision-making applications, a key requirement in the design of ERDS is that, besides the correctness of the decisions, the delays should be also taken into account. The various test collections created for the different tasks that were executed over the years consist of sets of documents posted by users on the popular social website Reddit.

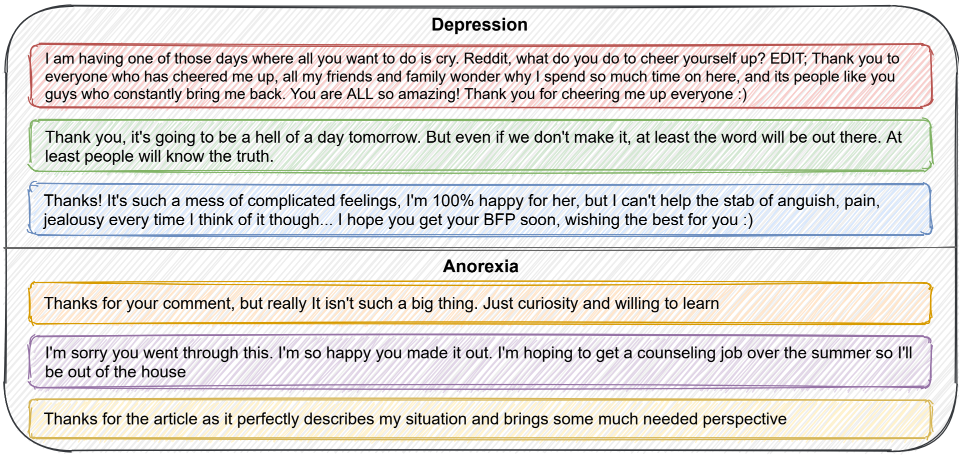
Reddit is a social news aggregation, web content rating, and discussion website. It consists of various *subreddits*, each of which focuses on a different topic. Users can post pictures, web links, or other types of content. On the contrary to Twitter, Reddit does not enforce any extreme restrictions on the length of posts.

In order to build more realistic collections, control groups also contain users who are active on Reddit's communities devoted to discussing about mental disorders or psychological issues, but are not suffering from any of them. For instance, a mental health practitioner giving support to other forum members or people interested in the subject because they have some relative suffering from certain disorder. This type of control users makes the datasets realistic and challenging because their topics of interest are highly related to those of the positive group. In this way, any automatic method designed to detect signs of psychological risks should not be merely based on topic-like classification but, instead, it needs to incorporate more subtle forms of evidence.

For each user up to their most recent 2000 submissions were retrieved and included in each corpus. It should be noted that on Reddit, users submit content in the form of posts and comments. While posts are used to start an online conversation (called a *thread*), comments are nested responses to comments or posts. Reddit's API supplies a maximum 1000 posts and 1000 comments per user. An example of the submissions included in the collections is shown in Figure 3.2. When creating the collection, eRisk's organisers were interested in retrieving a complete view of users' language, which includes discussions and concerns about a wide range of topics. For this reason, all available submissions were included in the collection irrespective of the subreddit in which they were published. A summary of the eRisk collections is shown in Table 3.2.

3.3 Self-assessment Inventories

Self-assessment inventories comprise a set of formalised psychological tests, also known as questionnaires, developed to provide an objective and standardised measurement of a sample of human behaviour [231]. These psychometric tests



Depression

I am having one of those days where all you want to do is cry. Reddit, what do you do to cheer yourself up? EDIT; Thank you to everyone who has cheered me up, all my friends and family wonder why I spend so much time on here, and its people like you guys who constantly bring me back. You are ALL so amazing! Thank you for cheering me up everyone :)

Thank you, it's going to be a hell of a day tomorrow. But even if we don't make it, at least the word will be out there. At least people will know the truth.

Thanks! It's such a mess of complicated feelings, I'm 100% happy for her, but I can't help the stab of anguish, pain, jealousy every time I think of it though... I hope you get your BFP soon, wishing the best for you :)

Anorexia

Thanks for your comment, but really It isn't such a big thing. Just curiosity and willing to learn

I'm sorry you went through this. I'm so happy you made it out. I'm hoping to get a counseling job over the summer so I'll be out of the house

Thanks for the article as it perfectly describes my situation and brings some much needed perspective

Figure 3.2. Example submissions from the eRisk dataset for Depression and Anorexia. Only users' submissions are shown as labels were omitted to protect their privacy. The example submissions include positive and control users.

have been extensively used as a reliable way to collect high-quality data from several sources, including online ones [83; 108]. In the case of CPA, in addition to provide a language sample (such as an essay or a set of utterances) individuals participating in the data collection process are usually requested to compile a standard self-report questionnaire on the Big Five to collect the ground truth. In essence, they are asked to assess on a Likert scale how well their personality matches a series of descriptions. Examples of these questionnaires are the BFI [97] (Big Five Inventory), the S5 [107] (*Short Five*) or the TIPI [77] (Ten Item Personality Measure). The collections described below are used in Chapter 6.

3.3.1 EAR Conversations

Mehl et al. [140] conducted a user study to shed light on how personality manifests and is perceived in everyday life. To this end, they created a dataset comprised of a set of daily-life conversation extracts from 96 subjects wearing a device called Electronically Activated Recorder (EAR) for a span of two days. To guarantee the anonymity and privacy of the participating subjects, random snippets of conversation were recorded, and only the subject's conversation utterances[4] were transcribed, hindering any possibility to reconstruct complete con-

[4] In the context of this thesis, an *utterance* is the verbatim transcription of a spoken word or statement.

Table 3.2. Summary of eRisk 2018 and 2019 collections (Reddit). The activity period represents the number of days passed from the first to last the document collected for each user. On average, a user's corpus spans over a period of roughly one year and half. The oldest documents in the collections date from the middle of 2006, while the latest ones are from 2017.

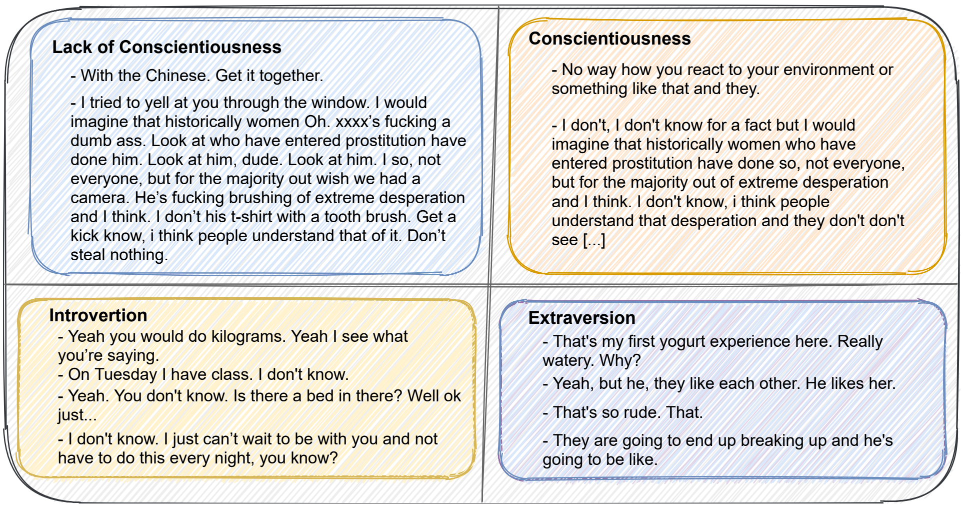
	Depression		Anorexia		Self-Harm	
	Positive	Control	Positive	Control	Positive	Control
# of Users	214	1,493	61	411	41	299
# of Documents	89,999	982,747	24,776	227,219	7,141	161,886
Avg. # of Documents/User	420.5	658.2	406.16	552.84	174.17	541.42
Avg. # Words/Document	45.0	35.3	64.6	31.4	39.3	28.9
Avg. Activity Period (Days)	≈ 658	≈ 661	≈ 799	≈ 654	≈ 504	≈ 785

versations. This setting is desirable for our goal, since we are interested only in assessing the personality of a subject who could potentially be interacting (e.g. chit-chatting) with a conversational agent, and therefore only the subject's conversation utterances are useful for our purposes. It should be noted that only the conversation transcripts and no other material, such as the original recordings, are available. An example of the utterances included in the collection is shown in Figure 3.3. It is noteworthy, that to date there are no other datasets which provide conversation utterances along with personality ratings on the Big Five Inventory.

The conversation extracts are accompanied by self and observer ratings of personality. The self-ratings were obtained by asking participants to compile a standard self-report questionnaire. Complementary, 18 independent judges, who listened to all of a participant's audio recordings, were requested to render a personality rating following the descriptions of the BFI [97]. The outcome of each observer was averaged to obtain the final personality scores. The dataset contains 118,259 words and 15,269 conversation utterances. Table 3.3 depicts the statistics of the conversations dataset. It is noteworthy, that to date there are no other datasets which provide conversation utterances along with personality ratings on the Big Five.

3.3.2 Stream-of-Consciousness Essays

Pennebaker et al. [171] conducted a study to gain insights on how various psychometric language variables, such as those included in hand-tuned dictionary-



Lack of Conscientiousness

- With the Chinese. Get it together.

- I tried to yell at you through the window. I would imagine that historically women Oh. xxxx's fucking a dumb ass. Look at who have entered prostitution have done him. Look at him, dude. Look at him. I so, not everyone, but for the majority out wish we had a camera. He's fucking brushing of extreme desperation and I think. I don't his t-shirt with a tooth brush. Get a kick know, i think people understand that of it. Don't steal nothing.

Conscientiousness

- No way how you react to your environment or something like that and they.

- I don't, I don't know for a fact but I would imagine that historically women who have entered prostitution have done so, not everyone, but for the majority out of extreme desperation and I think. I don't know, i think people understand that desperation and they don't don't see [...]

Introvertion

- Yeah you would do kilograms. Yeah I see what you're saying.
- On Tuesday I have class. I don't know.
- Yeah. You don't know. Is there a bed in there? Well ok just...
- I don't know. I just can't wait to be with you and not have to do this every night, you know?

Extraversion

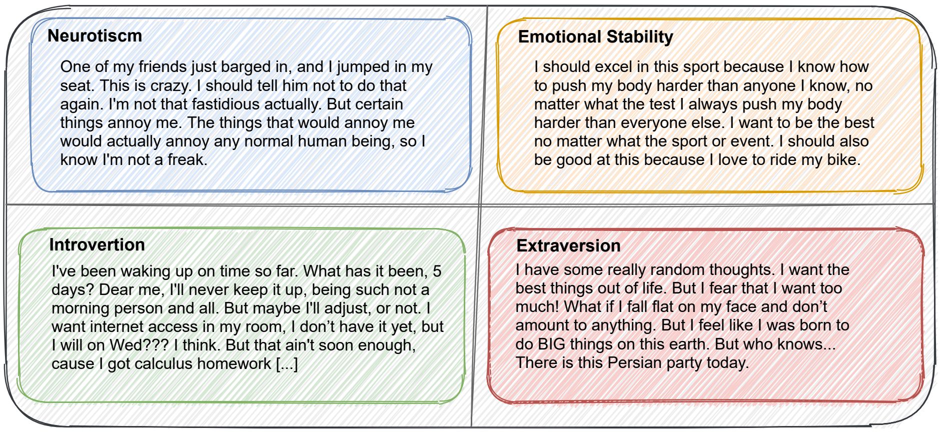
- That's my first yogurt experience here. Really watery. Why?
- Yeah, but he, they like each other. He likes her.
- That's so rude. That.
- They are going to end up breaking up and he's going to be like.

Figure 3.3. Snippets from the EAR conversations dataset, for participants with large scores on conscientiousness, lack of conscientiousness, extroversion and introversion. Only the participants' utterances are shown.

based tools as LIWC, correlate with five-factor measures of personality, with various demographics measures and with several health-related behaviour markers. Their ultimate goal was to examine how language use can reflect individuals differences, in particular personality styles. To this end, they collected a corpus comprising 2, 479 essays. Participating subjects were asked to think about their thoughts, sensation, and feelings on the spot and write whatever came to their mind for 20 minutes. An example of the essays included in the collection is shown in Figure 3.4. The dataset contains 1.9 million words. Moreover, individuals were requested to compile a standard self-report questionnaire on the Big Five. Table 3.4 depicts the statistics of the essays dataset. We decided to also include this dataset among our data resources, as later shown in Section 6.3, because we are interested in studying and comparing the manifestation of personality through spoken and written language. Therefore, we strive to discover linguistic cues of personality from both sources.

Table 3.3. Summary of the EAR conversations dataset.

Trait	# of Users	Avg. # of Utt./User	Avg. # Words/Utt.
AGR	47	43.23	10.85
-AGR	49	43.94	10.89
CON	58	39.33	10.37
-CON	38	50.18	9.03
EXT	48	48.08	10.71
-EXT	48	39.17	8.47
NEU	46	45.89	10.72
-NEU	50	41.54	10.23
OPN	47	47.60	10.68
-OPN	49	39.82	9.10



Neurotiscm

One of my friends just barged in, and I jumped in my seat. This is crazy. I should tell him not to do that again. I'm not that fastidious actually. But certain things annoy me. The things that would annoy me would actually annoy any normal human being, so I know I'm not a freak.

Emotional Stability

I should excel in this sport because I know how to push my body harder than anyone I know, no matter what the test I always push my body harder than everyone else. I want to be the best no matter what the sport or event. I should also be good at this because I love to ride my bike.

Introvertion

I've been waking up on time so far. What has it been, 5 days? Dear me, I'll never keep it up, being such not a morning person and all. But maybe I'll adjust, or not. I want internet access in my room, I don't have it yet, but I will on Wed??? I think. But that ain't soon enough, cause I got calculus homework [...]

Extraversion

I have some really random thoughts. I want the best things out of life. But I fear that I want too much! What if I fall flat on my face and don't amount to anything. But I feel like I was born to do BIG things on this earth. But who knows... There is this Persian party today.

Figure 3.4. Extracts from the Stream-of-Consciousness essays dataset, for participants rated as extremely neurotic, emotionally stable, introvert, and extrovert.

Table 3.4. Summary of the Steam-of-Consciousness Essays dataset.

Trait	# of Users	Avg. # of Sen./Essay	Avg. # Words/Essay
AGR	1254	49.45	283.49
-AGR	1214	49.28	283.25
CON	1233	51.41	288.38
-CON	1235	47.33	278.38
EXT	1277	47.88	281.67
-EXT	1191	50.96	285.20
NEU	1310	49.13	281.97
-NEU	1158	49.63	284.96
OPN	1272	50.40	289.26
-OPN	1196	48.27	277.11

Part I

Online Mental State Assessment

Chapter 4

Depression Detection and Tracing via Social Media Interactions

4.1 Introduction

During the last decade, the recognised importance of mental health has motivated the search for cutting-edge and innovative computational methods for identifying the onset of mental disorders at early stages. This is particularly important since many cases still go undetected and the lack of a timeliness treatment can lead to disability, psychotic episodes, thoughts of self-harm and, at its worst, suicide.

Thanks to their increasing popularity, online social media platforms have provided new opportunities for innovative methods for detecting different mental disorders, like depression. Moreover, the availability of such user-generated content can enable an exploratory screening process to automatically identify people who might be struggling with psychological problems, provide a preliminary assessment and, if needed, call for professional action. By leveraging user-generated content, risk-assessment and decision-making technologies have the potential to really make a difference by offering low-cost unobtrusive mechanisms for early screening of mental disorders at large-scale.

In this chapter, we take a few steps forward on this matter and present an initial attempt to exploit the latent semantic structure of social media users' textual posts to identify individuals potentially at risk of depression. Semantic structure analysis has been previously applied to determine a reduction in the semantic coherence in patients who suffered from schizophrenia. Such analysis has been shown to achieve a diagnostic accuracy comparable to clinical ratings [58].

Despite the great importance of mental health research through the analysis of online social media activities, datasets for the automatic detection of depression are limited and resources are very scarce [83]. This is mainly due to the need for the collection of large amounts of labelled data which is usually a complex and time-consuming endeavour. Motivated by this, we present in this chapter a weak-supervision framework for collecting such data. As a result, design of data-driven solutions to this problem becomes feasible. To the best of our knowledge, this is the first step towards automatically deriving large samples of data for the study of depression on online social media settings. We empirically validate our methodology and show it can be effectively used for automatically collecting posts samples. Furthermore, we release the dataset created and present a benchmark to foster the research on data-driven approaches for automatic identification of depression.

Our contributions in this chapter can be summarised as follows:

1. We build a novel an early risk-assessment system that measures the semantic proximity between user's textual posts and a set of words with topical relevance to depression and use such information to identify the onset of depression.

2. We perform various experiments to assess the effectiveness of the system trained on the semantic proximity features on a real-world dataset and analyse its performance. In particular, we study the effect of the decision threshold on its effectiveness and we explore the temporal spread of the cues that indicate the onset of depression.

3. We present a methodology for automatically gathering depression and non-depression post samples from weak-supervision signals[1].

4. We introduce the dataset created following the proposed methodology which can advance research on depression detection.

5. We build a series of depression post-classifiers using the automatically gathered data. We present a range of models for this purpose to serve as benchmarks and points of reference for researchers who are interested in using it.

6. Based on the developed models and the gathered data, we present a case study showing the potential of this research to identify users at risk of depression.

[1] Code available at `https://github.com/earissola/umap20`

The remainder of the chapter is organised as follows. Section 4.2.1 details how semantic proximity is computed to derive a set of features which are later used to build an early depression assessment system. Experimental design and classification results analysis are presented in Sections 4.2.2 and 4.2.3, respectively; Section 4.3.1 outlines the methodology designed to automatically select posts providing evidences of depression and characterises the dataset created; the empirical evaluation conducted to validate the proposed methodology is described in Section 4.3.2; a case study of depression using time-series analysis to demonstrate the potential use that the depression post-classifier could have is presented in Section 4.3.3; Section 4.4 concludes the chapter.

4.2 Semantic Proximity as a Proxy for Depression

As stated in the introduction (Chapter 1), research on psycho-linguistics has provided evidence that the ways in which people use words could act as a reliable source to assess a wide array of behaviours [172]. As a matter of fact, language attributes provide a unique *window* into thoughts and feelings, enabling direct assessment of mental-state alterations [46; 22]. In this section, we show how semantic structure analysis can be applied to spot early traces of depression from social media users' submissions. To this aim, we propose to leverage a methodology previously used to assess how mental-state changes can be detected after drug induction and which has achieved a diagnostic accuracy comparable to clinical ratings on the assessment of the consequences of schizophrenia [58].

4.2.1 Feature Extraction

Recall from Chapter 2.1, that meaning can be understood as emerging from mutual dependencies of words within the language. Semantically related words cooccur in texts with coherent topics at a higher frequency than unrelated words. Using this property, the relation between two words can be quantitatively measured by the frequency of the co-occurrence patterns that each of them present. Latent Semantic Analysis (LSA) [53] is an associative model that captures the meaning of words by means of linear representations in a high-dimensional semantic space. The semantic content of a word is encoded as a vector and this vectorial representation can be used to estimate how *semantically close* other words are.

Bedi et al. [14] quantified semantic and structural facets of speech to assess how mental-state changes can be detected after drug induction. They examined a set of transcribed interviews from individuals who had been administered with different drugs and discovered that effectively speech semantic content is affected after drug intoxication. They concluded that such semantic alterations can be accurately used to discriminate between the different drugs tested. Social media users might have their own way to express their moods or feelings by choosing different words, though these might have a high semantic similarity. Motivated by this, we are interested in studying whether this method can be tailored for identifying the onset and development of depression from social media submissions. To this end, we follow the method proposed in [14] and use LSA to compute the similarity of a set of depression-related words with respect to every word in a user's collection of posts. When the similarity of a word (ranging between 0 and 1) is above a threshold of 0.1 it is converted to 1 and when it is below the threshold to 0, producing a binary trace. As its original definition, we consider appropriate to keep the threshold of 0.1 as we want to omit words marginally related with depression without being too stringent. Subsequently, the mean value of this trace is computed for each depression-related word in the set. We use an LSA model trained on the TASA corpus, that is a collection of educational materials compiled by Touchstone Applied Science Associates. TASA is comprised of general reading texts believed to be common in the US educational system up to college, including a wide variety of short documents from novels, newspapers, and other sources. It includes $37,651$ documents and $12,190,931$ words, from a vocabulary of $77,998$ distinct words. In particular, we make use of the freely available[2] TASA 4 LSA model developed by Stefanescu et al. [220].

To build the set of depression-related words we use the words from the depression lexicon created in [38]. In essence, this lexicon is comprised of words closely associated with texts written by individuals discussing mental disorders or its symptoms on online settings and also include words related to names of medications. Additionally, we extend this set by looking for related concepts by running the query with the word “depression” in the well-known lexical database WordNet[3]. The final set is comprised of 96 words. These are all the words closely connected to the concept of “depression”, including *anxiety*, *withdrawal*, *delusions*, *blues*, *megrims*, among others. It is worth mentioning that during the preprocessing stage stopwords[4] and *hapax*[5] were remove due to the fact that LSA does not work properly for rare words and overly common words distort the

[2] Available at `http://semanticsimilarity.org/`

[3] See `https://wordnet.princeton.edu`

[4] Experiments were conducted using the NLTK's stopword list, avaialble at `https://gist.github.com/sebleier/554280`

[5] A *hapax legomenon*, or just hapax, is a word that occurs only once within a context.

representation of all words.

As stated in the introduction of this chapter, one of our goals is to estimate the presence of depression as early as possible. To this end, we use the 96 resulting values from the semantic proximity computation as features to train a Support Vector Machine classifier (SVM). We also consider the total number of words in a user's collection of posts as a feature. The rationale behind this is that previous studies has shown that increases in word count were positively associated with depression [190].

4.2.2 Experimental Design

Here, we outline the evaluation framework followed to assess the effectiveness of the model trained on the semantic proximity features. Moreover, we are also interested in sensing how different aspects involved in the training phase of the assessment algorithm, such as the decision threshold and the number of training examples, impact its on effectiveness and, especially, what can be learnt from the experiments to improve the knowledge on the task.

We use eRisk depression dataset (see Section 3.2.2) to conduct the various experiments. This collection is divided into a train and a test split. The train split contains the full history of a set of training users and, for example, it allows you to build predictive technology (e.g., text classifiers) from the entire threads of posts written by the training users. In the test split, each collection of user's posts is divided into ten chunks (in chronological order, based on the time each post was written). We employ the splits defined by eRisk's organisers which are shown in Table 4.1.

Table 4.1. Summary of eRisk dataset train and test splits.

	Train		Test	
	Positive	Control	Positive	Control
# of Subjects	83	403	52	349
# of Documents	30,851	264,172	18,706	217,665

Any system implementing an early risk detection algorithm is given one chunk of user data (oldest posts are given first) and it has to decide whether to classify a user as either depressed or control or to wait for the next chunk. Based on this decision, performance is assessed in terms of Recall (R), Precision (P) and F_1.

A key requirement for evaluating early detection algorithms is that, besides the correctness of the decisions, the delays should be also taken into account. Motivated by this, eRisk's organisers developed an error metric called ERDE [119] that penalises late decisions even when they are correct. In other words, when a system receives a chunk of data it can either emit a decision or choose to wait for the next chunk in order to take a *more informed* decision. If the system opts to wait then it gets a penalty for its delayed decision. In the case of ERDE, the delay is measured by counting the quantity of documents observed before providing the final answer. As a measure of error, the ultimate goal is to minimise it.

Let U be the set of users in the collection. Let d be a binary decision taken by a system for user u with delay k. Given the ground truth, d can be one of the following: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) or *False Negative* (FN). Taking into account these four cases, *ERDE* is defined as:

$$ERDE_o(u) = \begin{cases} c_{fp} & \text{if } d \text{ is FP} \\ c_{fn} & \text{if } d \text{ is FN} \\ c_{tp} \cdot lc_o(k) & \text{if } d \text{ is TP} \\ 0 & \text{if } d \text{ is FN} \end{cases}$$

The factor $lc_o(k)$ encodes a cost associated with the delay taken in spotting a TP, and it is defined as a monotonically increasing function of k:

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}}$$

The parameter o governs the point in which the cost starts to rapidly increase. Somehow, it represents a sort of *urgency* in detecting positive cases, and the lower the value of o, the higher is the urgency in identifying TP cases. Given $|U|$ individuals in a collection, a system has to take $|U|$ decisions and its final ERDE is defined as the mean error of computed over the entire users set. The proper values for the cost weights should be chosen considering the application domain and the practical consequences of detected and undetected cases. For example, late detection can be considered as equivalent to not identifying the case at all and, thus, we should set c_{tp} to be equal to c_{fn}. Thus, eRisk organisers decided to set c_{fn} and c_{tp} to 1 and fixed c_{fp} according to the proportion of positive cases in the test data.

4.2.3 Sensitivity Analysis and Temporal Spread of the Cues

The eRisk dataset is built for the time-dependent identification of the early signs of depression. Using the entire history of posts written by each of the training users we trained an SVM based on the 97 features outlined in Section 4.2.1. This trained classifier is then used at test time for the detection of depression on each separate chunk of data corresponding to a time slot. When a chunk of data is received, our model classifies a user as depressed only when the estimated class membership probability exceeds a certain threshold. Otherwise, the decision is delayed until the next chunk arrives. The probabilities estimates are calibrated using Platt scaling [174]. In essence, it works by fitting a logistic regression model to the SVM's scores.

We run different experiments to analyse how the value of the threshold affects the effectiveness of the early detection. The values we evaluate range from 0.5 to 0.9. It is noteworthy that, at this initial stage, we use the same threshold for every user, while as shown later it should be different. Furthermore, given that the classes in the corpus are not balanced, we study the effect of undersampling the majority class, i.e., reducing the size of the non-depressed user set at training time. Table 4.2 presents the results obtained after conducting the various experiments. In order to enhance clarity and to avoid overloading the table, we only include two values in the evaluated range which we consider representative for the purpose of this analysis.

A first observation reveals that when either the threshold becomes larger or the number of non-depression examples in the training set gets smaller recall improves. Conversely, precision diminishes. However, it should be noted that ERDE starts to grow. As the threshold becomes more conservative and stringent decisions are taken in the latter chunks. This delay is highly penalised by ERDE and highlights the trade-off between taking *early* decisions at the risk of making more mistakes or waiting to receive more data to take more informed decisions. To better understand this, Figure 4.1 depicts a boxplot based on the number of chunks that a particular configuration needed to take a decision on a certain user. Whereas under a low threshold (such as 0.5) most of the decisions are made in the first chunks, a larger one (such as 0.9) forces the system to wait until the lasts chunks.

Finally, we explore the temporal spread of the cues that indicate the onset of depression. Figure 4.2 depicts two examples of this analysis. Each figure presents a comparison of two users (randomly selected), all of them affected by depression. We observe that the evidence that some users show is very close to the threshold (black dotted-lined) but does not surpass it until a breaking point

Table 4.2. Sensitivity analysis on the effectiveness of semantic proximity features for the early assessment of depression. ERDE and F_1 are considered to gauge the effect of varying the decision threshold and undersampling the majority class (non-depressed). The best performance figures achieved for each metric are highlighted in bold.

Non-depressed	$threshold = 0.5$			$threshold = 0.7$		
(Set size)	$ERDE_5$	$ERDE_{50}$	F1	$ERDE_5$	$ERDE_{50}$	F1
403	**12.62** %	11.67 %	0.19	**13.11** %	**11.67** %	0.32
200	12.63 %	**9.7** %	**0.45**	14.68 %	12.35 %	**0.45**
100	15.22 %	10.85 %	0.34	27.88 %	24.05 %	0.39
50	20.08 %	14.48 %	0.24	28.61 %	22.48 %	0.26

(Subject B in the figure and identified with blue dots). This means that the user might show some signs of depression, but these are not severe yet and could easily disappear. Therefore, more chunks need to be processed in order to identify the true onset of depression. Conversely, there are other cases wherein after processing the first chunks we can see that there is already sufficient evidence, as the line consistently surpasses the threshold, to conclude that the user is rapidly developing depression (Subject A in the figure and identified with orange dots).

These observations suggest that effectiveness could be enhance if the threshold is defined on a user-dependent basis, to capture the very subjective behaviour of each user. As a practical alternative, users with similar characteristics could be grouped in order to create different *stereotypes* or *profiles*. In this way, each group of users would have its own threshold. When a new users arrive they would be associated with the group that better suit their characteristics and the corresponding threshold would be selected. However, different issues need to be addressed in this case, such as how to set the thresholds for each group of users (criteria to follow) or how to deal with the cold-start problem.

4.3 Leveraging Weak-supervision Signals

In the previous section, we showed how the latent semantic structure of textual posts can be exploited to identify evidence that could suggest the onset of depression. To this aim, we use LSA to compute the similarity of a set of depression-related words with respect to every word in users' chronology of posts and derive

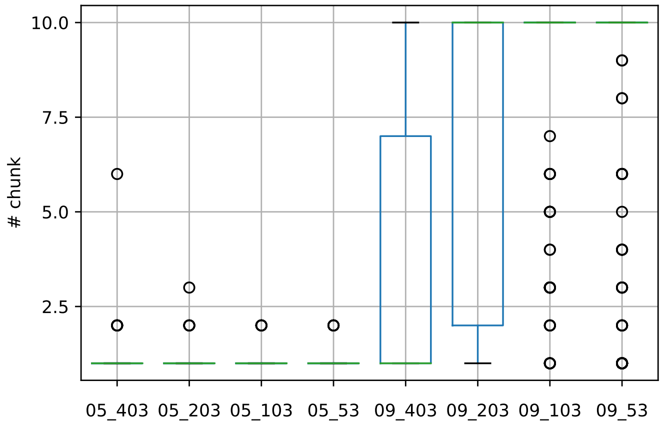


Figure 4.1. Number of chunks that a particular system configuration needed to take a decision on a certain user (i.e., to classify him/her as depressed). For instance, 05_403 refers to a configuration where the non-depressed class in the training set comprises 403 examples and the threshold is 0.5.

a set of semantic proximity features. We evaluate the effectiveness of the features and study the sensitivity of various parameters of the resulting assessment algorithm using eRisk collection.

One drawback of some of the studies conducted using the collections such as those built at eRisk or CLPsych is that users are usually represented by the concatenation of all their posting history [135]. This assumption might not be completely accurate. In fact, it is unlikely that every message posted by an individual potentially affected by a mental disorder could provide some evidence that could be used to identify the onset of such disorder [31]. If processed automatically, such messages might introduce unnecessary noise in the development and tuning of classification systems used for risk-assessment and decision-making.

Our goal is to build a post-level annotated dataset large enough to enable the development of robust detection models. Such models should be able to extract evidences from each message posted by a user who could be potentially tracked in order to anticipate the onset of depression. The strength and trend of such evidences can be utilised to detect the point where the mental disorder is starting to manifest [197]. Furthermore, post-level annotations provide a useful way to discover life circumstances related with mental disorders not captured by traditional depression diagnostic criteria [8]. Finally, we sought to alleviate the great

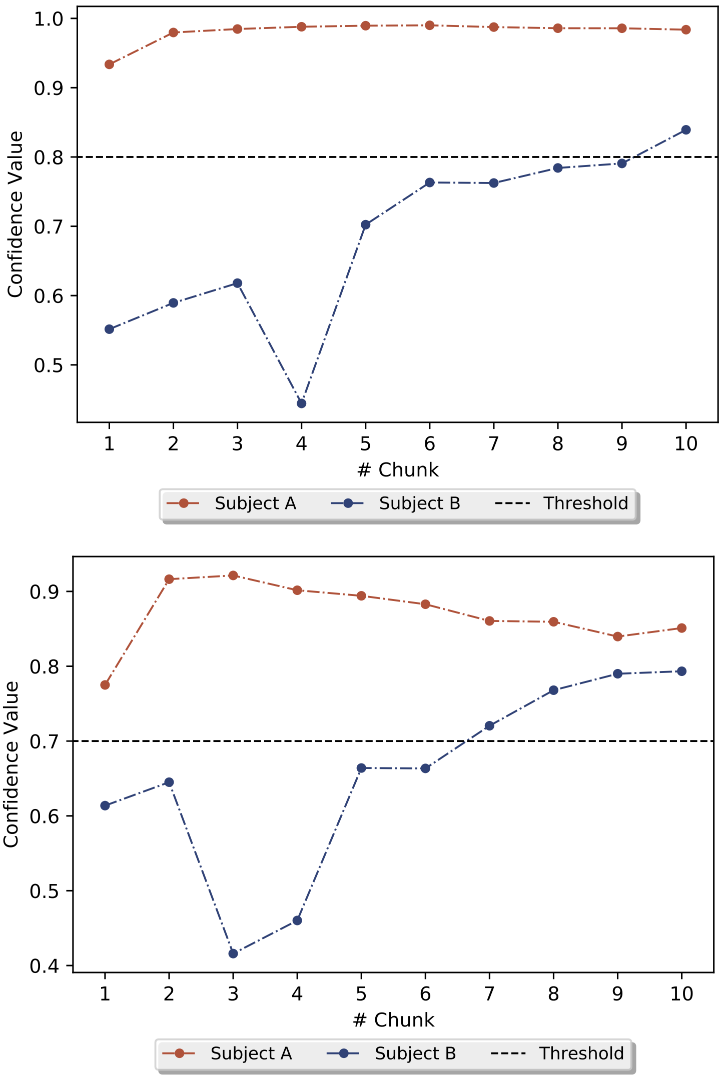


Figure 4.2. Temporal spread of the cues that indicate the onset of depression. Each figure is comparing different users affected by depression. While for some users the evidence already surpasses the threshold after the first chunks (Subject A identified with orange dots), in other cases more information is needed (i.e, more chunks) to determine the onset of depression (Subject B identified with blue dots).

effort and time that involves manually annotating individual posts by proposing a set of heuristics to automatically gathering samples of posts providing some evidence of depression. Thus, we are able to obtain higher-level supervision over unlabelled data.

4.3.1 Methodology

Here, we outline the methodology we propose to automatically generate large datasets of depression and non-depression posts. Consider a set of social media users on whom we have definitive knowledge that they are suffering from depression. This knowledge could come from a formalised psychological test (such as the PHQ-9 [109][6]) or it could be self-declared. Given a chronology of textual posts and based on previous findings in the literature, we propose different heuristics to characterise depression signs and use this information to automatically select posts for building the dataset.

Let D^+ be the candidate set of positive posts samples. We retrieve such posts from a set of users suffering from depression. Since the goal is to filter out less useful messages, we define two heuristics:

- Filtering posts by their sentiment polarity score.
- Filtering posts by their topical similarity with a *depression taxonomy*.

Let D^- be the control posts samples, which are posts not providing any reference to depression signs. Such posts are randomly collected from a set of users who are not affected by the mental disorder. In the following subsections we explain in detail how we apply the heuristics to create the dataset.

Sentiment Polarity Score (H_s): Sentiment analysis constitutes the computational study of opinions, feelings and subjectivity in text. Research has shown that the sentiment polarity score (also referred to as *semantic orientation* or *valence*) of a post can be linked with the emotions evoked by a piece text [164]. The study of sentiment through lexical analysis has been previously applied in different psychology-related tasks, such as memorability [10], and particularly in the domain of automatic depression detection [38; 165; 95]. Based on these findings and our own empirical analysis, we hypothesise that when this value is negative (i.e., below zero) it can be a good indicator of distress or unhappiness, especially when the posts are written by users experiencing depression. The sentiment polarity score of a post ranges from -1 to 1 and is calculated using a lexicon-based approach. Given the computed scores, the posts in D^+ are sorted from lower to higher polarity retaining those messages whose score is less/equal to 0. We use TextBlob[7] to obtain the polarity score of a post.

[6] *PQH* stands for *Patient Health Questionnaire*.

[7] See `https://textblob.readthedocs.io/en/dev/index.html`

According to the *Diagnostic and Statistical Manual of Mental Disorders* [8] (DSM) depressive moods are characterised by the predominance of two emotions: *sadness* and *disgust*. Based on psychology literature, we refine this heuristic by mandating the presence of words related with these emotions in the posts. The goal is to clean the set of messages obtained after filtering by sentiment polarity score from *false positive* cases, such as "*That's the worst name I ever heard*" (this message also has a low sentiment polarity score). To this end, we use the terms included in EmoLex [151]. Each word in this dictionary is associated with the emotions it evokes.

Additionally, we compute for each post a *sadness score*. To calculate this score we use the NRC Affect Intensity Lexicon (AIL) [150] which associates words with real-valued scores of intensity for different emotions. Given a word and an emotion, the intensity ranges from 0 to 1, where an intensity of 1 means that the word evokes the highest intensity of that emotion. The sadness score is computed as the average of the intensities of the words which evoke "sadness" found in the text of a post. It should be noted that the lexicon only provides scores of intensity for four basic emotions: anger, fear, joy, and sadness. The rationale of calculating this score is to be strict with the words considered to evoke sadness. Moreover, the AIL includes words that might not predominantly convey a certain emotion and still tend to co-occur with words that do. For instance, the words *failure* and *death* describe concepts that are usually accompanied by sadness and, thus, they denote some amount of sadness.

Hence, following heuristic H_s a post in D^+ is considered a valid candidate for the final set if it contains at least one word related with either sadness or disgust and a sadness score higher than a certain threshold. We decide to use a threshold of 0.1 as we want to omit words marginally related with depression without being too stringent.

Topical Similarity (H_t): Considering the success that topic models have shown two extract meaningful features to detect depression [194], we define H_t as a heuristic in which less useful posts in D^+ are filtered out based on their topical similarity with a *depression taxonomy*. We first build this taxonomy using the lexicon released by Choudhury et al. [38]. This lexicon consists of words closely associated with texts written by individuals discussing depression or its symptoms on online settings and also include words related to names of medications. In essence, the authors built a lexicon comprised of the tokens with high probability of being associated with postings from individuals discussing about depression-related themes or its symptoms. To this end, they mined a sample of a snapshot of the "Mental Health" category of Yahoo! Answers.

We enlarge this set by collecting all possible online vocabularies with concepts and terms commonly related to depression. The goal is to select from all these words those that are considered closely associated with this mental disorder and define a compact but accurate list. For instance, generic terms present in the dictionary such as *relationship* or *family* might introduce significant noise as they are in general very frequent terms in any context and not only when writing about depression-related themes. On the other hand, terms like *grief* and *sorrow* are normally used with higher frequency by individuals suffering from this mental disorder. With the aid of three human experts (one experienced clinical psychologist and two computational linguistic experts, none co-authoring this paper) we selected a subset of the words and produced a list of 78 depression-related terms which comprised the depression taxonomy, shown in Figure 4.3.

```
150mg
40mg
addictive
adhd
antidepressant
anxiety
appetite
attacks
beautiful
bible
blurred
care
chemical
church
clinical
counteract
dating
delusions
diagnosis
discuss
dizziness
doctor
doses
drowsiness
drowsy
drugs
dysfunction
episodes
fatigue
friends
god
hate
headache
heaven
hell
helped
helpful
hold
hospitalization
imbalance
inhibitor
insomnia
irritability
jesus
leave
lord
love
medication
nausea
nervousness
neurotransmitters
patients
pill
pills
prescribed
prescriptions
psychosis
psychotherapy
relationship
religion
sedative
seizures
severe
sexual
side-effects
sleep
someone
stimulant
style
suffer
suicidal
swings
therapy
tolerance
toxicity
weaned
weight
withdrawal
```

Figure 4.3. Depression-related terms included in the depression taxonomy.

Topic models are hierarchical Bayesian models of discrete data, where each topic is a set of words drawn from a fixed vocabulary representing a high-level concept. The Latent Dirichlet Allocation [17] (LDA) is a topic model that discovers latent topics present in a given text collection. LDA represents each topic as a probability distribution over words in the documents. For each document, a multinomial distribution θ over topics is randomly sampled from a Dirichlet function with parameter α (which influences the shape of the distribution). We use LDA gensim's[8] implementation to obtain the topics that emerged from the posts in D^+ and compute the cosine similarity between each post and the depression taxonomy. The goal is to obtain a ranked list of the posts according to their similarity with the taxonomy of depression. Thus, the higher ranked posts are those with the highest association to depression. In order to compute the ranked list, each post is mapped into the LDA space and compared with the depression taxonomy as a reference point. Given that the number of topics (K) is an unknown parameter in the LDA model, we follow the method of Griffiths et. al. [79]. Basically, it consists in keeping the LDA parameters (commonly known as α and η) fixed, while assigning different values to K and run the LDA model each time. We selected the model that minimise $log\,P(W|K)$, where W contains all the words in the vocabulary. This procedure is performed until the optimal number of topics has been obtained. In our case, we train the LDA model with K equals to 50 up to 200 at steps of 50, where optimal value is 200.

Resulting Dataset: In order to automatically select posts of our dataset we use the eRisk depression collection (see Section 3.2.2). Recall that this publicly available collection consists of two groups of Reddit users, namely depressed and non-depressed. Labels are assigned to users but not to individual posts. It is noteworthy that the methodology we propose can be potentially applied to any collection where information about the depression status of the users is available. For example, in the case of eRisk collection some users have self-declared being diagnosed with depression.

Following H_s the candidate posts in D^+ are ranked according to their polarity and sadness scores. We select the top 3500 posts to build the final set, which is defined as D_s^+. In a similar way, applying H_t we create a second set of posts, defined as D_t^+, where we select the top 3500 posts with the highest topical similarity to the depression taxonomy. Instead, D_s^- and D_t^- are two sets of 3500 posts each that are randomly sampled from set of users who are not diagnosed with depression and used them as control sets.

To have a better understanding on how the data is distributed, Figure 4.4 depicts the Gaussian kernel density estimation (KDE) of the automatically derived sets. We observe that there is some overlap between the curves. This intersection is desirable since it enables to improve the decisions in the boundary cases. We empirically found that selecting samples with no such overlap has a negative impact on the performance of a depression post classifier, caused by an overfit. For example, if the vocabularies obtained from the posts in D^+ and D^- are disjoint

[8] See https://radimrehurek.com/gensim/models/ldamodel.html

and too specific they hamper the classifier to learn a useful representation.

The goal of Figure 4.5 is to study whether there exists some relationship between the posts selected with H_s and H_t. A Pearson's correlation coefficient of 0.1484 (significant at p-value < 0.001) denotes a very small positive correlation between the topic similarity score and the polarity/sadness scores (top plot). We also analyse whether some relationship holds between the polarity and sadness scores both used in the definition of H_s. A Pearson's correlation coefficient of -0.1388, shows a small negative correlation (significant at p-value < 0.001) between them (bottom plot). In conclusion, there is no considerable correlation between variables analysed.

4.3.2 Empirical Validation

Here, we describe the process conducted to validate both the methodology and the resulting dataset. The goal is to determine whether the two automatically generated sets, created following H_s and H_t, can be used to train a classification model. Such model should be able to effectively distinguish posts providing some evidence of depression signs from those which do not.

Validation Set: We created a validation set of manually annotated posts. To this end, we randomly sampled a total of 400 posts, 200 from each class, in the eRisk 2018 collection[9] and asked three human experts to label them. The annotation process followed a similar procedure to that defined by Moreno et al. [153]. Annotators were asked to determine which posts can be considered as a reference to depression following the DSM criteria. The references found should point to symptoms or feelings experienced by the individual and not by a third person. Furthermore, general comments about daily ordeals and common experience of having a bad day do not meet the criteria as a depression symptom. Each message was assigned with one of the following codes: (1) No depression reference is expressed; (2) One or more depression references are expressed; (3) Unable to make a judgement.

The three annotators achieved a pairwise Cohen's Kappa score ranging between 0.577 and 0.749. Achieving a high inter-rater agreement can be a difficult task. In some cases taking a decision is complex without any additional information. Posts were retained when two out of three annotators agreed on the label. The final validation set comprised of 55 positive posts (i.e., references to depression) and 93 control posts.

[9] These posts were removed from D^+ and D^-.

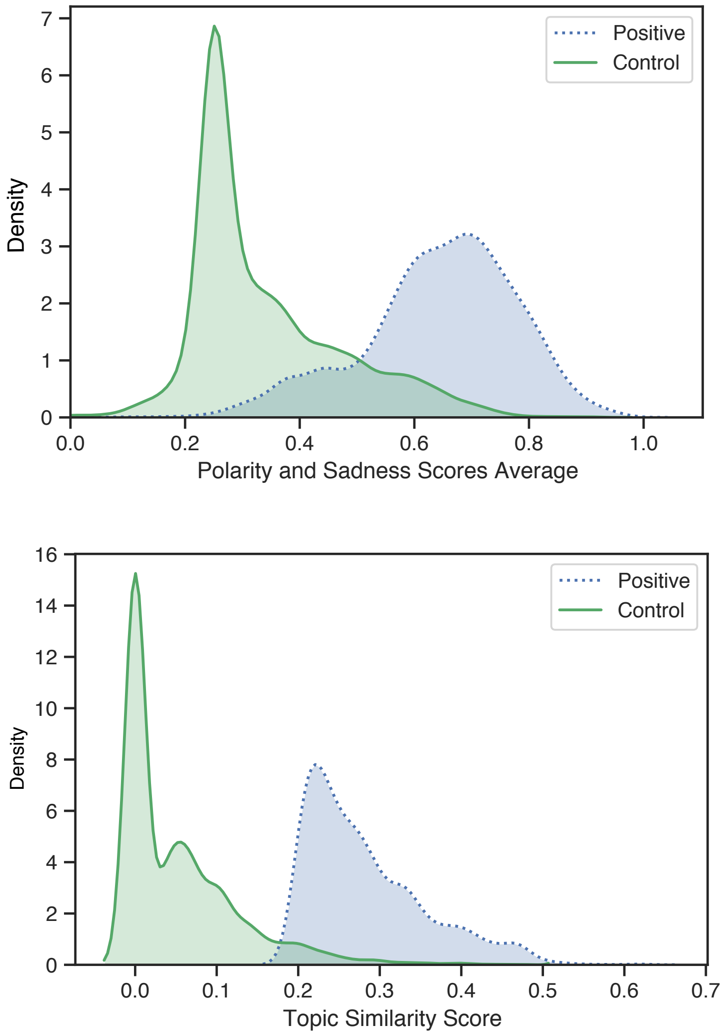


Figure 4.4. KDE computed for the automatically derived sets. On the top, candidate posts in D^+ are selected based on H_s. While on the bottom, the posts are selected according to H_t. In both cases, control posts samples are randomly picked from the set of users who are not affected by depression.

Benchmark Models: Once we have automatically derived a set of depression and non-depression post samples, we proceed to build a benchmark by developing a series of depression post-classifiers and evaluating various features to identify posts that show some evidence of depression signs from those which do not. As previously stated, sentiment analysis refers to the study of positive and nega-

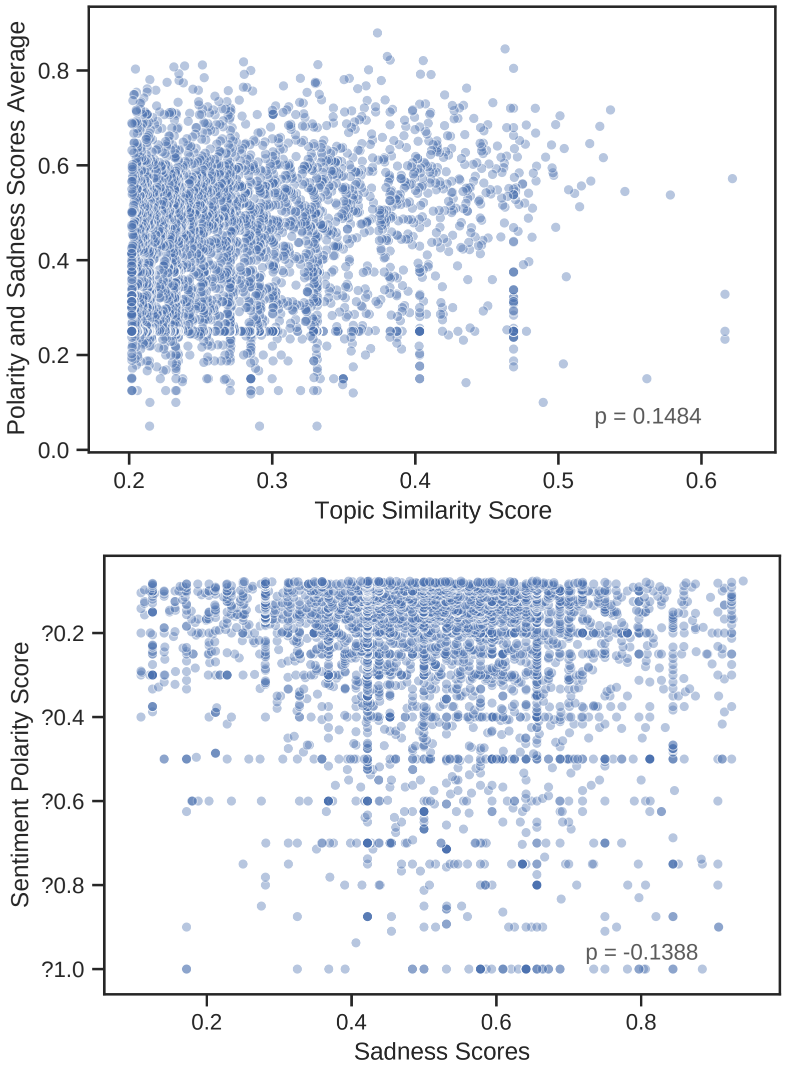


Figure 4.5. Correlation study. The top plot analyse the relationship between the posts selected following H_s and H_t, respectively. The bottom plot studies the association between sentiment polarity and sadness scores both used in the definition of H_s. Pearson's correlation coefficients are reported (both significant at p-value < 0.001).

tive feelings expressed in a piece of text. Determining whether a post contains an evidence of depression or not, can be thought as a specific case of sentiment analysis given that it involves the analysis of the emotions and mood communicated by the post. For this reason, we include in the benchmark a baseline that is commonly used for sentiment analysis [164]. This baseline is trained using only unigram binary features and an SVM (named SVM_Unigrams_B in Table 4.3). In addition, we considered three sets of features, bag-of-words, LIWC [222] and word embeddings (context-free and contextual). In the first case, each post is represented with the raw frequency of the unigrams extracted from the textual content of the posts (named Unigrams in Table 4.3). As a complement, we include a model variant where the unigrams are extended with four extra features, the word count, the polarity score, the sadness score (Section 4.3.1) and the happiness score [55] of the post (named Unigrams++ in Table 4.3). LIWC psychometric word categories counts collected from the posts.

To produce an embedding representation of the posts, we use the words embeddings obtained from the eRisk training partition, using word2vec [145] (named W2V in Table 4.3) and fastText [18] (named FTT in Table 4.3) methods. Furthermore, GloVe [173] pre-trained word embeddings are also considered (named GloVe in Table 4.3). For the sake of comparison, we set the number of dimensions of W2V and FTT word embeddings to 200 as that is the largest number available for GloVe. Finally, we also use Bidirectional Encoder Representations from Transformers (BERT) [54] to derive contextual language representations from the posts (named BERT_Embeddings in Table 4.3). Embedding representations of the words found in each post are averaged column-wised to obtain a k-dimensional representation of it.

The different features sets are used to train various Logistic Regression (denoted as LR in Table 4.3) classifiers. Let π be the probability that the response variable equals the case in which a post is a reference of depression given some linear combination of the predictor variables $(x_1, \ldots, x_p)$, for instance the frequency of the unigrams extracted. The g *logit* function is expressed as:

$$g(x_1, ..., x_p) = \ln(\pi/(1-\pi)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

This allows to model the response probabilities and estimates the parameters $(\beta_0, \beta_1, \ldots, \beta_p)$ by solving a regression problem. We choose this learning method since it has state-of-the-art effectiveness on a range of text categorisation tasks [67]. The LR classifiers are trained with L2 regularisation.

Experimental Results: Table 4.3 shows the results of the various benchmark models trained separately on each automatically derived set and evaluated on the validation set previously described. The effectiveness is evaluated using classic metrics such as Precision (P), Recall (R), F_1 and Area Under the Curve ROC (AUC).

The highest precision is achieved by the unigram-based features. In the case of the posts gathered based on H_s there is an improvement of 6% over the sentiment baseline. Overall, the use of word embeddings increases the recall, F_1

and the AUC performance. For instance, the use of GloVe pre-trained embeddings as well as task-specific word2vec embeddings provide an improvement over the baseline of more than 18% in Recall and 5% in both F_1 and AUC, for both automatically derived training sets. We observe that, overall a higher recall is achieved following heuristic H_s. While a higher precision is obtained when applying H_t. Given the nature of the task a higher recall is usually preferable. For instance, under certain circumstances, producing a considerable amount of false alarms (*false positives*) can be tolerated at the benefit of discovering most of the *real* cases. Finally, F_1 allows to conclude that any of the heuristics can effectively be used to automatically derived a set of posts samples and used such set to train a classification model. We hope this benchmark to serve as a starting point for further research.

Table 4.3. Benchmark models trained on the automatically derived sets and evaluated using the validation set (manually annotated posts). The best performance figures achieved for each metric are highlighted in bold.

H_s	**Precision** (%)	**Recall** (%)	F_1 (%)	**AUC** (%)
SVM_Unigrams_B (Baseline)	72.72	72.72	72.72	78.29
LR_Unigrams	**78.94**	81.81	80.35	84.45
LR_Unigrams++	75.92	74.54	75.22	80.28
LR_LIWC	64.78	83.63	73.01	78.37
LR_W2V	62.33	87.27	72.72	78.04
LR_FTT	68.05	89.09	77.16	82.17
LR_Glove	71.62	96.36	82.17	86.89
LR_BERT_Embeddings	72.97	**98.18**	**83.72**	**88.33**

H_t	**Precision** (%)	**Recall** (%)	F_1 (%)	**AUC** (%)
SVM_Unigrams_B (Baseline)	80.76	76.36	78.50	82.80
LR_Unigrams	80.39	74.54	77.35	81.89
LR_Unigrams++	**83.33**	72.72	77.66	82.06
LR_LIWC	65.00	94.54	77.03	82.21
LR_W2V	75.36	94.54	**83.87**	**88.13**
LR_FTT	69.86	92.72	79.68	84.53
LR_Glove	67.94	96.36	79.69	84.74
LR_BERT_Embeddings	73.61	**96.36**	83.46	87.96

4.3.3 Case Study of Depression using Time-Series Analysis

Recall from the beginning of the section that our ultimate goal is to foster the development of robust assessment algorithms to anticipate the onset of depression. Such models should be able to extract evidences from each message posted by a potentially tracked user. For this purpose, we proposed a series of heuristics to characterise depression signs and use this information to automatically selecting posts for building a dataset. Using such dataset, we presented novel methods for assigning depression scores to each post written by the user. Motivated by this, we show an important application of our depression post-classifier by conducting a case study for analysing users' posts over time. To accomplish this, we randomly chose eight users from the eRisk test partition, such that four belong to the positive class (i.e., depressed) and the other four belong to the control class (i.e., non-depressed). This case study can reveal how the moods of depressed/non-depressed users evolve over time.

Figures 4.6 and 4.7 illustrate the results of this case study by showing the smoothed signal of depression/non-depression (obtained from the confidence scores that our classifier assigned to each post) of the four positive subjects and the four control subjects, respectively. In particular, we use the model trained on the posts selected based on H_s and unigram features to conduct this user study. The signals generated by the confidence score produced by the classifier are smoothed by computing the moving average of the series with a window equal to the 10% of the user's posts. The horizontal line (dotted) represents the decision threshold which separates the depressed from the non-depressed posts for each user. We observe in the two figures that the behaviour of the positive and control users is very distinctive. In particular, the signals corresponding to the positive users are most of the time above the 0 decision threshold, except for the last individual. In this particular case, the individual shows in the beginning what can be defined as a "normal" behaviour until a turning point where signs of depression becomes evident. This could mean that the user has developed depression over time. Conversely, the signals corresponding to the control users always remain by far below the 0 decision threshold.

Another feature that is clear in the positive group is a much higher mood swing as compared with the control class. Such analysis is useful to anticipate the point where a user is developing depression and as we have observed can distinguish between depressed and non-depressed users effectively.

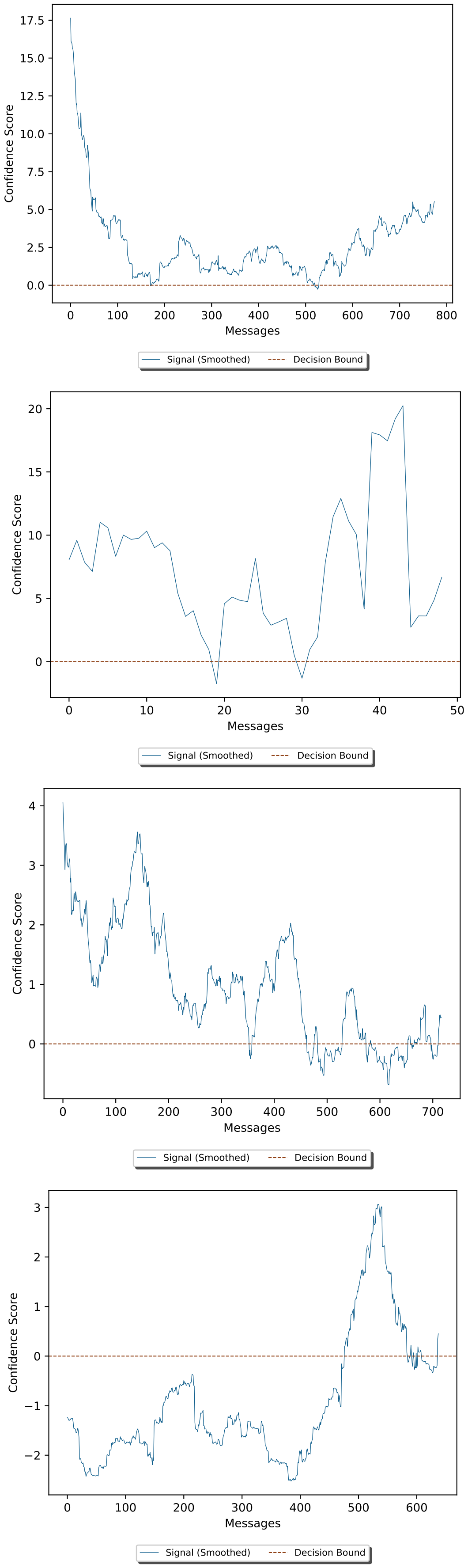


Figure 4.6. Smoothed signal generated for users in the positive class (Depressed).

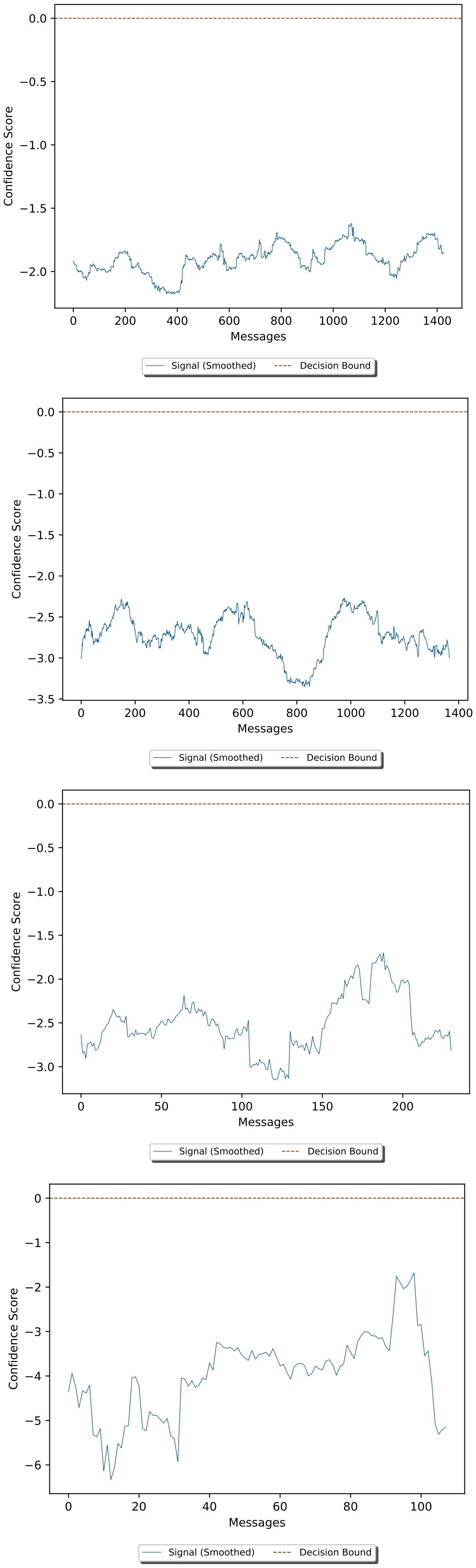


Figure 4.7. Smoothed signal generated for users in the control class (Non-depressed).

4.4 Discussion and Summary

Recently, there has been increasing research interest in the identification of mental state alterations through the exploitation of online digital traces. By leveraging user-generated content, language-based technologies have a great potential to provide low-cost unobtrusive mechanisms for early screening of mental disorders.

In this chapter, we showed how semantic structure analysis can be applied to identify traces of depression from social media users' submissions. To this aim, we leveraged a methodology previously employed to assess how mental-state changes can be detected after drug induction. Although the effectiveness achieved by our early risk-assessment system did not outperform the top-performing system presented at the eRisk 2017 edition [120], it fared modestly at early detecting individuals at risk of depression. In fact, our system yielded lower ERDE (both when o is equal to 5 and 50) than the most robust system at the cost of losing effectiveness in terms of F1. Yet, the results of the evaluation highlight the value of developing automatic screening assistants to aid mental health practitioners by providing prognostic information about individuals at risk of mental disorders on social media, like depression. In particular, outcomes from our temporal spread of the cues analysis spotlighted that performance could be boosted if the assessment algorithm's decision threshold is defined on a user-dependent basis, thus capturing the very subjective behaviour of each user. Also, we learnt that as model's threshold becomes more conservative and stringent, decisions are taken in the latter chunks. This delay is highly penalised by ERDE and highlights the trade-off between taking *early* decisions at the risk of making more mistakes or waiting to receive more data to take more informed decisions.

Early risk detection is challenging because multiple objectives are involved (ERDE, Precision, Recall), and therefore one would have to decide which metrics to focus when optimising the various parameters of a specific system. It becomes essential to understand the practical implications of using such systems in the clinical practice. For instance, one could consider the difference that the two evaluation measures of $ERDE_5$ and $ERDE_{50}$ could have in a real-life scenario. A mental health agency might decide to set the penalty costs on the consequences of late detection. This leads to a natural trade-off depending on whether the goal is to be more conservative and raise a handful of timely alerts or to maximise the recall at the cost of issuing false alarms. As concluded from our experiments, such systems should be customised for each user in order to provide more personalised and precise services. Implications resulting from our work serve as a starting point for future research dedicated to understanding the degree to which the latent depression patterns encoded in users' writing semantics could be leveraged to anticipate the onset of depression.

Finally, given the lack of available resources, such as datasets, that might hinder the development of cutting-edge technologies to assist health practitioners in their daily labour, we introduced a methodology to automatically gather post samples of depression by taking advantage of weak-supervision signals. Furthermore, we used the dataset derived to train models which are able to determine whether a post is conveying evidence of depression. Our results showed that this methodology is very effective for gathering large quantities of data from social media. Moreover, we showcased the potential of post-depression classifiers in identifying latent depression patterns via a case study using time-series analysis. In the light of this, we released the automatically created dataset, as well as the validation set of manually annotated posts, in order to contribute to the research on automatic detection of depression.

Chapter 5

Beyond Prediction: Mining Insights on Mental Disorders

5.1 Introduction

The proliferation of online social media platforms is changing the dynamics in which mental state assessment is performed [216; 33; 224]. Individuals are using these platforms on a daily basis to share their interests [246] and personal life events [104] as well as to disclose their feelings and moods [180]. With this in mind, online social media platforms have become promising means for researchers and health practitioners to proactively identify linguistic markers or patterns that correlate with mental disorders.

As previously stated, OMSA deals with the unobtrusive identification of mental state alterations through the analysis of online digital traces, in particular examining the patterns of language use. Gaining insights on the language and online behaviour of individuals affected by mental disorders could lead to identify new predictive markers up to now not considered in the medical literature, motivating new inquiries into the study of behavioural traits of mental health disorders as observed on social media. For this reason, we sought to gain new insights and extend the current knowledge on the manifestation of mental disorders to online settings, particularly social media platforms. Despite the collective efforts in the community to develop models for identifying potential cases of mental disorders on online social media, not much work has been done to provide insights that could be used by a predictive system or a health practitioner in the elaboration of a diagnosis. In this chapter, we present a series of analytical studies which aimed at gaining a better understanding on the language and behaviour that characterise individuals affected by mental disorders.

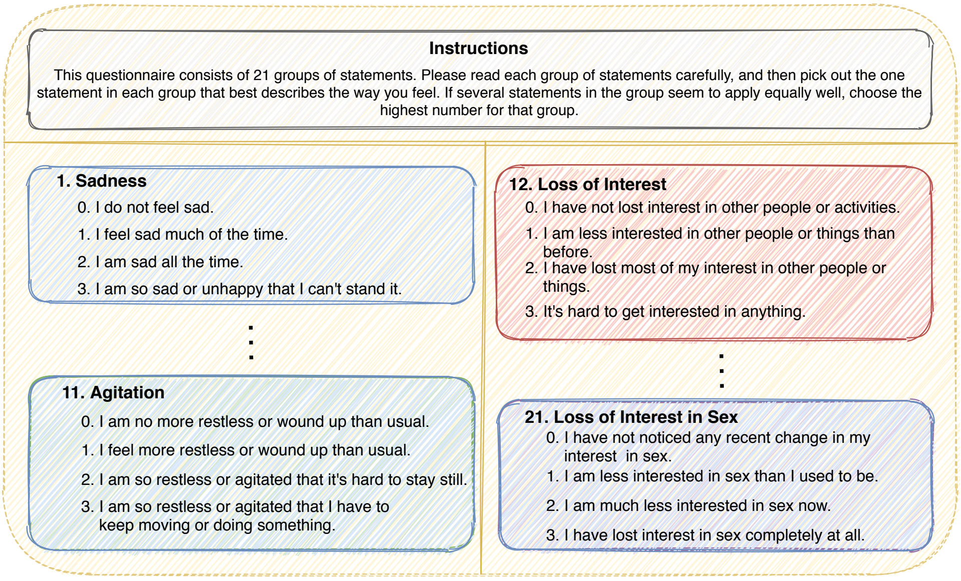
Initial efforts to address the automatic identification of potential cases of mental disorders on social media have mainly modelled the problem as classification [135; 201]. Researchers participating in eRisk and CLPsych workshops have examined a wide variety of methods to identify positive cases[1] [26; 27]; however, not much insight has been given as to why a system succeeds or fails. With such insights, the models and features used in those studies could be analysed and motivated more deeply. Therefore, we argue that even though achieving an effective positive detection performance is important, tracking and visualising the development of the mental disorder is equally relevant. In fact, an accurate detection system can be more useful if it provides a way of understanding the factors that lead to a certain decision. As argued by Walsh et al. [235], this issue, along with other factors, might prevent most of the risk-assessment and decision-making technologies from ever been used in real-life settings.

Bearing this in mind, we believe it is necessary to carry out experiments providing insights on how individuals' language attributes and online behaviour are distinctive among users suffering from mental disorders as well as between different disorders. We focus on several language dimensions as well as on the social engagement and posting trends of social media users. We especially emphasise the importance of finding innovative ways to *visualise the onset and development of a mental disorder*. Thus, systems oriented at visualisation for risk-assessment and decision-making could be complemented with preliminary step-by-step directions for practitioners to identify high-risk individuals based on statistical and visual analyses. Furthermore, it is essential for practitioners to count on a way to assess the degree of the differences identified between individuals; hence visual outcomes should not be presented as standalone objects but supported with the significance that the observed trends have.

The eRisk 2019 and 2020 editions have introduced a novel task aimed at exploring the feasibility of automatically estimating the severity of the multiple symptoms associated with depression, as measured by the Beck's Depression Inventory [12] questionnaire, BDI. The BDI belongs to a group of formalised psychological tests, also known as questionnaires or inventories, developed to provide an objective and standardised measurement of a sample of human behaviour [231]. These psychometric tests have been extensively used as a reliable way to collect high quality data from several sources, including online ones [38; 83]. In essence, the BDI is a self-report rating inventory for measuring characteristic attitudes and symptoms of depression. Through a set of 21

[1] Recall that the term *positive* refers to individuals who are potentially suffering from mental health concerns such as depression, anorexia, self-harm or post-traumatic stress disorder (PTSD). While *control* refers to individuals not affected by any of the aforementioned mental disorders.

multiple-choice questions it gauges the presence and intensity of feelings like sadness, hopelessness, self-dislike, social withdrawal, and loss of energy. The BDI relies on the theory of negative cognitive distortions as central to depression [12]. It was developed by collecting and combining patients' descriptions of their symptoms and using them to structure a scale which could reflect the severity of a given symptom [13]. An example of the questionnaire[2] is shown in Figure 5.1.



Instructions

This questionnaire consists of 21 groups of statements. Please read each group of statements carefully, and then pick out the one statement in each group that best describes the way you feel. If several statements in the group seem to apply equally well, choose the highest number for that group.

1. Sadness

0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.

⋮

11. Agitation

0. I am no more restless or wound up than usual.
1. I feel more restless or wound up than usual.
2. I am so restless or agitated that it's hard to stay still.
3. I am so restless or agitated that I have to keep moving or doing something.

12. Loss of Interest

0. I have not lost interest in other people or activities.
1. I am less interested in other people or things than before.
2. I have lost most of my interest in other people or things.
3. It's hard to get interested in anything.

⋮

21. Loss of Interest in Sex

0. I have not noticed any recent change in my interest in sex.
1. I am less interested in sex than I used to be.
2. I am much less interested in sex now.
3. I have lost interest in sex completely at all.

Figure 5.1. Beck's Depression Inventory (BDI).

As an initial outcome of the eRisk challenges, it was shown that it is feasible to automatically estimate the severity of some symptoms, as measured by BDI items, by extracting evidence from users' interactions on social media. However, the performance achieved by the participating teams is still modest and we are still far from an effective depression screening tool. As already stated, we consider that to gain new insights on the task and the data it is important to carry out experiments to get a further understanding on systems outcomes, in particular on the reasons they fail. Is it because of the lack of evidence about certain BDI items? Is it because the systems misunderstand the existing evidence? By answering these questions we could guide future research on this area and we could suggest avenues for achieving better estimates of the severity of the different symptoms related to depression. For example, depending on the availability of social media evidence (defined as *incidence*), systems could focus on certain items (and ignore

[2] Complete questionnaire available at `https://bit.ly/3drfpVg`.

others) in their attempt to understand the psychological aspects of social media users. Furthermore, we are also interested in studying whether the amount of available evidence regarding BDI items varies over social media sources, and whether this variation results from the mental health condition of the individuals or if it is a consequence of the characteristics of the social media source.

Our main research questions, therefore, are:

- **RQ1**: How different is the social media language of users with mental health disorders compared to control individuals?

- **RQ2**: How do social engagement and posting trends of users with mental health disorders differ from control individuals?

- **RQ3**: To what extent are the language and online behavioural traits of various mental disorders different?

- **RQ4**: How can language-specific and emotional information be visualised for mental health practitioners during the diagnosis process?

- **RQ5**: How can users' engagement in relationship with their mental health condition be visualised as it develops?

- **RQ6**: How different are the language and online behavioural traits of users with mental health disorders among different social media platforms? In this respect, how much are the constraints imposed by social media influencing such users?

- **RQ7**: Given a group of users who suffer from depression (positive group), how much information is available on their social media feeds about each BDI item?

- **RQ8**: To what extent do item incidence levels and other possible features correlate with the effectiveness achieved by automated methods that infer the item's response based on social media data?

- **RQ9**: In comparison with positive users, how much information is available on the feed of control users about each BDI item?

To the best of our knowledge, this is the first time that a study of mental health disorders on social media is analysed in depth based on two platforms with highly diverse characteristics. Moreover, it is the first time that the relationship between BDI items and social media is scrutinised. This analysis is a valuable step towards developing better risk-assessment and decision-support tools.

The remainder of this chapter is organised as follows. Section 5.2 details the approach followed to answer the different research questions; Section 5.3.1 presents the results and analyses of the study on language and behaviour. Similarly, Section 5.3.2 presents the results and analyses of the incidence of depression symptoms; Section 5.4 concludes the chapter.

5.2 Objectives and Methods

Here, we describe how we design the experiments to study social media posts to answer the research questions posed in the introduction of the chapter. We also outline what can be learned from each experiment, focusing on the specifics of each research question.

5.2.1 Open-Vocabulary

Vocabulary Uniqueness: One variable we analyse to answer **RQ1** and **RQ6** is the similarity (or diversity) of the unique sets of words that compose the vocabulary of positive and control classes. Analysing such dimension tells us up to which extent classes have a common vocabulary and which words, if any, could be specifically used by users belonging to a certain class.

Considering each vocabulary as a set, we inspect the relative size of their intersection. To this end, we use Jaccard's index [115] to measure the similarity between finite sample sets. Formally, let P be the unique set of words obtained from positive users, e.g. self-harm, and C be the unique set of words obtained from control users. We compute Jaccard's index as follows:

$$J(P,C) = |P \cap C| / |P \cup C| \ .$$

As we see, the index gives us the ratio of the size of the intersection of P and C to the size of their union. The index ranges from 0 to 1, where an index of 1 indicates that the sets completely intersect, and thus, have the same elements. As the value approaches 0 the sets are more diverse among themselves.

Word Usage: An important aspect when studying the language of different groups, in addition to vocabulary similarities and differences, is to understand the patterns of word usage. Here, we attempt to answer **RQ1**, **RQ3**, **RQ4**, **RQ6** by computing and comparing the language models for each class. This analysis aims to quantify the differences that might emerge between the classes in terms of the probability of using certain words more than others.

Language models are processes that capture the regularities of the language across large amounts of data [47]. In its simplest form, known as a unigram language model, it is a probability distribution over the terms in the corpus. In other words, it associates a probability distribution of occurrence with every term in the vocabulary for a given collection. In order to estimate the probability for a word w_i in a document D in a collection of documents S we use:

$$P(w_i|D) = (1 - \alpha_D)P(w_i|D) + \alpha_D P(w_i|S) ,$$

where α is a smoothing coefficient used to control the probability assigned to out-of-vocabulary words. In particular, we use the linear interpolation method (also referred to as Jelinek-Mercer smoothing) where $\alpha_D = \lambda$, i.e., a constant. To estimate the probability for word w_i in the collection we use $s_{w_i}/|S|$, where s_{w_i} is the number of times a word occurs in the collection, and $|S|$ is the total number of words occurrences in the collection. In this chapter, D identifies all the documents in a specific class, i.e., we concatenate all the documents of a particular class, such as self-harm. While S is the union of all the documents of two classes in a corpus, i.e., positive and control.

Once we computed the language models for each class, we plot the probability distributions obtained and analyse to which extent the distributions differ. Furthermore, we support our observations by computing the Kullback-Leibler divergence (KL), a well-known measure from probability theory and information theory used to quantify how much two probability distributions differ. In essence, a KL-divergence of 0 denotes that the two distributions in question are identical. The KL-divergence is never negative and is larger for distributions that are more different. Given the *true* probability distribution P and control distribution C, the KL-divergence is defined as:

$$KL(P||C) = \sum_x P(x) log \frac{P(x)}{C(x)} .$$

5.2.2 Psychometric Attributes and Linguistic Style

A common method for linking language with psychological variables involves counting words belonging to manually-created categories of language [38; 45; 39]. Conversely to the experiment described in section 5.2.1, such method is known as "closed-vocabulary" analysis [213]. In essence, we address **RQ1**, **RQ3**, **RQ4**, and **RQ6** by studying *function words*, and topic-specific vocabulary. On the one hand, the goal of conducting such analysis is to quantify specific stylistic patterns that could differentiate positive instances of a mental disorder from control individuals. For example, individuals suffering from depression exhibit a higher tendency to focus on themselves [8], and thus, it is expected that the use of personal pronouns such as "I" would be higher. On the other hand, certain positive classes might exhibit a higher use of specific topically-related words. As we show later for the case of anorexia when compared to depression and self-harm.

It should be noted that we decide to keep the stop-words since many words such as pronouns, articles and prepositions reveal part of people's emotional state, personality, thinking style and connection with others individuals [39]. As a matter of fact, such words, called function words, account for less than one-tenth of a percent of an individual's vocabulary but constitute almost 60 percent of the words a person uses [39].

Linguistic Inquiry and Word Count [222] (LIWC)[3] provides mental health practitioners with a tool for gathering quantitative data regarding the mental state of patients from their writing style. In essence, LIWC is equipped with a set of dictionaries manually constructed by a psychologist which covers various psychologically meaningful categories and is useful to analyse the linguistic style patterns of an individual's way of writing. In our study, we measure the proportion of documents from each user that scores positively on various LIWC categories (i.e., have at least one word from that category). In particular, we choose a subset of the psychometric categories included in LIWC where we found significant differences between positive and control users. Subsequently, we plot the distributions obtained using box-plots and compare them.

Furthermore, we extend this analysis utilising Empath[4] [60]. This text analysis tool shares a high correlation with gold standard lexicons, such as LIWC, yet it covers a broader and dynamic set of emotional and topical categories. Conversely to hand-tuned dictionary-based tools, Empath's data-driven and human-validated categories are derived from existing knowledge bases and literature on human emotions available on the web by means of unsupervised language modelling. Similarly to the analysis conducted with LIWC, we utilise box-plots to visualise the distributions obtained from a subset of categories where we found significant differences between positive and control users and contrast them.

5.2.3 Emotional Expression

Individuals usually convey emotions, feelings, and attitudes through the words they use. For instance, *gloomy* and *cry* denote sadness, whereas *delightful* and

[3] See http://liwc.wpengine.com/

[4] See http://empath.stanford.edu/

yummy evoke the emotion of joy. Here, we address **RQ1**, **RQ3**, and **RQ4**, **RQ6** by studying how individuals, suffering from mental disorders, emotionally express themselves in their social media posts. Furthermore, we investigate how such emotional expression could differentiate between affected and non-affected users.

We use the emotion lexicons built by Mohammad et al. [151; 150] where each word is associated with the emotions it evokes to capture word-emotion connotations. In addition to common English terms, the lexicons include more prominent words on social media platforms. Moreover, they include some words that might not predominantly convey a certain emotion and still tend to co-occur with words that do. For instance, the words *failure* and *death* describe concepts that are usually accompanied by sadness and, thus, denote some amount of sadness.

5.2.4 Social Engagement and Posting Trends

This dimension intends to profile certain attributes of users' documents, such as the length or the time they were created, as well as individuals' level of participation and interaction on the social media platform. As stated by Coppersmith et al. [46] such attributes provide a proxy, although imperfect, for significant findings in the mental health literature which may manifest and be measured on social media. In particular, we analyse the usage of the following elements to address **RQ2**, and **RQ6** by studying:

- Mentions: A document containing another account's username, preceded by the "@" symbol. For instance, `Hello @earissola!`.

- Hashtags: Unbroken word or phrase preceded by "#". on Twitter, when a hashtag is used in a tweet, it becomes linked to all of the other tweets that include it. For instance, `#WishYouWereHere`.

- All-Caps: Words completely written in capitals. Usually, when uppercase letters are used in emails, text messages, and social media, they are intended to emphasise as if the individual was speaking in an assertive voice or shouting to convey an emotion such as anger or dismay.

- ASCII emoticons: Emoticons written as plain text, such as `>:(`, `:-P`, `<3`.

- Emojis: Ideograms and smileys used in electronic messages and web pages. They are quite similar to emoticons, although emojis are pictures rather than typographic approximations.

- Other annotated elements: It includes words with emphasis (such as `a *great* time`, `I don't *think* I...`), censored words (such as `f**k`, `s**t`) and repeated words.

- Retweets (RT): On Twitter, usually people type "RT" at the beginning of a tweet to indicate that they are re-posting someone else content, i.e., quoting another person's tweet.

- Submission type: On Reddit, users submit content in the form of posts and comments. While posts are used to start an online conversation (called a *thread*), comments are nested responses to other comments or posts.

In our study, we measure the proportion of documents from each user that scores positively on any of these elements (i.e., have at least one occurrence of that element). Subsequently, we plot the distributions obtained using box-plots and compare them.

Finally, we study the time-gap between two consecutive documents for each user in the collection to address **RQ5** and **RQ6**. By analysing the mean and variance of this variable on a monthly basis and aggregating it according to the corresponding group (e.g., PTSD), we analyse and compare the posting regularity of each class. In this way, we can study how users' engagement on the social media platform develops in relationship with their mental health condition. Formally, let $t_1, t_2, \ldots, t_n$ be the timestamps of the elements that comprise the chronology of documents of a user. The average document time-gap for a user is computed as:

$$\frac{1}{n-1}\sum_{t=2}^{n-1} t_{n-1} - t_n,$$

where n represents the number of documents that a user has written. Once we have computed the time-gap for each user, we aggregate the values obtained according to the class and month of the year and plot the distribution.

5.2.5 BDI Item Incidence

In order to answer **RQ7** and **RQ9**, we quantify the *incidence* that each of the BDI items has on social media. In essence, we measure the amount of evidence that can be retrieved about each item from a collection of social media posts. Our assumption is that the strength of such evidence varies across different items and, thus, we can rank and compare the BDI items based on this strength.

Incidence Level: Formally, let s_{ijk} be the *incidence score* computed as the average relevance score of the documents retrieved for BDI item q_i using ranking function

f_j from collection c_k. Based on the incidence scores, we sort the BDI items creating an ordered list of 21 elements (total number of items in the questionnaire), defined as r_{jk}, and where each item occupies position p_{ijk}. Since ranking r_{jk} will comprise of 21 positions, we divide it into four bands. In this way, each band will be indicative of the amount of evidence that could be retrieved for the items in it (as measured by s_{ijk} and relative to the rest of items in the questionnaire). Bearing this in mind, let $I(q_i)$ be a function that assigns an *incidence level* (or category) to each item according to its position p_{ijk} in the ranking r_{jk}, defined as:

$$I(q_i) = \begin{cases} \textit{High } (+), & \text{if } p_{ijk} \geq 1 \wedge p_{ijk} \leq 5 \\ \textit{Middle-High } (\wedge), & \text{if } p_{ijk} \geq 6 \wedge p_{ijk} \leq 10 \\ \textit{Middle-Low } (\vee), & \text{if } p_{ijk} \geq 11 \wedge p_{ijk} \leq 15 \\ \textit{Low } (-), & \text{if } p_{ijk} \geq 16 \wedge p_{ijk} \leq 21 \end{cases}.$$

The rational of dividing the resulting ranking into four bands is that with this approach each incidence level will comprise of the same number of items (except for the last category which has one more item). This categorisation scheme helps to rank BDI items in terms of their relative incidence on social media. Thus, an item whose incidence is high has more available online evidence compared to other items in the psychological questionnaire and, ideally, automatic screening tools would benefit from the higher availability of information. Overall, knowledge about incidence could be leveraged to improve the way in which depression-screening systems are constructed. For example, systems could focus on the items with higher incidence, where more information can be extracted, and ignore items showing sparse evidence, which potentially lead to unreliable estimates. Finally, it is worth mentioning that this is just one way to categorise and compare the incidence of BDI items on social media. For instance, one could directly use the incidence score (i.e., average relevance score) defining a set of thresholds and incidences levels accordingly.

Search methods: For retrieval purposes, we employ Okapi-BM25 (**BM25**) [204] and Query Likelihood with Dirichlet smoothing (**QLD**) [176], two extensively-used ranking functions in the Information Retrieval literature [47]. The document collections used in this study come from two different social media platforms, Reddit and Twitter. As previously stated, we study two social media platforms with very different characteristics and constraints. The reason is that we are interested in determining whether the trends and patterns we observed are held on both platforms, despite the particularities that each one has. The incidence score is computed based on the document [5] relevance scores obtained for each BDI item (post-level analysis). We apply two levels of analysis, according to the number of ranked documents (k) whose scores are averaged from the retrieved lists: *global-incidence* ($k = 1000$) and *top-incidence* ($k = 10$). In this way, we can analyse incidence at different levels.

Guan et al. [82] studied the effect of the relevant document position for navigational and informational search tasks, and found that users mostly prefer top-ranked results almost regardless of their relevance. In fact, Google's first results page attracts 95% of internet users' attention[6]. On most current commercial web search engines, results pages comprise of 10 elements. As top incidence focuses on the existence of a few social media submissions that are highly relevant to the target BDI item, regardless of how well the BDI item matches the entire collection of documents, we decided to pick $k = 10$. Global incidence, instead, takes a more general view of the presence of the BDI item in the document collection. In order to ensure a considerable larger search coverage, without comprising the relevance of the retrieved documents, we decided to use $k = 1000$. A high top incidence score for a given BDI item suggests that a few users intensively discussed the topics and used words related to the BDI item. While a high global incidence score suggests a more sustained prevalence of the BDI item in the available collection of social media submissions (i.e, the item-related evidence spread over a larger set of documents). In particular, we are interested in studying the relative differences in terms of top and global incidence for all the 21 BDI items.

Query Formulation: In order to perform the retrieval experiments, we derived one query from each BDI item. To this end, we considered each item of the questionnaire and constructed a query by concatenating the item title and the content of the available responses[7]. We opted to extract query words from both parts because the title represents the general topic covered by the item (e.g., sadness) while the possible responses express in different ways how people might feel about the topic of the item. Preliminary experiments utilising short queries built from the item title showed that these general expressions (e.g., sadness, agitation or indecisiveness) are hardly used by people to express their feelings. Therefore, we decided to build larger queries using all the words available (title+responses). The resulting 21 queries are shown in Table 5.1. The application of term-matching strategies might overlook some indirect indicators in users'

[5] The term *document* is used henceforth in this chapter to refer to a social media post.

[6] See `http://bit.ly/38OKy4z`

[7] Stopwords were removed in this process. All the experiments were conducted using the NLTK's stopword list, avaialble at `https://gist.github.com/sebleier/554280`

posts for the BDI items. Nonetheless, this simple yet high-precision approach could be considered as a baseline for future experiments and studies as such strategies have proven to be effective in traditional search scenarios. More sophisticated query construction and search strategies which consider other types of relationships between the terms, such as semantic similarity, as well as the limitations of this approach are discussed in the conclusions.

Table 5.1. Queries generated by concatenating the title and content of the available responses for each BDI item.

#	Query Generated
1	sadness time feel unhappy sad stand much
2	get future worse discouraged expect used work things hopeless feel pessimism
3	failure look past total failures person lot back, failed feel see
4	get used ever things loss little pleasure much enjoy
5	feelings particularly time done things many feel quite guilty
6	punishment feelings expect punished may feel
7	lost dislike ever feel self-dislike confidence disappointed
8	usual critical used criticize everything faults happens self-criticalness bad blame
9	killing myself, kill suicidal chance carry wishes thoughts
10	crying, used thing crying every feel anymore cry little
11	wound usual restless agitated something stay agitation hard feel keep still moving
12	get interest lost less anything activities interested things hard loss
13	usual trouble difficult difficulty making used decisions find ever make indecisiveness greater well much
14	useful consider utterly worthwhile used worthless compared worthlessness feel
15	energy used less anything ever loss enough much
16	somewhat experienced pattern usual day hours early 1-2 get less sleeping changes sleep lot wake back change
17	usual irritable time irritability much
18	somewhat experienced usual time less changes food appetite crave change greater much
19	usual difficulty long concentration find mind anything ever hard keep well concentrate
20	usual get easily used fatigue fatigued things tired lot tiredness
21	noticed recent used interest less lost interested sex loss change completely much

5.2.6 Depression Severity Estimation

In order to address **RQ8**, we need some measure to assess the effectiveness of systems that infer BDI responses from social media evidence. To this aim, we adopt two of the official measures proposed by Losada et al. [122; 123]. These measures compare the BDI questionnaire filled by a real user with the BDI questionnaire filled by a system as follows:

- **Average Hit Rate (AHR)**: Hit Rate (HR) averaged over all users. HR is a stringent measure that computes the ratio of cases where the automatically estimated questionnaire has exactly the same answer as the real questionnaire.

- **Average Closeness Rate (ACR)**: Closeness Rate (CR) averaged over all users. CR takes into account that the responses of the BDI questionnaire represent an ordinal scale and computes a distance-based estimation of performance.

For each BDI item we study the relationship between the effectiveness of the automated systems that attempted to respond to the BDI questionnaire and certain BDI features, such as incidence on social media. To this end, we obtained the runs submitted to eRisk 2019 and eRisk 2020 (20 and 18, respectively)[8] and plot the AHR and ACR (averaged over all participating systems) against incidence and other variables. Furthermore, we also compute the Pearson's correlation coefficient between the analysed variables.

5.3 Results and Analysis

This section presents the results obtained from the experiments outlined in Section 5.2 on the eRisk and CLPsych collections described in Chapter 3.2. Moreover, we analyse the corresponding outcomes towards answering the proposed research questions.

5.3.1 Language and Behaviour

Here, we present the results of a thorough study of various dimensions of language and online behavioural traits to characterise users affected by mental disorders (depression, anorexia and self-harm) on two popular social media platforms, namely, Reddit and Twitter. Also, we provide several methods for visualising the data in order to provide useful insights to mental health practitioners. We are interested in investigating whether the trends and patterns we observed on users affected by mental disorders in one social media platform hold in another one. Especially when users are restricted to various types of constraints, such as space limitations, which could influence their writing style and expression [72]. This could reveal whether the language employed by individuals affected by mental disorders is independent of the social media platform they participate in. Based on the restrictions, objectives, and features people tend to behave differently on different platforms. Therefore, it is of high importance to investigate if the features under analysis can generalise to all social media platforms.

[8] We thank the eRisk organisers for providing these submissions and the ground truth.

To this end, we first compared users affected by a particular disorder against control individuals in each platform separately. Then, we studied whether different mental disorders share the same characteristics and if they are clearly different in terms of the dimensions analysed. Finally, we investigated whether the trends and patterns hold on both social media platforms in order to identify and quantify potential differences and better understand which analysis techniques are more suitable for each case. With this study we sough to answer **RQ1** to **RQ6**.

Vocabulary Uniqueness: Table 5.2 compares the vocabulary of the different groups of positive users (i.e., depression, anorexia and self-harm for Reddit and depression and PTSD for Twitter) against that of control users. We analyse their union, intersection and difference. On Reddit, we observe that positive cases of depression and anorexia exhibit a high similarity with their respective control groups, with a Jaccard index of 59% and 65%, respectively. On the other hand, self-harm positive cases use a more diverse set of words in their documents compared to the control group (44%). On Twitter, differences between positive and control groups are considerably larger. A Jaccard index of 26% for depression and 27% for PTSD indicates a more distinctive vocabulary. However, it should be noted that such noticeable differences could in part due to the informal essence that distinguishes microblogging activity [37; 144; 156] along with Twitter's character limit. Therefore, abbreviations, compound words hashtags, mentions, internet slangs and misspellings are common, in many cases deliberate, causing an almost linear growth of the vocabulary size [202]. In fact, the number of unique terms on Twitter is substantially larger than on Reddit, possibly causing a smaller overlap between the sets compared.

Moreover, the vocabulary size of the various positive groups gives us an idea about the words that control users have never used but used by the positive groups. Among the terms that are unique to the positive groups, we find the following ones interesting (on Reddit): self-harm, trazodone[9] (Depression); anorexics, depersonalization[10], emetrol, pepto[11] (Anorexia).

Word Usage: Figure 5.2 compares the language models obtained for the different classes and collections. Note that the smoothing is necessary since, as shown before, there are terms that are present only in the positive class vocabulary but not in the control one and vice-versa. Figures 5.2(a) to 5.2(c) contrast the lan-

[9] *Trazodone* is an antidepressant medication.

[10] Depersonalization is a mental disorder in which individuals feel disconnected or detached from their bodies and thoughts.

[11] *Emetrol*, and *Peptol* are medications used to treat the discomfort of the stomach.

Table 5.2. Vocabularies comparison between positive and respective control users. KL-divergence computed across the language models obtained for the documents of positive and control users. As a reference, the KL-Divergence is also calculated between the different control groups on Reddit.

(a) Reddit (eRisk)

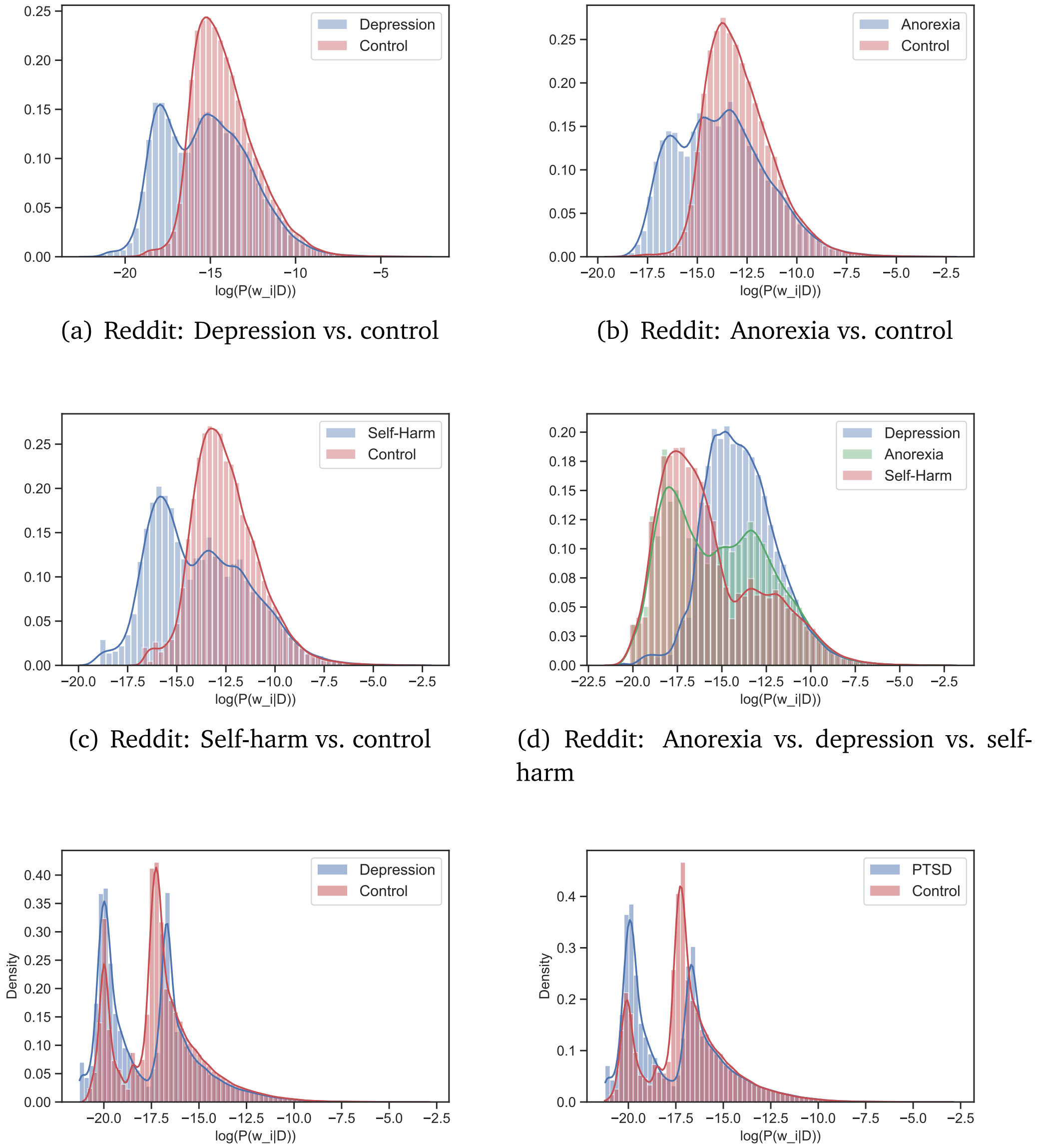
	Depression	Anorexia	Self-Harm
# of Unique Words Positive	41,986	21,448	11,324
# of Unique Words Control	70,229	31,980	25,091
Jaccard's Index (Positive vs. Control)	0.59	0.65	0.44
Difference Size (Positive vs. Control)	218	229	49
Difference Size (Control vs. Positive)	28,461	10,761	13,816
KL(Positive||Control)	0.18	0.18	0.18
KL(Control||Positive)	0.21	0.31	0.20
KL(Control||Control)	0.08	0.07	0.10

(b) Twitter (CLPsych)

	Depression	PTSD
# of Unique Words Positive	150,508	149,393
# of Unique Words Control	238,712	238,712
Jaccard's Index (Positive vs. Control)	0.26	0.27
Difference Size (Positive vs. Control)	70,492	66,945
Difference Size (Control vs. Positive)	158,696	156,264
KL(Positive||Control)	0.08	0.09
KL(Control||Positive)	0.09	0.09

guage model of Reddit positive users against their respective control groups. We note that there are clear differences in terms of language use. However, as Figures 5.2(e) and 5.2(f) reveal, on Twitter such differences are much smoother as there is a notable overlap between the distributions.

These observations are supported by the computation of KL-divergence. Table 5.2 shows the value of KL-divergence computed across the language models obtained for the documents of positive and control users on the different platforms. We note that the KL-divergence confirms the difference between the positive and control language models observed on Reddit in the plots. In fact, as we compare the distribution of different control groups, we observe smaller KL-divergence values (≈ 0), indicating that these distributions are very similar. Moreover, when compared to Reddit, the positive groups on Twitter show a KL-divergence with respect to controls of almost half. Interestingly, the divergence observed between positive and controls groups and vice-versa is practically equal. Recall that KL-divergence is a distribution-wise asymmetric measure that allows



(a) Reddit: Depression vs. control

(b) Reddit: Anorexia vs. control

(c) Reddit: Self-harm vs. control

(d) Reddit: Anorexia vs. depression vs. self-harm

(e) Twitter: Depression vs. control

(f) Twitter: PTSD vs. control

Figure 5.2. Language models probability distribution comparison on Reddit (subfigures a to d) and Twitter (subfigures e and f) (best viewed in colour).

to exactly calculating how much information is lost when we approximate one distribution with another. Therefore, the fact that this difference appears to be symmetric, emphasises the similarity observed between the distributions.

Finally, comparing the language models of the three positive classes on Reddit as in Figure 5.2(d) is harder to identify noticeable differences between the distributions. Depression and anorexia language models are rather similar. While the largest noticeable difference is observed for self-harm when compared with either depression or anorexia. On Twitter, we observe an even larger overlap between the distributions[12]. This reinforces the idea that the word probability distribution between the different positive classes is very similar and how they use words. In this way, a system could compare the language model of a patient with both control and positive groups to provide the mental health practitioners with assistance in determining whether they are positive or not. The practitioner can further examine the patient to determine which disorder they are diagnosed with.

Psychometric Attributes and Linguistic Style: Using a selected set of categories[13] from LIWC and Empath, we demonstrate that language use of Reddit and Twitter users, as measured by LIWC and Empath, is significantly different between positive and control individuals. Figures 5.3, 5.4, 5.5 and 5.6 show the proportion of documents from each user that scores positively on various LIWC and Empath categories (i.e., have at least one word from that category). Selected categories include function words (like pronouns and conjunctions), time orientation (like past focus and present focus), emotionality, drives (like affiliation, achievement and power), as well as cognitive (like insight, tentative and certainty), social (like family and friends) and biological processes (like body and health). Bars are coloured according to the positive and control classes they represent.

The most remarkable case is the difference found in the use of the pronoun "I" between positive and control users on both social media platforms, which in the case of depression replicates previous findings [45; 38]. Moreover, the proportion of messages using words related to positive emotions (denoted *posemo*) is larger than negative (denoted *negemo*) ones, even for the positive classes, as measured by LIWC on Reddit and Twitter. Such circumstance could be related to the fact that English words, as they appear in natural language, are biased towards positivity [106]. Except for categories *we* and *she/he*, differences reach statistical significance using Welch two-sample t-test (p-value < 0.001) from each corresponding control group in Figures 5.3(a) and 5.4(a). No significant differences between depression, anorexia and self-harm were found for any of the LIWC and Empath categories analysed in Figure 5.3(a) and 5.4(a). Similarly, in most of the

[12] Given that the figure is not providing any observable trend and to avoid needlessly overloading the document, we decided not to include it.

[13] For a comprehensive list of LIWC categories see `http://hdl.handle.net/2152/31333`

categories analysed in Figures 5.5 and 5.6 (except for *optimism* and *friends*) statistically significant differences are observed between at least one positive group and its respective control. Moreover, we also see significant differences between individuals suffering from depression and PTSD (such as cognitive and social processes, violence, death and swearing, just to mention a few).

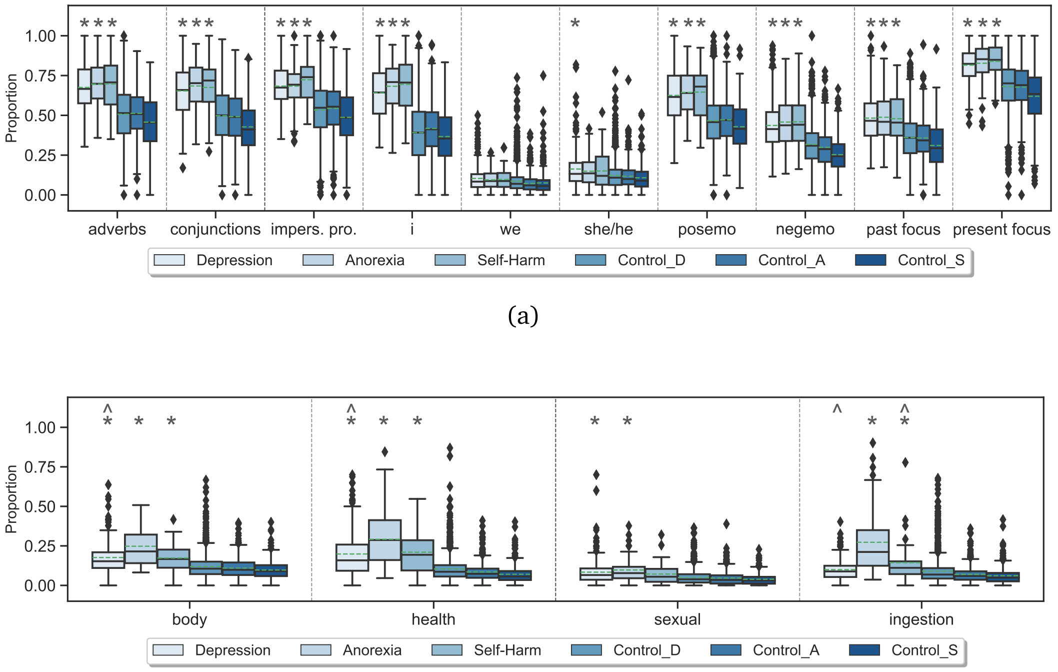
Figure 5.3(b) depicts categories related to *biological processes*. In essence, this category includes words directly associated with the body and its main functions. We note that individuals within the anorexia group show certainly different behaviour compared with depression and self-harm groups. Intuitively, this result is expected, given that anorexia is characterised by an intense fear of gaining weight and a distorted perception of weight. Moreover, LIWC observations are complemented by Empath's categories shown in Figure 5.4(b). Overall, people with anorexia place a high value on controlling their weight and shape, using extreme efforts that tend to significantly interfere with their lives. Therefore, it is reasonable that such individuals talk more about themes related to their body and its function. Statistical significance (p-value < 0.001) between individuals suffering from anorexia and those affected by depression is achieved for categories *body*, *health* and *ingest*. While only for the latter category the differences are statistically significant when comparing self-harm and anorexia users.

In addition to the default categories that LIWC includes, we study other domain-specific lexicons. The first of them is the well-known depression lexicon[14] released by Choudhury et al. [38]. It consists of words closely associated with texts written by individuals discussing depression or its symptoms in online settings. The use of such words is not significantly different with respect to anorexia and self-harm groups for Reddit as well as PTSD for Twitter. This suggests that those words are also frequently used by users affected by anorexia, self-harm or PTSD. The same situation was observed when considering the set of absolutist terms[15] derived from the work of Al-Mosaiwi et al. [4], who concluded that the elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation.

Overall, we observe less quantifiable differences between positive and control groups on Twitter than on Reddit. It appears that the use of language on Twitter is less distinguishable between users potentially suffering a mental disorder. This could be related to the limited length or the depth of topics that are discussed on Twitter, as opposed to Reddit. As literature has shown brevity has an impact on the writing style exhibiting distinctive linguistic features; for instance, length

[14] Examples of words included in the lexicon: *insomnia*, *grief*, *suicidal*, *delusions*.

[15] Examples of absolutist terms: *absolutely*, *constantly*, *definitely*, *never*.



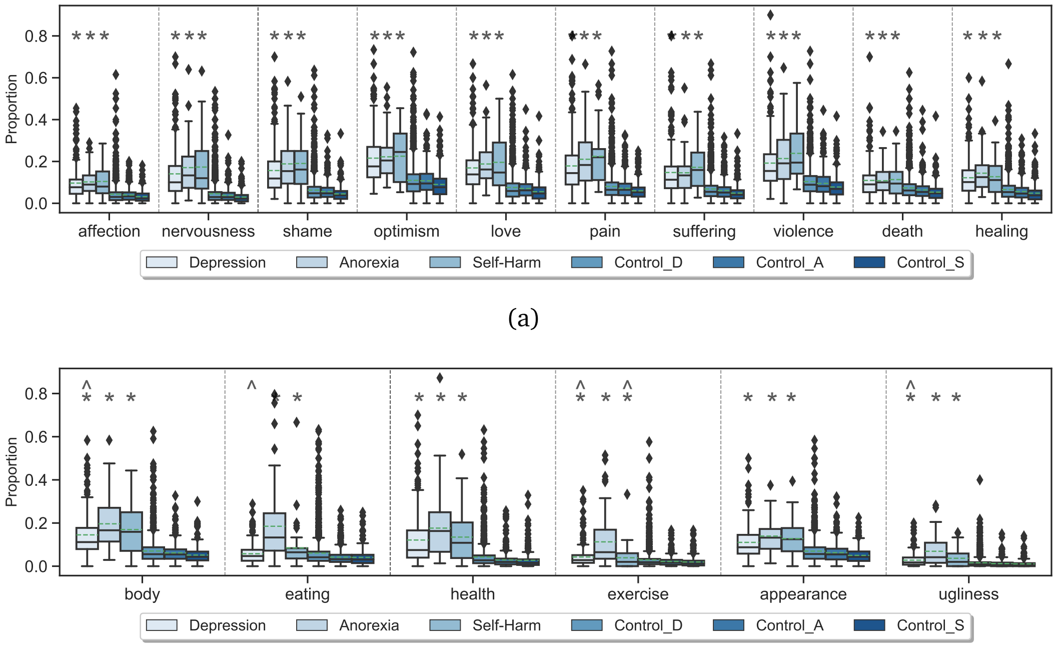
(a)

(b)

Figure 5.3. Box and whiskers plot of the proportion of documents each user has on Reddit (y-axis) matching various LIWC. Statistically significant differences between each positive and their respective control groups are denoted by * (p-value $<$ 0.001). Also, statistically significant differences between Depression/Self-Harm and Anorexia are denoted by ˆ (p-value $<$ 0.001).

constraints disproportionately preserve negative emotions and articles, adverbs, and conjunctions have the highest probability of being omitted [73]. However, we see that among the two disorders in the data (i.e., depression and PTSD), individuals diagnosed with PTSD are potentially easier to identify by their language use. Also, in general we observe on Twitter similar trends to Reddit, where individuals with depression use a language that resembles that of control users. Even though both PTSD and depression groups are exhibiting similar language use on Twitter, we see that the depression group is even more similar to the control group.

As it was shown, statistically significant differences are observed between positive and control groups (on both social media platforms) as quantified by several psychometric and linguistic style attributes. Utilising two different although complementary dictionary-based tools, we have shown that this close-vocabulary



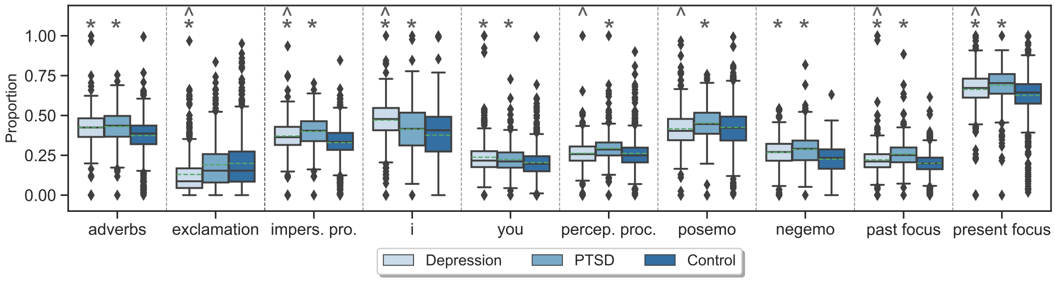
(a)

(b)

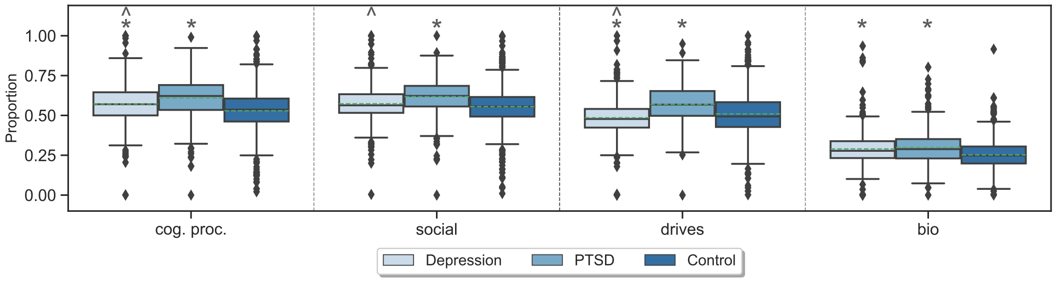
Figure 5.4. Box and whiskers plot of the proportion of documents each user has on Reddit (y-axis) matching various Empath. Statistically significant differences between each positive and their respective control groups are denoted by * (p-value < 0.001). Also, statistically significant differences between Depression/Self-Harm and Anorexia are denoted by ˆ (p-value < 0.001).

analysis provides a reliable way to collect quantitative data from user-generated content about the mental state of individuals on social media.

Emotional Expression: Figure 5.7 depicts for each class and corresponding collection the average number of documents that contain at least one word associated with a particular emotion, including the polarity (positive or negative). In Reddit's case, we note considerable differences in the expression of emotions between positive and their respective controls. One way to interpret such results is that on average affected users tend to share emotions more regularly than control individuals. Conversely, depressed and control individuals exhibit an overlapping trend on Twitter, being users affected by PTSD the ones likely to share on average slightly more emotions. Comparing the social media platforms, we observe that terms related to `sadness` are less frequent on Twitter. Furthermore, words that relate to the positive sentiment are the most frequent ones among the various emotions on both social media. These observations on emotional expres-



(a)



(b)

Figure 5.5. Box and whiskers plot of the proportion of documents each user has (y-axis) on Twitter matching various LIWC categories. Statistically significant differences between each positive and the control group are denoted by * (p-value < 0.001). Also, statistically significant differences between Depression and PTSD are denoted by ˆ (p-value < 0.001).

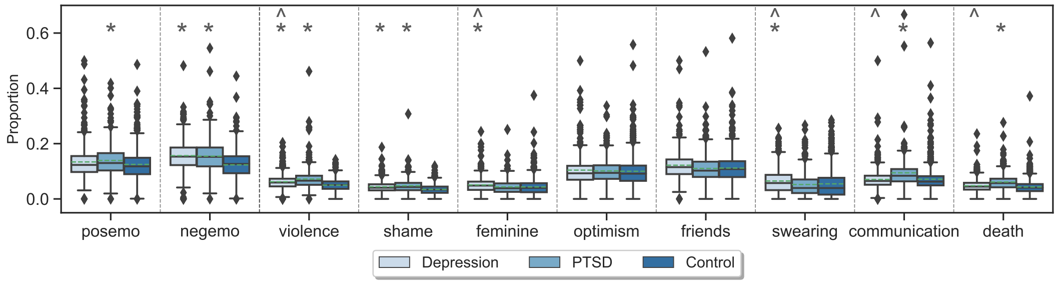
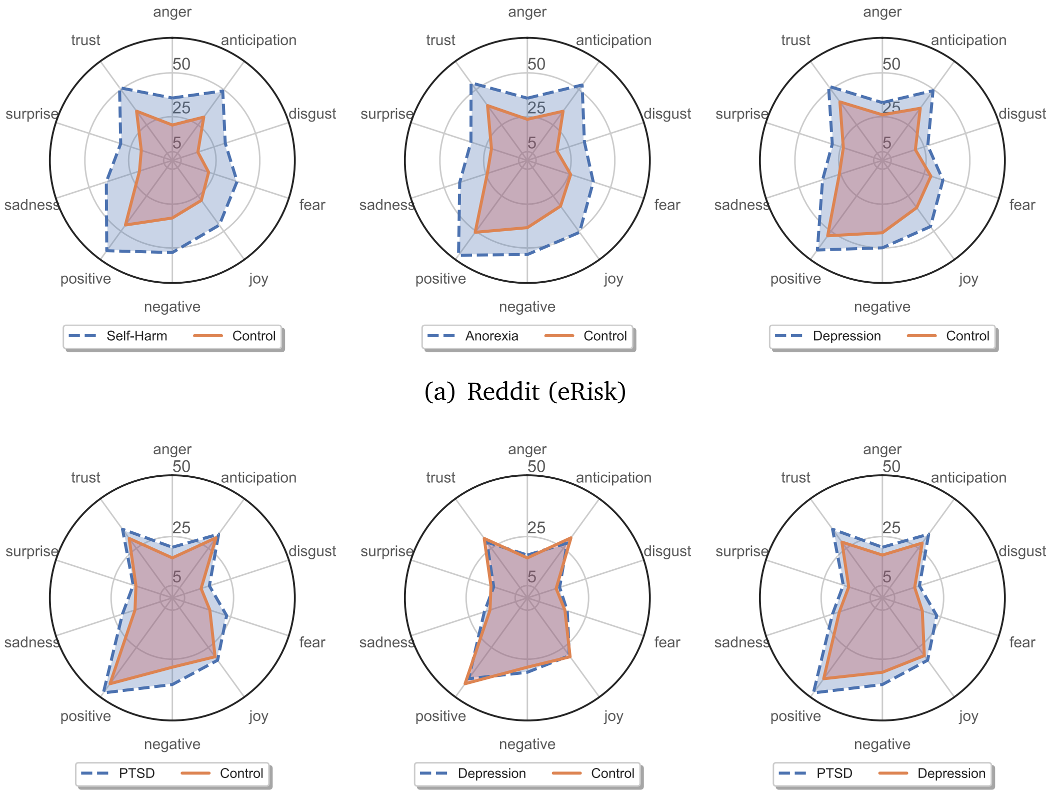


Figure 5.6. Box and whiskers plot of the proportion of documents each user has (y-axis) on Twitter matching various Empath categories. Statistically significant differences between each positive and the control group are denoted by * (p-value < 0.001). Also, statistically significant differences between Depression and PTSD are denoted by ˆ (p-value < 0.001).

sion support our earlier findings indicating that there are considerable smaller differences in the use of language between positive and control groups on Twitter when compared to Reddit. Moreover, as it was also previously noted, we see that on Twitter users suffering from PTSD reveal a more distinguishable use of language with respect to controls than depressed individuals.



(a) Reddit (eRisk)

(b) Twitter (CLPsych)

Figure 5.7. Radar plot representing the average number of documents that contain at least one word associated with a particular emotion, including the polarity (positive or negative) for each positive group and its respective control.

In addition, we also analyse the frequency correlation of the different emotions for each class and collection in Figures 5.8 and 5.9. We note that certain emotions show different correlations depending on the class under observation. For instance, in self-harm's corresponding control class (Control_S in Figure 5.8(c)), `surprise` reveals a larger positive correlation with `trust`, `joy` and `positive` and `negative` orientation when compared with self-harm class. Conversely, with `surprise` and `disgust` as well as with `fear` and `disgust`. Interesting to note is that in the case of depression (Figure 5.8(a)), `sadness` exhibits a negative correlation with `joy` and `positive polarity`. Such correlation does not hold for the respective controls. This kind of study allows us to better understand how different emotional patterns emerge from the use of emotions.

In order to compare both social media platforms we take the particular case of depression. We see in both platforms that `anger` correlates positively with `disgust`, `fear`, `negative polarity` and `sadness`. Interestingly, we observe that the positive correlation between `anger` and `sadness` is nearly double on Twitter compared to Reddit. In addition, we observe that `anticipation` correlates positively with `joy`, `positive polarity`, `surprise`, and `trust` in both social media platforms. We also find notable correlations between `disgust` and `fear`, `negative polarity`, and `sadness` (larger positive correlation on Twitter). Among other interesting emotions we can point out the positive correlation that `fear` has with `negative polarity`, and `sadness`. We also observe more intuitive correlations such as the positive correlation between `positive polarity` and `joy`, `surprise`, and `trust`. More on the negative sentiment, we see that `sadness` positively correlates with `negative polarity` (stronger on Twitter). Overall, we note that emotion correlations are quite similar on both social media platforms for the various mental disorders. As we observed in other experiments, Reddit users affected by a mental disorder show a more distinctive behaviour from control ones than what we observe on Twitter. However, observing similar correlations implies that even though the bag-of-words analysis of the emotions lead to subtle differences, more sophisticated analyses can reveal more distinctive differences.

Social Engagement and Posting Trends: Tables 3.1 and 3.2 show various statistics of the different collections studied in this chapter. We note that among all the eRisk's collections, an average positive user generates fewer documents than an average control user. However, as we see the average length of the documents is longer for the positive cases. In particular, it is interesting to highlight the case of users who suffer from anorexia. They even write longer documents than users affected by depression and self-harm. It is worth noting that we spotted no meaningful differences among the various control groups. This observation is repeated for various experiments and is expected since each control group is a random sample of Reddit posts at different temporal periods. Therefore, each control group is a representative sample of users on Reddit.

In the case of CLPsych collections, we observe very subtle differences between positive and control individuals. The average number of documents generated by each group is nearly the same, being marginally larger for the positive ones.

Individuals affected by PTSD write on average slightly longer documents than depressed and control users. Finally, we observe a noticeable difference in the average number of documents per individual between the platforms. Regardless of whether they are potentially affected by mental disorders, Twitter users consistently generate more documents than Reddit users. Furthermore, the difference observed in terms of average document length is clearly a result of Twitter's character limit, which does not exist on Reddit.

Figure 5.10 depicts the proportion of documents that each user has considering various social engagement indicators (i.e., there is at least one occurrence of that element in a document). We considered a variety of elements modelling various aspects of users' interaction and engagement with the social media platform, such as hashtags, mentions, and ASCII emoticons. Bars are coloured according to the positive and control classes they represent.

We observe a noteworthy difference between the members of each social media platform. While Reddit users hardly utilise most of the elements studied, their usage is quite common among Twitter members, in particular mentions and hashtags. Interestingly, the majority of the submissions from individuals suffering from depression on Reddit are comments (instead of posts). This suggests that, with respect to control users, most of the time they tend to reply to existing submissions (either another comment or a post) instead of initiating the conversation thread themselves. Hence, in comparison with users who are estimated to be healthy, one could assert that they exhibit a more *passive* behaviour in terms of content creation. On Twitter, hashtags are significantly less often by depressed users. A closer inspection reveals that, on average, users affected by depression exhibit a significantly smaller hashtag ratio (≈ 1/%), meaning that out of ten words one is a hashtag. While for PTSD and control users out of ten words two are hashtags (≈ 2%). Moreover, PTSD users show a significantly larger usage of mentions than depressed and control individuals. This could be interpreted as a way of looking for peer support to overcome their condition. For instance, sharing narratives about the event that triggered their trauma to friends, pals support, or other individuals suffering from PTSD as well. However, there are not significant differences in the ratio of mentions per document among the three groups, being slightly larger for PTSD users (≈ 7%). Therefore, despite that on average PTSD users tend to include mentions more regularly than depressed and control individuals, the frequency of mentions per document is practically equal for the three groups.

An appealing feature of the collections studied in this thesis is that the average lapse time between users' first post (i.e., oldest interaction) and users' last post (i.e., newest interaction) cover more than a year. In this way, we can study how users' engagement in the social media platform develops in relationship with their mental health condition. Figure 5.11 plots the time-gap behaviour of users belonging to different groups in both social media platforms. We plot the average posting time-gap per month for both platforms. In Figure 5.11(a), there is a clear difference in terms of posting regularity of users in different groups on Reddit. In particular, users diagnosed with self-harm exhibit the highest variance on a monthly basis. We see that while the time-gap in some months (e.g., April, August, and November) is very high, the users post more frequently in other months. Moreover, we observe from the highest standard deviation in the behaviour of users that individuals who suffer from self-harm exhibit the least posting regularity. This could be due to the so-called *self-harm cycle* [16]. On Reddit data, we also see that users who suffer from Anorexia exhibit a larger time-gap between their posts. Moreover, the monthly posting variance is higher than the control groups. In general, we observe that the control groups show a relatively stable posting behaviour. We cannot see any noticeable differences between the three control groups on Reddit. Similarly to other measures, we see less distinction between the depression group and the control groups on Reddit. However, as we see in Figure 5.11(b), users on Twitter who suffer from depression, seem to post more frequently. We observe that the control group posting trend is more similar to the PTSD group. Interestingly, towards the end of the year (i.e., November and December), the control group users behave more like the depression group users. Finally, we observe in the plot that users belonging to the PTSD group exhibit a higher variance (i.e., the higher standard deviation in the graph), indicating less stability compared to the control group, even though the average time-gap is nearly the same.

5.3.2 Incidence of Depression Symptoms

Here, we present the results of a thorough study of the 21 BDI items and their incidence on social media. This represents an attempt towards understanding to what extent the answers of this psychological test can be inferred using automated means. To this end, we first categorised the incidence of BDI items by taking into account the amount of evidence that can be retrieved from social media and following the categorisation scheme previously introduced (Section 5.2.6). Secondly, we analysed the incidence together with other features in an attempt to understand what makes a BDI item easier to estimate. We start by

[16] Inflicting some sort of pain can bring a momentary sense of calm to the negative feelings experienced by the individual. However, this is often temporary and can lead to feelings of guilt and shame which can stimulate again the original emotions leading to further self-harm.

analysing the BDI items incidence obtained for each group of users individually (positive and control), and subsequently we contrast this variable between the two groups. Afterwards, we analyse the depression severity estimation. It is important to recall that retrieval units are individual posts. With this study we sough to answer **RQ7** to **RQ9**.

It should be noted that for this study we used both eRisk and CLPsych depression collections as the BDI is designed to measure the presence of depression symptoms. Moreover, eRisk depression control group contains also users who were active on Reddit's depression threads but had no depression (e.g., a doctor giving support to others or a husband whose wife is depressed). To address **RQ9** avoiding a possible bias, we decided to replace eRisk depression control group with the control group of eRisk self-harm, which is a purely random set of users.

Positive Groups: Table 5.3 compares the top and global incidence for the two groups of positive users (i.e. Reddit and Twitter). It should be noted that the retrieval units considered for this experiment are individual posts. This means that each post in a particular collection is considered as a single document. Thus, the more individual posts that are on-topic, the higher the incidence of the BDI item. Overall, we observe that there are noticeable differences in the incidence that various BDI items exhibit on social media. At global level, certain items such as 13 (*indecisiveness*) and 16 (*sleeping patterns*) show a consistent behaviour across different ranking functions and collections. This suggests that these topics are more frequently discussed and more evidence about these items can be extracted and analysed. Conversely, items 1 (*sadness*), 6 (*punishment feelings*) and 17 (*irritability*) exhibit the opposite behaviour, showing fewer available pieces of evidence. Other items such as 7 (*self-dislike*) are more prominent in Twitter than in Reddit. Interestingly, this item and item 8 (*self-criticalness*) show an incidence in Reddit than is higher at the top level (when compared with their incidence at global level). This suggests that, in Reddit, these two topics were not globally prevalent but a few users posted highly relevant contents about these items. Similarly, item 11 (*agitation*) has moderate global incidence in Twitter while its top incidence in this platform is much higher.

A cautionary note should be added here. The fact that an item shows a low incidence level does not imply that there is no evidence available on social media for that item. Instead, it means that the existing evidence, if any, is not easily retrievable by standard search methods. This difficulty to retrieve relevant contents sets a barrier for those systems that try to learn about the symptom represented by the BDI item. For items with higher incidence, instead, obtaining documental evidence does not represent a complex endeavour.

Table 5.3. BDI item's incidence level comparison of positive groups at global and top scales. *R* stands for Reddit and *T* for Twitter. High (+), Middle-High (∧), Middle-Low (∨) and Low (−) incidences categories.

		Global ($k = 1000$)				**Top** ($k = 10$)			
#	BDI Item	BM25$_R$	QLD$_R$	BM25$_T$	QLD$_T$	BM25$_R$	QLD$_R$	BM25$_T$	QLD$_T$
1	Sadness	−	−	−	−	−	−	−	−
2	Pessimism	+	∨	+	−	+	∧	∨	∧
3	Past Failure	∧	∨	∧	∧	∧	∨	∧	∧
4	Loss of Pleasure	∧	−	∧	∨	∨	−	∨	−
5	Guilty Feelings	∧	∨	∧	∨	∧	∨	∧	−
6	Punishment Feelings	−	−	−	−	−	−	−	∨
7	Self-Dislike	∨	∧	+	+	+	+	+	+
8	Self-Criticalness	−	∧	−	∧	∨	+	∨	+
9	Suicidal Thoughts	−	∧	−	∧	−	∧	∨	∨
10	Crying	+	+	+	−	+	+	+	−
11	Agitation	∨	−	∧	∨	∨	∨	+	+
12	Loss of Interest	∧	∧	∨	∧	∨	−	∧	∧
13	Indecisiveness	+	+	+	+	∧	∧	∧	∨
14	Worthlessness	−	∧	−	+	−	∨	−	−
15	Loss of Energy	∨	−	∨	−	−	−	−	−
16	Sleeping Patterns	+	+	+	+	+	+	+	∧
17	Irritability	−	−	−	−	−	−	−	+
18	Changes Appetite	∨	+	∨	+	∨	∧	∨	∨
19	Concentration	+	∨	∧	∨	+	∨	+	∧
20	Tiredness/Fatigue	∨	∨	∨	∨	∧	+	−	+
21	Loss Interest Sex	∧	+	∨	∧	∧	∧	∧	∨

Control Groups: Table 5.4 compares the incidence levels at global and top scales for the two groups of control users (i.e., Reddit and Twitter). As happened with the positive groups, items 13 (*indecisiveness*), 16 (*sleeping patterns*) exhibit high global incidence over different retrieval models and collections likewise items 15 (*loss of energy*) and 17 (*irritability*) show a consistent low incidence (both global and local). Items 2 (*pessimism*), 19 and 21 (*loss interest sex*) show an overall lower incidence than that achieved with the users in the positive groups. Interestingly, item 10 (*crying*) which could be thought as a theme highly discussed by positive users evinces a similar behaviour between both groups. All users consistently express concerns related to this item.

Positive vs. Control Groups: Figure 5.12 presents a boxplot of the incidence scores of the BDI items (for the two collections and user groups). The goal of this analysis is to study the variation of incidence scores over collections and groups. To this aim, we selected three groups of BDI items (those with consistently high, moderate and low incidence, respectively). In addition, as suggested

Table 5.4. BDI item's incidence level comparison of control groups at global and top scales. R stands for Reddit and T for Twitter. High (+), Middle-High (∧), Middle-Low (∨) and Low (−) incidences categories.

		Global ($k = 1000$)				**Top** ($k = 10$)			
#	BDI-Item	BM25_R	QLD_R	BM25_T	QLD_T	BM25_R	QLD_R	BM25_T	QLD_T
1	Sadness	∨	−	−	−	−	∨	−	−
2	Pessimism	∧	−	∧	−	−	∧	∧	+
3	Past Failure	+	∧	∨	∧	∧	∧	∧	+
4	Loss of Pleasure	∧	∨	∧	∨	∨	−	∧	−
5	Guilty Feelings	∧	∨	∧	∨	∨	∨	∨	−
6	Punishment Feelings	−	−	−	−	−	∨	−	∨
7	Self-Dislike	−	∧	+	+	+	+	+	+
8	Self-Criticalness	−	∧	∨	+	∧	∧	∨	+
9	Suicidal Thoughts	−	∧	−	∨	−	∨	−	∧
10	Crying	+	+	+	∧	+	+	+	∨
11	Agitation	∨	∨	∧	∧	∨	∧	+	∧
12	Loss of Interest	∧	∧	∨	∨	∨	−	∧	−
13	Indecisiveness	+	+	+	∧	∧	−	+	∧
14	Worthlessness	−	∨	−	+	−	∨	−	∨
15	Loss of Energy	∨	−	−	−	∨	−	−	−
16	Sleeping Patterns	+	+	+	+	+	+	+	∧
17	Irritability	−	−	−	−	−	−	−	+
18	Changes Appetite	∨	+	∨	+	∧	+	∨	∧
19	Concentration	+	∨	+	∧	+	∧	∧	∨
20	Tiredness/Fatigue	∨	∨	∧	∨	∧	+	∨	∨
21	Loss Interest Sex	∧	+	∨	∨	+	−	∨	−

by Fuhr [63], we computed the effect size using Cohen's D [225] along with the corresponding statistical significance (using Welch two-sample t-test with p-value < 0.001) to quantify the differences observed. Table 5.5 shows the effect sizes computed for the three groups of BDI items.

First, comparing the positive groups of Twitter and Reddit, we observe that, on average, the documents retrieved from Twitter have a higher relevance score than those from Reddit. In fact, by looking at Table 5.5 (columns labelled as **RP vs. TP** and **TC vs. RC**) the differences observed can be categorised as having a "very large" effect size ($\Delta > 1.2$) [225]. This suggests that more evidence in the form of tweets (compared with Reddit posts) is available about the topics of the BDI items. Intuitively, one explanation could be found in the characteristics and constraints (e.g., space limitations) of the social media platforms under study. Twitter users are constrained to a limited number of characters, while Reddit users write much longer texts on average (see Tables 3.2 and 3.1). Tweets are much more focused messages and, if they are on-topic for a given BDI item, their relevance score is not severely penalised in terms of length normalisation.

Another reason could be the size of the collections under analysis. The Twitter collection is almost twelve times bigger than the Reddit collection and, thus, the chances of retrieving relevant content increases.

Second, we notice an interesting trend between the incidence scores of the positive and control groups. When comparing eRisk positive and controls groups (**RP vs. RC** in Table 5.5) we notice very small effect sizes and differences are not significant in all the cases. For Twitter groups (**TP vs. TC** in Table 5.5), effect sizes are larger. Interestingly, Twitter control group presents, on average, a slightly larger incidence score when compared to the positive group. Although, this trend is not apparent for Reddit. One interpretation of this result is that the themes covered by the BDI items are highly discussed irrespective from the mental health state of the users. However, as shown earlier in this chapter most of the differences between positive and control groups are quantifiable considering the way individuals express and use language and emotional indicators and not to a greater extent to the themes they discussed about.

Table 5.5. Effect sizes computed for the incidence score of BDI items in Figure 5.13. Statistically significant differences between the compared groups are denoted by * (p-value < 0.001). RP stands for Reddit Positive, RC stands for Reddit Control, RP stands for Reddit Positive and TC stands for Twitter Control.

#	RP vs. RC	TP vs. TC	TP vs. RP	TC vs. RC
10	0.1511*	0.8984 *	2.5979*	2.9270*
13	0.0507	0.7558*	1.4040*	2.1710*
16	0.1767*	0.5193*	1.3933*	2.1344*
3	0.0200	0.5474*	1.2608*	1.8440*
20	0.1708*	0.6113*	2.0717*	2.5042*
12	0.0340	0.5279*	1.3584*	2.0211*
1	0.1478*	0.5347*	1.7741*	2.1913*
6	0.0673	0.7303*	1.9533*	2.6199*
17	0.2294*	0.6469*	2.5242*	2.9391*

Depression Severity Estimation: Figure 5.13 depicts the incidence score of the BDI items against the mean ACR and AHR achieved by the participating systems in eRisk 2020[17]. The Pearson's correlation coefficients, −0.5046 for ACR and

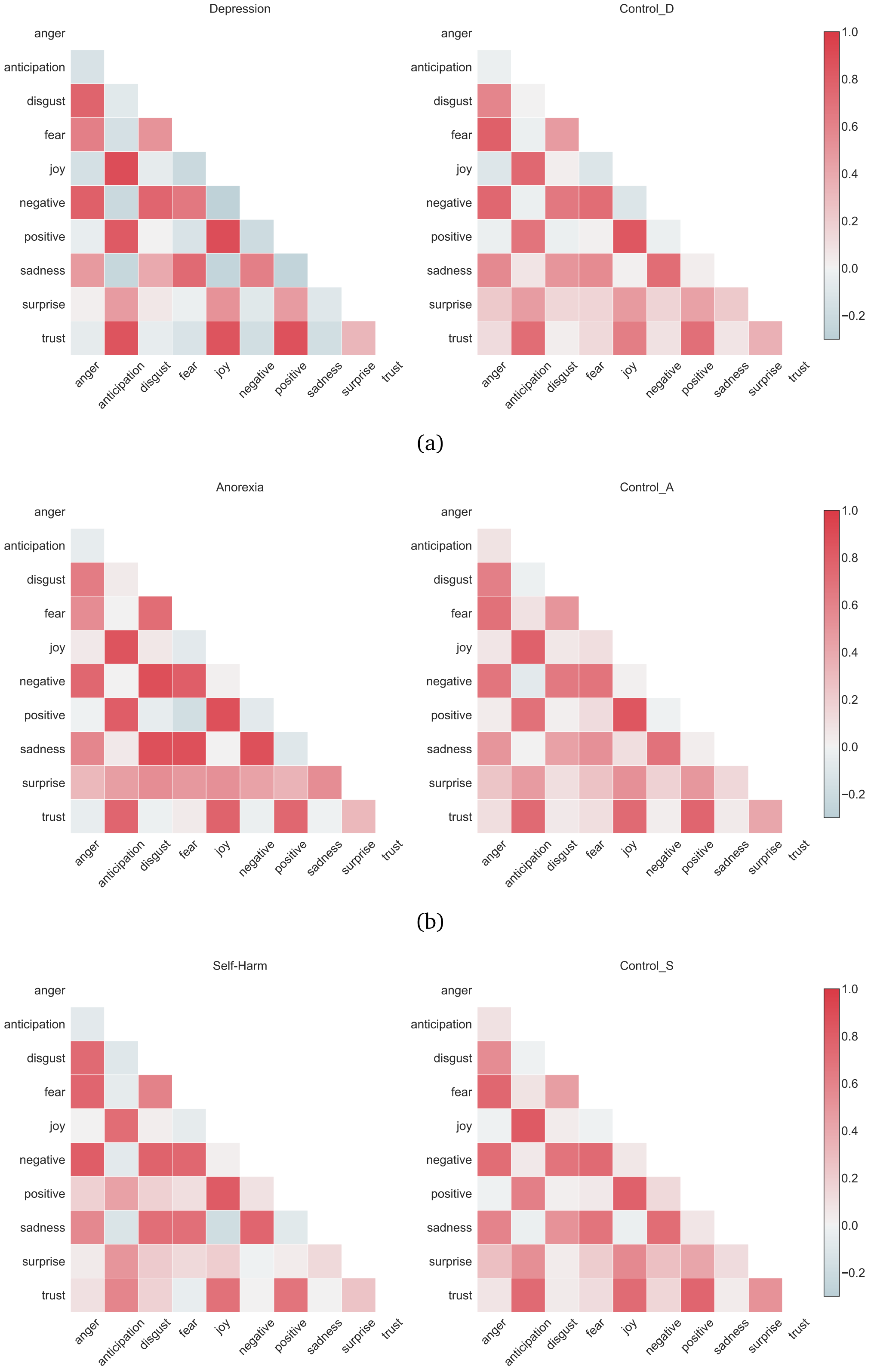
[17] Given that similar conclusions were obtained for eRisk 2019 and to avoid needlessly overloading the document, we only present ACR and AHR plots for eRisk 2020.

-0.5092 for AHR, show a negative correlation (adopting a Bonferroni correction both are significant at p-value $< 0.025\,(0.05/2)$). This suggests that the more evidence available, the less effective the systems are. In fact, items such as *punishment feelings* and *sadness* (items 6 and 1 in Table 5.3) show the highest ACRs but low incidence scores. Conversely, high incidence items like *sleeping patterns* (item 16) have low ACR. More documents that are estimated to be relevant for a given BDI item also means more noise, and our analysis suggests that current systems fail to filter out noisy information.

We also performed a query-length analysis. This revealed that items which yielded the highest effectiveness tend to be those whose representation has fewer words (Pearson's correlation coefficients of -0.5228 for AHR and -0.4370 for ACR, both statistically significant). The longer the query derived from the BDI item is, the more off-topic terms are introduced, causing an issue known as query drift. Query words are obtained from both the item's title and the item's responses. These textual elements are the only pieces of information available to the systems for representing the BDI item. Most participants produced queries or representations from the questionnaire without any further post-processing. Our correlation study between query length and effectiveness suggests that *long* BDI items are potentially noisy.

We also analysed the average inverse document frequency (IDF) of each query[18]. In essence, IDF is a measure of how much information a word provides. The goal of this experiment was to determine whether the discriminative power of the terms obtained from the BDI items had some effect on the effectiveness of the systems. However, we did not find any clear association between average IDF and performance.

[18] We employed MS MARCO Passage Retrieval collection to obtain the term statistics needed to compute IDF. Available at `https://microsoft.github.io/msmarco/`.

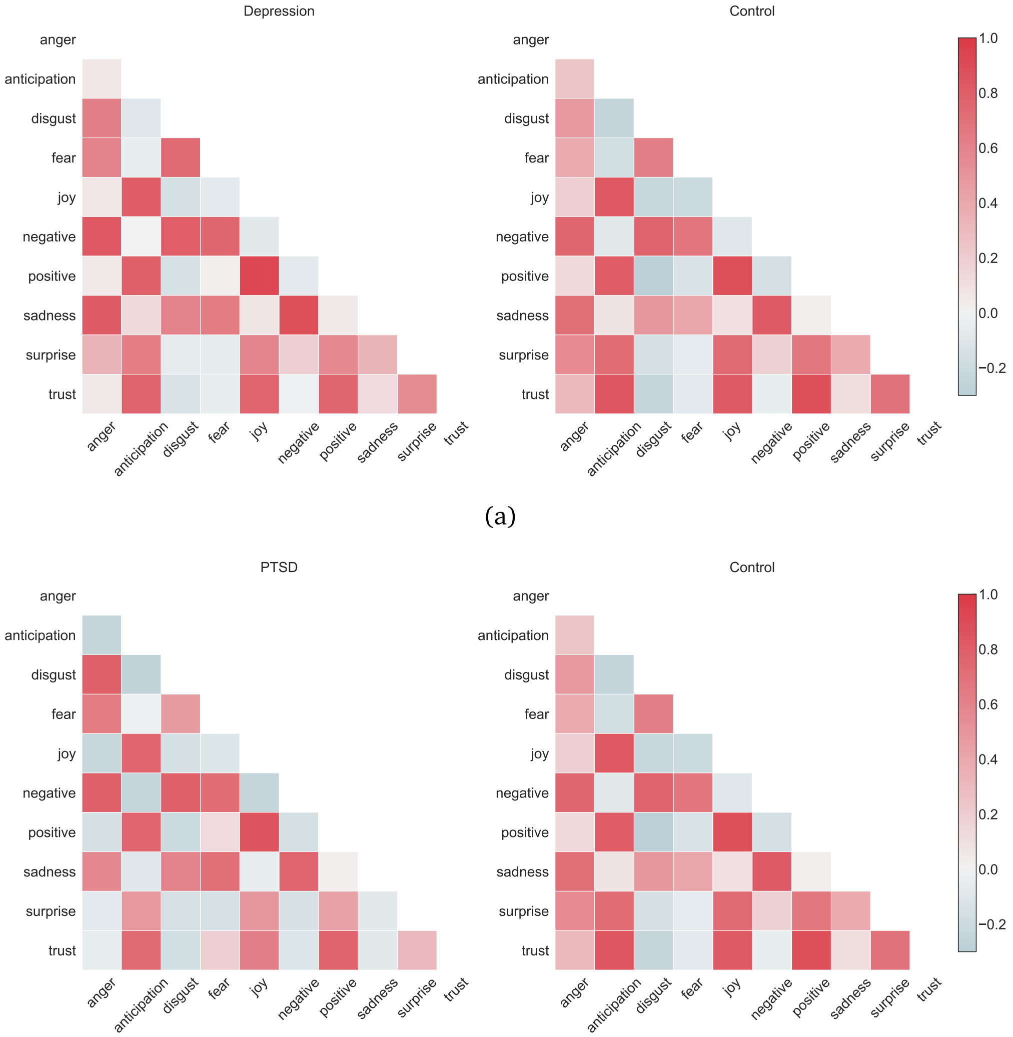


(a)

(b)

(c)

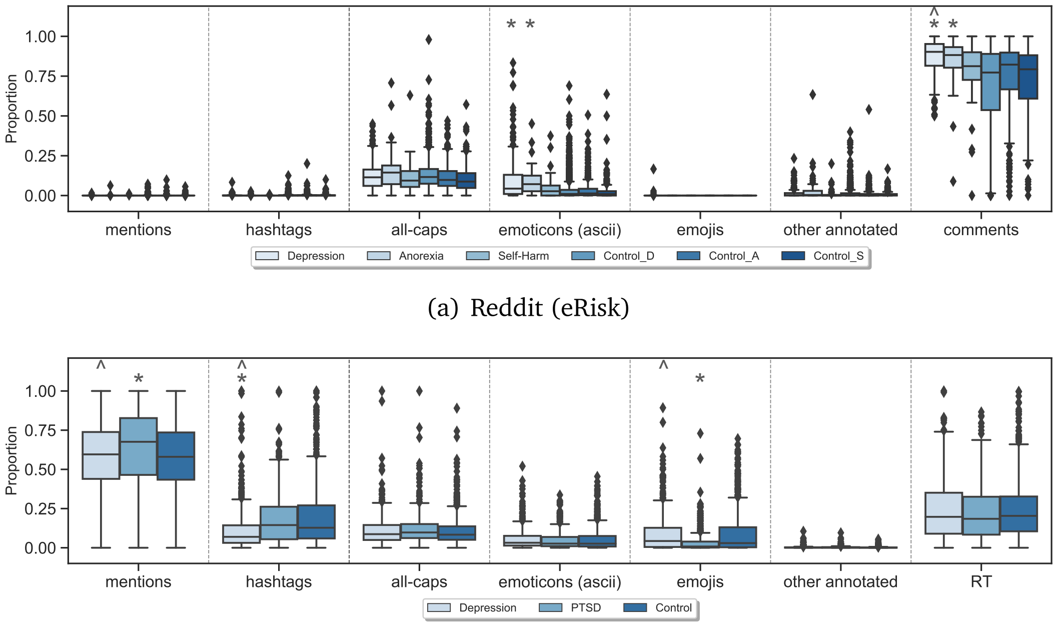
Figure 5.8. Heatmap depicting the frequency correlation of the different emotions for positive (left) and control (right) groups on Reddit (best viewed in colour).



(a)

(b)

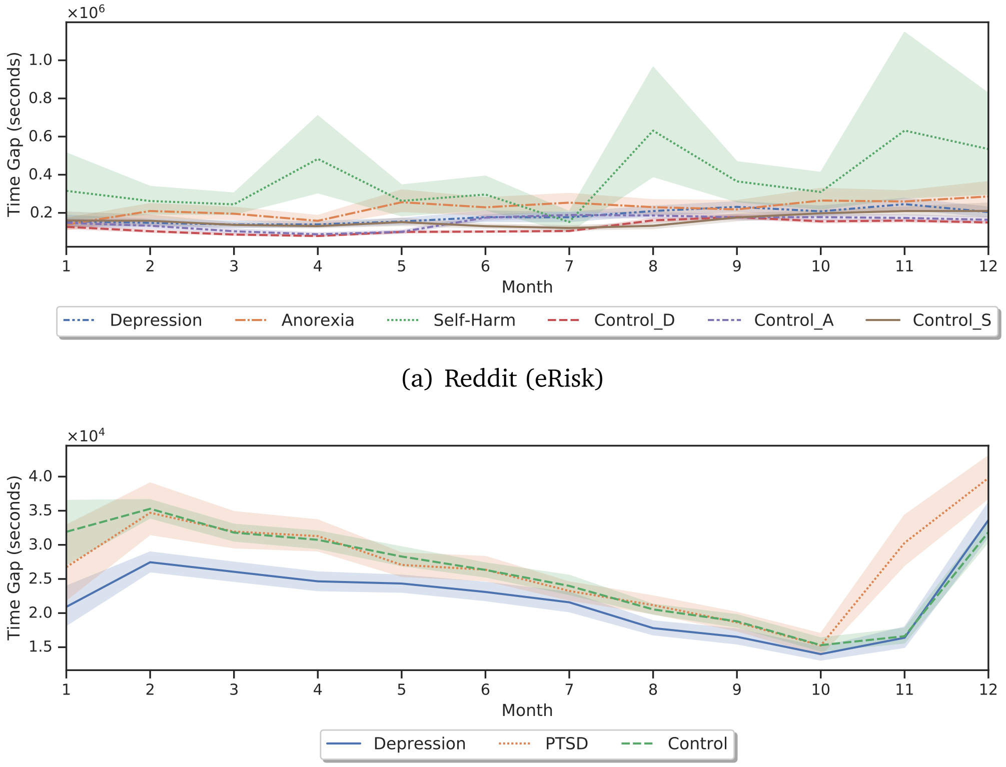
Figure 5.9. Heatmap depicting the frequency correlation of the different emotions for positive (left) and control (right) groups on Twitter (best viewed in colour).



(a) Reddit (eRisk)

(b) Twitter (CLPsych)

Figure 5.10. Box and whiskers plot of the proportion of documents each user has (y-axis) matching various social engagement and posting trends indicators. Statistically significant differences between each positive and their respective control groups are denoted by * (p-value < 0.001). Also, statistically significant differences between Depression and Self-Harm (eRisk) and Depression and PTSD (CLPsych) are denoted by ^ (p-value < 0.001).



(a) Reddit (eRisk)

(b) Twitter (CLPsych)

Figure 5.11. Time-gap analysis by month.

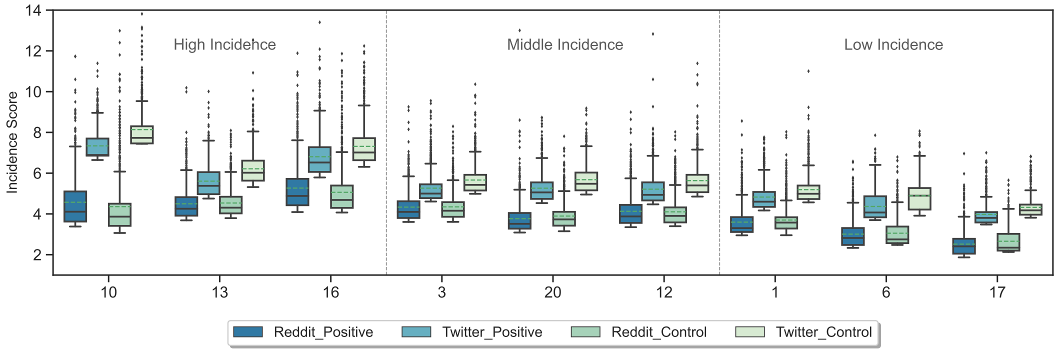
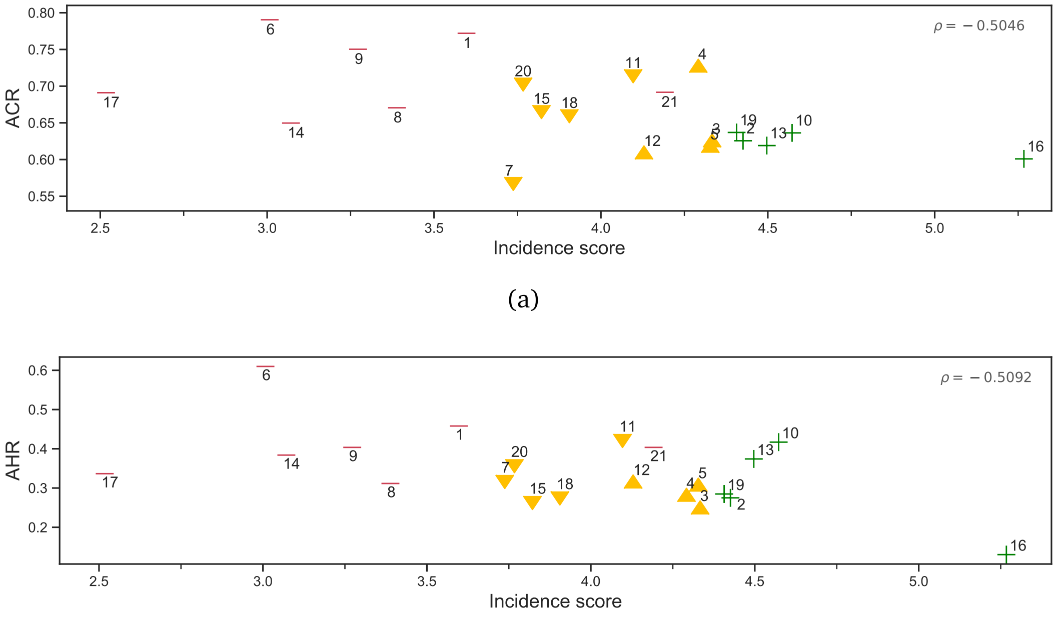


Figure 5.12. Distribution of incidence score of BDI items for the positive and control groups.



(a)

(b)

Figure 5.13. Correlation plot between incidence (computed utilising BM25 on Reddit) and average effectiveness (measured by ACR and AHR, with the eRisk 2020 systems). The Pearson's correlation coefficient is reported (adopting a Bonferroni correction significance it achieved at p-value $< 0.025\,(0.05/2)$). High (✚), Middle-High (▲), Middle-Low (▼) and Low (▬) incidences categories.

5.4 Discussion and Summary

The wealth of information encoded in continually-generated social media content is eager for analysis. In particular, social media data, that naturally occurs in a non-reactive way, is becoming a valuable complement for more conventional assessment instruments (such as questionnaires) used to determine the potential presence of mental disorders.

In this chapter, we reported results of a thorough analysis to show how users affected by mental disorders differ significantly from control individuals in the way they write and behave on social media. We investigated the writing style, how people express their emotions, and their online behavioural patterns on social media via visualising certain probabilistic attributes. To this aim, we analysed and visualised the activity, vocabulary, psychometric attributes, and emotional indicators in people's posts on two very different social media platforms. Studying and visualising such dimensions, we discovered several interesting differences that could *help a predictive system* or a health practitioner to determine whether someone is affected by a mental disorder. Across different mental disorders, however, we could not find any significant indicators to be able to distinguish one from the other. Moreover, we obtained evidence suggesting that the use of language in micro-blogging platforms, such as Twitter, where users are subjected to constraints influencing their writing style, is less distinguishable between users suffering a mental disorder than other less restrictive platforms, like Reddit. Therefore, we can confirm that analysing social media posts could help a system identify people that are more likely to be diagnosed with a mental disorder. However, determining the exact disorder is a much more difficult task, requiring human expert judgement. Also, we found that psychometric attributes, emotional expression and social engagement provide a quantifiable way to differentiate individuals affected by mental disorders from healthy ones.

In addition, we presented a detailed analysis towards improving the understanding of how depression assessment inventories, in particular the BDI, could be automatically estimated based on the evidence available on social media. We investigated the relative incidence of the 21 BDI items on users' feeds, as well as how incidence and other features influence the effectiveness of automatic tools that extract depression symptoms from social media posts. To quantify the incidence of different depression symptoms, we derived queries from the BDI items, retrieved documents from social media collections and analysed their relevance scores. We discovered that certain items have a consistent incidence (over different search methods and platforms), suggesting that the associated symptoms are more prevalent on these collections. However, the existence of more documental evidence about these topics also introduces more noise. In fact, we found that systems that attempt to infer the severity of depression symptoms from user data tend to perform poorly on high incidence BDI items. Low incidence items turned out to be the parts of the BDI questionnaire whose answers are easier to estimate using automated means. Furthermore, we discovered that the effectiveness of automatic systems is also inversely correlated with other features, such as the length of the BDI item.

Our findings open several possible avenues for further research. In particular, we are interested in the development of strategies to formulate meaningful queries from the BDI items. Here, we used a simple yet high-precision approach that could be considered as a baseline for future experiments. The generation of succinct and on-topic queries from BDI items represents a challenging task. The quality and specificity of the derived queries directly affects the relevance of the documents retrieved and, subsequently, the effectiveness of any system that mines depression symptoms from the retrieved data. Moreover, it would be interesting to compare automatically derived queries with those developed by human experts, such as psychologists and practitioners, in order to find differences and similarities as well as to study how one could complement the other. Another insightful avenue for future research is to investigate to what extent incidence is related with the weight a particular item has on the questionnaire. As the BDI was conceived, all the items share the same weight when estimating the depression severity of an individual. However, by examining the inventory's items one could question that certain items are more *important* than others. For instance, if an individual self-assessment scores higher on the BDI item related to suicide than that assessing the loss of interest in sex, there might be higher chances that the person is suffering from depression and, even more, that an intervention is quickly needed to avoid further consequences. Bearing this in mind, studying how items' importance on depression severity estimation is related to their incidence and specificity can contribute to better address the research and development of automatic assessment tools towards certain items more than others.

It is also important to note the limitations associated with the methodological choices done to conduct the study of depression incidence and its generalisability. As discussed before, this study provides a relative comparison between the evidence of the various BDI items readily available on social media. The categorisation scheme we chose allowed us to organise BDI items into different incidence groups facilitating the incidence analysis. Such scheme can be constructed in different ways, for instance by directly employing the incidence score (i.e., average relevance score) and defining different thresholds and incidences levels accordingly. One limitation we noted with our categorisation scheme is that, as observed in Figure 5.12, there could be items assigned to a certain category where those in a category immediately above (or below) might present a very narrow margin in terms of the retrieval scores. Thus, an item with middle-high incidence could also be considered to have high incidence (or the other way around), it is just that given its position in the ranking it end up in a lower category. Yet, with our categorisation scheme we can still paint a picture of the trends depicted by the information about each BDI item that can be retrieved from social media. Another limitation we observed underlies on the analysis granularity. We decide to conduct the study considering single posts as retrieval units (post-level). Alternatively, one could study the incidence at user-level by concatenating all the posts of a user into a single document. This could provide complementary information to better understand the incidence levels observed. For instance, the fact that an item has higher incidence is it because many users are discussing about this topic? Or is it because a few, very active users are recurrent on the topic? Although still informative such level of granularity would be limited by the number of available users in each collection. At a higher level, our goal is to shed light, using a finer-grain scale, on the feasibility and limits that language predictive technology, in particular, systems designed for automatically filling BDI questionnaire from online sources could present. In essence, we are interested in exploring and improving the understanding on how information can be extracted from social media sources and integrated with tools that could help health practitioners to perform a more comprehensive assessment of the individuals complementing traditional diagnostic methods. For this reason, we consider that our proposal is a good first approximation to tackle the problem providing useful insights into it, opening several avenues for future work and posing broader research questions.

The studies we presented throughout this chapter have a high practical impact since research could be steered towards building new metrics that can correlate with a disease before traditional symptoms arise and which clinicians can use as leading indicators of traditional later-onset symptoms.

Part II

Computational Personality Assessment

Chapter 6

Personality Cues From Conversations

6.1 Introduction

Recent years have witnessed the emergence of powerful technological platforms supporting expression of self. An increasing number of companies providing different services, such as recommendation or web search, seek to offer more effective and personalised solutions to their customers by taking into account various aspects of users' profile information. The goal of such personalised systems is to adapt the content, the interface or the services in general to match the preferences and characteristics of each individual user [226]. Traditionally, explicit and implicit user feedback[1] has been used to model users' tastes and behaviour. Thanks to the advent of robust models for unobtrusive analysis and profiling [234], personalised services can now leverage more detailed user profiles that include highly descriptive attributes, such as personality [198]. In fact, several works highlight the importance that personality profiling has gained over the years becoming an essential application for the marketing, advertisement and sales industries [94; 240]. As argued by Mehl et al. [140] the study of such latent mental-related attributes allow for better understanding of the reasons and motivations behind individuals' behaviour.

The Web is one of the many platforms where conversational agents (i.e.,chatbots) are increasingly communicating with humans to provide automatic services. Such agents are capable of interacting with humans through natural language interfaces usually to help them to complete various types of tasks. Tasks can range from simply setting an alarm to more complex cases like health advice [229]. As argued by Baeza-Yates, chat-based interactions provide deeper

[1] In this context, the term *feedback* refers to information collected from users regarding their reactions to a product, service, or website experience.

insights into users' intent and moods by recognising behavioural patterns and preferences[2]. Thus, conversational agents may be able to provide more effective services by improved dialogues, if they take into account user profile information [247] and personality traits of people who they are interacting with, and adapt their behaviour in response. In particular, it has been demonstrated that users exhibit different reactions to conversational agents depending on their own personality [30]. This introduces the need to develop systems which, apart from interpreting, and interacting with people, are able to adapt to their personality, akin to humans [139]. We envision more empathetic and naturalistic conversational agents and as a result of this vision, we take a first step in this avenue to improve over the state of the art of automatic personality recognition in conversations.

As stated in the introduction, individuals greatly differ from one another in their way of thinking, feeling and hence acting, constituting the basis of what is commonly referred to as personality. In this respect, the *lexical hypothesis* states that any trait relevant for describing human behaviour has a corresponding lexical token [96]. In fact, individuals exhibit certain specific patterns when they talk or write about other people or about themselves. Typically, these lexical tokens are adjectives which, once clustered together using factor analysis, reveal basic "traits" of personality. Despite its complexity and ambivalence, language can be highly informative for the study of personality [171; 157]. The main reason, as argued by Boyd et al. [20], is that language use is relatively reliable over time, consistent, and varies considerably between individuals.

Traditionally, personality is assessed with self-assessment questionnaires, such as the BFI [97] (Big Five Inventory), the S5 [107] (*Short Five*) or the TIPI [77] (Ten Item Personality Measure), sometimes complemented with one-to-one sessions with a professional. Although still useful, such methods rely on retrospective self-reports, whose validity might sometimes be affected by cognitive[3] and situations issues[4] [219], apart from being subjected to the observer-expectancy effect [75]. To overcome these drawbacks and find implicit measurements rather than explicit self-reports, personality researchers sought to complement traditional assessment approaches with novel solutions based on the application of computational methods [218]. Many useful cues about an individual's mental state, such as the individual's social and emotional conditions, and personality

[2] See `https://bit.ly/3wsckNn`.

[3] *Cognitive* issues address whether the respondents understand the question and whether they have the knowledge or memory to answer it accurately.

[4] *Situational* issues include the influence of the setting of the survey (at work, at home, etc.). Certain questions may have a socially desirable response.

can be captured by examining the language use of that individual [172; 39]. Therefore, the analysis of language use is beneficial for identifying and assessing human personality [242; 20] and to complement traditional assessment methods.

In this chapter, we present a novel approach to personality recognition in conversations based on the notion of "capsules" [88]. A capsule hosts a small group of neurons whose activities represent the various properties of a specific type of entity. The capsule-model learns a hierarchy of feature detectors through a routing-by-agreement algorithm [207]. In a conversation, these feature detectors can represent sets of words whose occurrence co-vary, thus revealing latent underlying personality traits. To the best of our knowledge, this is the first time a model based on capsule neural networks for the task of automatic recognition of personality in conversations is presented.

Moreover, motivated by the exploration of open-vocabulary features for gaining new psychological insights, we introduce a novel data-driven approach for the study of personality. More precisely, we aim at discovering linguistic patterns associated with a particular personality trait based on multiword expression. In essence, multiword expressions (or simply *collocations*) are sets of expressions consisting of two or more words that corresponds to some conventional way of expressing things.

Thus, our contributions in this chapter can be summarised as follows:

1. We propose a novel approach based on capsule neural networks for automatic recognition of personality in conversations.

2. We show and analyse the effectiveness of the proposed model on a real-world dataset of conversations.

3. We present a novel open-vocabulary approach based on multiword expressions to gain insights on personality and its manifestation through language.

4. We provide a quantitative and comparative analysis showing the potential of our method on two real-world datasets.

The remainder of this chapter is organised as follows. Section 6.2.1 details the components of the proposed model for automatic recognition of personality in conversations; Section 6.2.2 describes the experimental design and presents the results; Section 6.2.3 presents a qualitative analysis showing the interpretability potential of the capsule-based model; Section 6.3.1 details the procedure used to discover a set of candidate collocations and how they are ranked; Quantitative and comparative analysis of the results is presented in Section 6.3.2; finally, Section 6.4 concludes the chapter.

6.2 Personality Assessment using Capsule Neural Networks

As previously stated, the automatic identification of personality in conversations has many applications in natural language processing, like for example community role identification (e.g., group leader) in online social media conversations as well as meeting transcripts. Conversation utterances provide a lot of information about the parties involved in a conversation such as cues to the participants' personality traits, one of human's most distinguishable attributes. However, traditional computational personality assessment models rely on limited domain-knowledge and various psychometric indicators. As a matter of fact, none of the models proposed in the literature (see Section 2.2) for personality recognition approached this problem as a frequency co-variance between a number of sets of words modelled as capsules. Motivated by this, we propose an approach based on a fixed-sized document representation capable of modelling such changes in the data and hence predicting people's personality via their natural language conversations. As shown in Chapter 2, little work has been done in the domain of personality assessment in conversations.

6.2.1 Model Composition

In this section, we outline the model architecture we propose for performing the automatic assessment of personality in conversations. In essence, the model is comprised of two components and operates at two different levels. The first component, which works at the utterance level, is a capsule neural network trained to determine if a particular utterance is conveying evidences of a personality trait. The second component, which operates at the user level, is a shallow classifier which receives the predictions produced from the capsule-based component regarding the utterances of a particular user to perform the personality assessment of the individual.

Capsule-based Component: Capsule-based models have achieved state-of-the-art performance in image recognition [207] and has shown great potential for different text classification tasks [241], such as intent detection [239]. The activities of the neurons within an active capsule are supposed encode the important information about the state of the property of the particular entity they are detecting. Such information is stored as a vector. While the orientation of the vector represents the different components (know as “instantiation parameters”) of a particular entity (e.g., shape, size and colour of a face in an image), the length encodes the probability that the entity represented by the capsule exists in the current input (e.g., how likely a face with certain attributes is present in the image).

Each layer in a capsule neural network contains many capsules which through a mechanism called routing-by-agreement, learn a hierarchy of feature detectors. For detecting low-level properties (e.g., mouth, eyes), capsules broadcast their activations to high-level capsules (e.g., a face) only when there is a strong agreement of their prediction with such high-level capsules. This means that when multiple predictions agree, a higher level capsule becomes active. It is clear then that high-level capsules encode more complex entities with more degrees of freedom. The larger the dimension of a single capsule, the more properties of an entity it can represent. However, larger dimension increases the computational complexity.

We consider that given all the above-mentioned characteristics, capsule-based models could be suitable for modelling user’s utterances for personality assessment through conversation transcripts. Low-level capsules will encode different semantic attributes of an utterance which, given the presence of certain instantiation parameters, will collectively contribute to determine the existence of a more abstract entity representing a personality trait. Hence, the appropriate contribution of each semantic attribute can be determined in an unsupervised manner using a dynamic routing-by-agreement mechanism and finally aggregated to obtain the personality trait representation. The proposed architecture is shown is Figure 6.1.

The two capsule-based components comprising the network architecture are:

PrimaryCaps: Primary capsules are the lowest level of multidimensional entities. In the original proposal by Sabour et al. [207], tailored for image recognition, the first layer is composed of convolutional units in charge of outputting different vectors encoding various features from the input image. Following the same spirit, we compute different semantic features from the raw conversation utterance based on a bi-directional recurrent neural network with multiple self-attention heads [116]. The key idea is to extract different aspects of the semantics in the utterance and encode them into low-level vector representations. To

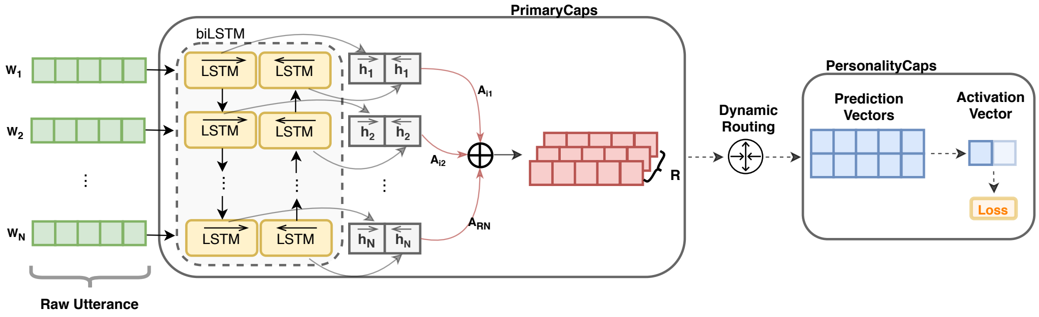


Figure 6.1. Capsule-based Component Architecture.

this end, each self-attention head focuses on a specific component of the input conversation utterance and produces a semantic feature that might not be expressed by words in proximity.

Suppose we have an input utterance composed of N tokens.

$$u = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)$$

Each token in the sequence is represented by a vector of dimension D_W as observed in Figure6.1. Such vector representation could be pre-trained, using some method such as word2vec [145] or FastText [18], or even learned while training. We use a recurrent neural network such as a bidirectional LSTM [89] (biLSTM) to sequentially process the utterance:

$$\begin{aligned} \overrightarrow{\mathbf{h}}_t &= LSTM_{fw}(\mathbf{w}_t, \overrightarrow{\mathbf{h}}_{t-1}), \\ \overleftarrow{\mathbf{h}}_t &= LSTM_{bw}(\mathbf{w}_t, \overleftarrow{\mathbf{h}}_{t+1}) \end{aligned} \tag{6.1}$$

For each token $\mathbf{w}_t$ in the conversation utterance, the hidden state $\overrightarrow{\mathbf{h}}_t$ obtained from the forward pass $LSTM_{fw}$ and the hidden state $\overleftarrow{\mathbf{h}}_t$ from the backward pass $LSTM_{bw}$ are concatenated to generate a hidden state $\mathbf{h}_t$. The hidden matrix obtained is defined as:

$$\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N) \in \mathbb{R}^{N \times 2D_H},$$

where D_H represents the number of hidden units in each unidirectional LSTM.

Our goal is to encode a variable length utterance into a fixed size embedding. To achieve this goal, we compute a linear combination of the N biLSTM hidden vectors in H following a multi-head self-attention framework. Each self-attention head is encouraged to focus on a specific component of the utterance, like a special set of related words or phrases.

A self-attention annotation matrix **A** is computed as:

$$\mathbf{A} = softmax\,(\mathbf{W}_{s2}\, tanh(\mathbf{W}_{s1} H^T)), \tag{6.2}$$

where $\mathbf{W}_{s1} \in \mathbb{R}^{D_A \times 2D_H}$ and $\mathbf{W}_{s2} \in \mathbb{R}^{R \times D_A}$ are weight matrices for the self-attention. D_A is a hyperparameter which denotes the number of hidden units of self-attention and can be set arbitrarily. R represents the number of self-attention heads, i.e., the different components in a conversation utterance to attend which reflect the overall semantics of the sequence. The softmax function ensures that the attentive scores on all the tokens sum to one for each self-attention head. Equation 6.2 can be seen as a 2-layer Multilayer Perceptron (MLP) without bias.

In total R semantic features are extracted from the input conversation utterance, each from a separate self-attention head: $\mathbf{M} = \mathbf{AH}$, where

$$\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_R) \in \mathbb{R}^{R \times 2D_H}$$

Each $\mathbf{m}_r$ is a $2D_H$-dimensional embedding vector. Every embedding vector posses a distinguishable orientation when the objective is properly regularised (see Equation 6.6), since it is desirable to prevent the different attention heads from suffering from redundancy by focusing on the same component of the conversation utterance.

We note that capsules exhibit an intrinsic ability to aggregate semantic features at lower levels which contributes to representing higher level latent personality distribution of the utterance. The orientation of a vector encodes semantic/personality attributes which, depending on the way individuals communicate, could slightly vary. The network encourages the learning of more general semantic embedding vectors, as less informative semantic attributes for a particular personality dimension may not be penalised by their orientation. They just have small length since their likelihood of existence is marginal.

PersonalityCaps: The basic idea behind the dynamic routing is to build a non-linear mapping in a iterative fashion guaranteeing that the output of each capsule is broadcast to an appropriate parent in the hierarchy. Therefore, the PersonalityCaps select among different R semantic features produced by the PrimaryCaps in order to construct a higher-level representation of each personality by means of an unsupervised routing-by-agreement mechanism.

Since the various semantic features might contribute with diverse intensities when detecting a personality dimension, the PersonalityCaps encode semantic features with regards to each personality. This is achieved by multiplying each $\mathbf{m}_r$ by a weight matrix $\mathbf{W}_{ri}$:

$$\mathbf{p}_{i|r} = \mathbf{W}_{ri}\,\mathbf{m}_r, \tag{6.3}$$

where $i \in \{1, 2, \ldots, I\}$, $r \in \{1, 2, ..., R\}$. $\mathbf{W}_{ri} \in \mathbb{R}^{2D_H \times D_P}$ is the weight matrix if the PersonalityCaps, $\mathbf{p}_{i|r}$ is the "prediction vector" from the layer below, i.e., the r-th semantic feature of a personality dimension i, and D_P is the dimension of the prediction vector.

The prediction vectors obtained from the PrimaryCaps are dynamically routed to the PersonalityCaps, which compute a weighted sum over all the incoming prediction vectors with:

$$s_i = \sum_r^R c_{ri}\mathbf{p}_{i|r}, \tag{6.4}$$

where c_{ri} denotes the coupling coefficient which establishes how much the r-th semantic feature is contributing to personality dimension i. The coupling coefficients between a PrimaryCap r and all the PersonalityCaps sum up to one and are computed by a so-called routing softmax. Its initial logits b_{ri} are log prior probabilities that PrimaryCap r should be coupled to PersonalityCap i. This mechanism known as dynamic routing-by-agreement [207] allows to iteratively refine the initial coupling coefficient. We outline the complete procedure in Algorithm 1.

Algorithm 1 Routing Algorithm

```
procedure ROUTING($\mathbf{p}_{i|r}$, $s$)
    for all PrimaryCaps $r$ and PersonalityCaps $i$: $b_{ri} \leftarrow 0$
    for $s$ iterations do
        for all PrimaryCaps $r$: $c_r \leftarrow softmax(\mathbf{b}_r)$
        for all PersonalityCaps $i$: $s_i \leftarrow \sum_r c_{ri}\mathbf{p}_{i|r}$
        for all PersonalityCaps $i$: $v_i = \mathtt{squash}(s_i)$
        for all PrimaryCaps $r$ and PersonalityCaps $i$: $b_{ri} \leftarrow b_{ri} + \mathbf{p}_{i|r} \cdot v_i$
    return $v_i$
```

In this, `squash(.)` is a non-linear function which guarantees that short vectors get shrunk to almost zero length and long vectors get shrunk to a length slightly below one, leaving the orientation unchanged and scaling down its magnitude.

It is applied to s_k to get an activation vector v_k for each personality dimension class i:

$$v_i = \frac{||s_i||^2}{1 + ||s_i||^2} \frac{s_i}{||s_i||} \tag{6.5}$$

The length (i.e., the norm) of the instantiation vector v_i determines the probability of existence of a certain personality dimension in the input utterance, while its orientation encodes the different attributes that personality dimension. To guarantee that the outputs of the PrimaryCaps are broadcast to the relevant PersonalityCaps, c_{ri} receives a low value when inconsistency exists between $\mathbf{p}_{i|r}$ and v_i.

Since our goal is that top-level capsule for personality dimension class i shows a long instantiation vector if and only if that personality dimension is evident in the utterance, the loss function considers the max-margin loss in each labelled utterance, and additionally it encourages the diversity between the components of the sequence to which the self-attention heads are focusing:

$$\begin{aligned} L = \sum_{i=1}^{I} \{ & Y_i \cdot max(0, m^+ - ||v_i||)^2 + \\ & \lambda(1 - Y_i) \cdot max(0, ||v_i|| - m^-)^2 \} + \\ & \alpha \left\| \mathbf{AA}^T - I \right\|_F^2 , \end{aligned} \tag{6.6}$$

where $Y_i = 1$ if the i personality dimension is evident in the utterance, λ is a down-weighting coefficient, m^+ and m^- are margins and α is a non-negative trade-off coefficient which rewards the differences between the attention heads [116].

Shallow Classifier Component: The second component of the model we propose takes advantage of the predictions generated by the capsule-based component to determine the personality of a user. We present two viable methods.

Score-based Classifier: We use the scoring function outlined in Equation 6.7, previously proposed by [11] to construct a context sensitive and self-maintaining sentiment lexicon. The function takes as input the predictions produced by the capsule-based component. That is, the number of utterances from a user classified as being of personality type X (positive), and the number of utterances classified as not being of personality type X (negative). For instance, extrovert versus non-extrovert. The function outputs a score which tell us the degree in which a certain personality is evident in the user language. As it can be observed the resulting score is in the range $\{-1, +1\}$.

$$Score(user) = \frac{\#of Positive - \#of Negative}{\#of Positive + \#of Negative} \tag{6.7}$$

If the score exceeds a threshold of 0.1, the user is classified as being of personality type X. We decide to use this value as we do not want to be too stringent with the decisions, at the cost of allowing more mistaken decisions. This model is named CapsScore in Table 6.1.

Statistical Classifier: The second alternative consists in averaging the norm of PersonalityCaps prediction vectors taking into account all the utterances of a user; thus, deriving a probability score for each personality dimension. The outcome of this procedure will be ten different scores. Recall from Chapter 1 that each dimension of the Big Five has its corresponding opposite and that is the reason for the ten scores. Finally, the scores are fed to a statistical classifier, such as a Decision Tree [118] (DT), to obtain the personality of a user under assessment. This model is named CapsStat in Tables 6.1 and 6.2.

6.2.2 Experimental Design and Results

In this section we outline the evaluation framework followed to assess the effectiveness of our proposed model presented in Section 6.2.1. We outline the dataset used for the experiments, the performance metrics assessed, the baselines and the corresponding analysis and interpretation of the results.

As a test bed for our experiments we use the EAR Conversations dataset gathered by Mehl et al. [140], described in Section 3.3.1. Recall that this dataset is comprised of a set of daily-life conversation snippets from a set of individuals collected for a span of two days and accompanied by personality ratings following the descriptions of the BFI. The approach we propose classifies subjects into the value of each of the Big Five dimensions (Chapter 1). More precisely, the main task, i.e. personality assessment, is subdivided in five binary classification sub-tasks. For example, in one of the sub-tasks the model is trained to determine whether a subject is extrovert or non-extrovert. The same procedure is applied for each of the five personality dimensions. This is the standard procedure followed when conducting personality assessment tasks and using Big Five scores. It allows to have a better control on class distribution and easier-to-interpret prediction tasks [32]. Ground truth labels in the conversations dataset were derived by converting the Big Five scores to nominal classes with a median split. The resulting classes are balanced in the Big Five set.

Since the five sub-tasks we intend to perform are binary classification problems, the effectiveness is evaluated using classic metrics, namely, Precision (P), Recall (R) and F_1. The overall performance is computed by taking the average of the five sub-tasks per each subject. In a real-life application (e.g., pre-employment personality assessments) computing such averaged performance would probably lack of a direct interpretation as each Big Five pair, comprised of a trait and its corresponding complementary dimension, is predicted individually. Nonetheless, this average still provides us with a general view of the effectiveness of any personality assessment approach when estimating an individual's personality (being this comprised of five traits and their complements).

We compare the proposed model with other text classification state-of-the-art alternatives used in the literature for automatic assessment of personality:

- N-grams [93; 32]: We use unigrams, bigrams and trigrams weighting them with TF-IDF to represent users' utterances and fit a DT classifier;
- LIWC [130; 93; 32]: We use LIWC [172] psycholinguistic lexicon to compute the percentage of words of each category present in the users' utterances. These features are used to train a DT classifier;
- NEU-14 [158]: We use the method proposed by Neuman et al. (Section 2.2) to obtain similarity scores between the users' utterances and personality dimensions vectors. The similarity is calculated using pre-trained FastText word embeddings [18]. The resulting scores are used to train a DT classifier.

It should be noted that in all the baselines and in our models, we use the same learning method, a DT classifier. The rationale behind this is to make a fair comparison among the different models without relying on a particular algorithm that might benefit a certain group of features.

The hyperparameters of the capsule-based component used for experiments are chosen using three-fold cross validation. D_W is 300 as FastText pre-trained word embeddings are used to represent the raw input utterance. Regarding the PrimaryCaps, the number of hidden units in each LSTM (D_H) is set to 32, the number of hidden units of self-attention (D_A) to 20 and the number of self-attention heads (R) to 3. The dimension of the prediction vectors of PersonalityCaps (D_P) is 10. In the loss function, the down-weighting coefficient λ is 0.5, margins m^+ and m^- are set to 0.9 and 0.1, respectively, and the trade-off coefficient α to 0.0001. The routing iterations number $iter$ (Algorithm 1) used is 2. Adam optimiser [105] is employed to minimise the loss. An input dropout layer [217] with a dropout rate of 0.8 is used to avoid overfitting issues.

Table 6.1 depicts the results obtained after conducting each individual experiment. The three evaluation metrics are averaged over the five personality dimensions. As observed from the experiments, both of our models CapsScore and CapsStat outperform the baselines. In particular, F_1 improves by a 10% with respect to LIWC and NEU-14 and over a 20% with respect to N-grams. Despite its simplicity, the CapsScore model achieves a reasonably good performance. One of the advantages of our model is that it does not depend on any domain-knowledge or external resource as LIWC or NEU-14 do.

Table 6.2 shows a comparison between our most effective model (CapsStat) and the best baseline (LIWC) for each of the Big Five dimensions. Overall, our model outperforms the baseline in terms of F_1 in all dimensions except for Extraversion. Intuitively, one explanation could be found in the attributes that characterise extrovert individuals. Overall, extraversion includes traits such as talkative, energetic, assertive, and outgoing. In terms of language expression, individuals who score highly on extraversion are chattier than their more introverted peers and their language tend to be more abstract and *loose*, while introverts tend to spoke in more concrete terms. Hence, it is possible that, even though the capsule-based component is able to extract distinctive semantic features from the conversation utterances, most of the time there is not enough agreement between the low-level capsules to activate a higher level capsule, and thus denote the presence of extraversion in the input. In other words, the probability of the various semantic attributes represented by each capsule, and comprised of different instantiation parameters, are not sufficient to identify the personality trait in question.

Table 6.1. Overall effectiveness achieved by the baselines and our models on recognising personality from the conversation dataset. Evaluation metrics are averaged over the five personality types. The largest values achieved for each metric are highlighted in bold.

Model	**R** (%)	**P** (%)	$\mathbf{F_1}$ (%)
N-grams	42.31	46.89	44.21
LIWC	60.15	51.61	54.43
NEU-14	50.91	56.74	53.43
CapsScore	61.48	68.10	58.83
CapsStat	**61.55**	**68.33**	**64.68**

Therefore, as a result of the two presented experiments we conclude that our proposed models outperform all the baselines in terms of personality identification in conversations.

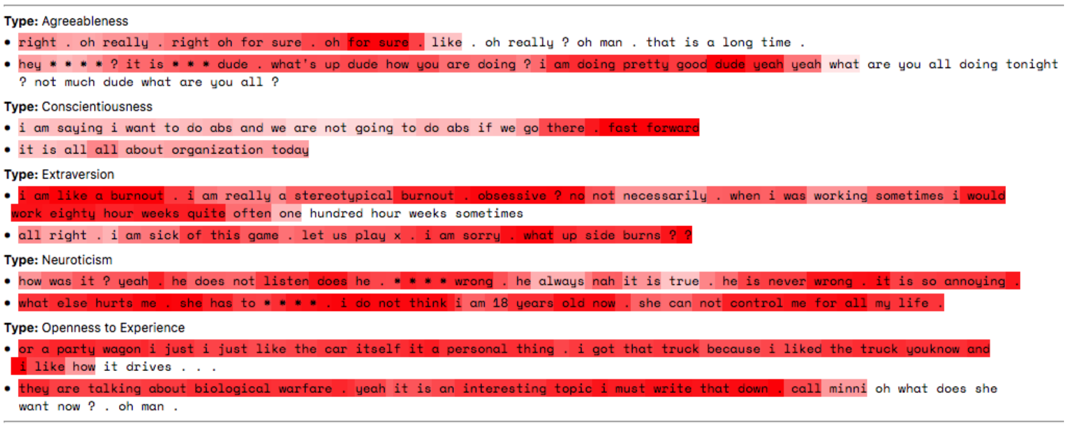
Table 6.2. Performance comparison between our most effective model (CapsStats) and the best baseline (LIWC) for each of the Big Five personality types. The largest values achieved for each metric are highlighted in bold.

Type	LIWC			CapsStats		
	R (%)	**P** (%)	$\mathbf{F_1}$ (%)	**R** (%)	**P** (%)	$\mathbf{F_1}$ (%)
AGR	**83.33**	45.45	58.82	61.54	**72.73**	**66.67**
CON	66.67	50.00	57.14	**75.00**	**75.00**	**75.00**
EXT	**72.73**	**72.73**	**72.73**	58.33	63.64	60.87
NEU	36.36	44.44	40.00	**54.55**	**66.67**	**60.00**
OPN	41.67	45.45	43.48	**58.33**	**63.64**	**60.87**

6.2.3 Interpreting Personalities

One interesting feature of capsule-based neural models is its inherent interpretability potential. As a step towards utilising this potential, we present a qualitative analysis of the self-attention annotation matrix **A** within the PrimaryCaps and visualise the attention weights for the various personality traits. Recall that each element in this matrix describes how much the biLSTM hidden state at a particular time step (i.e., representing a token in that time step) contributes to producing the resulting embedding. This provides a general view of what the network mainly focuses on. In essence, the annotation matrix allows to understand which words the network attends more and which ones receive less attention.

Figure 6.2 provides qualitative examples for each of the Big Five personality traits. We notice in the figure that the different self-attention heads learn to focus on particular sets of words/phrases of the utterance according to the personality under consideration. For instance, in Figure 6.2 we observe that an *Agreeableness* personality exhibits usage of words commonly used to express agreements and compliance, while a person with *Openness to Experience* for example, describes the car owned “as a party wagon” and shows interest in exploring new things (e.g. the individual has bought a new car). Moreover, an individual with *conscientiousness* personality is very self-conscious regarding future planning and shows interest in self-organisation.



Type: Agreeableness

- right . oh really . right oh for sure . oh for sure . like . oh really ? oh man . that is a long time .
- hey * * * * ? it is * * * dude . what's up dude how you are doing ? i am doing pretty good dude yeah yeah what are you all doing tonight ? not much dude what are you all ?

Type: Conscientiousness

- i am saying i want to do abs and we are not going to do abs if we go there . fast forward
- it is all all about organization today

Type: Extraversion

- i am like a burnout . i am really a stereotypical burnout . obsessive ? no not necessarily . when i was working sometimes i would work eighty hour weeks quite often one hundred hour weeks sometimes
- all right . i am sick of this game . let us play x . i am sorry . what up side burns ? ?

Type: Neuroticism

- how was it ? yeah . he does not listen does he . * * * * wrong . he always nah it is true . he is never wrong . it is so annoying .
- what else hurts me . she has to * * * * . i do not think i am 18 years old now . she can not control me for all my life .

Type: Openness to Experience

- or a party wagon i just i just like the car itself it a personal thing . i got that truck because i liked the truck youknow and i like how it drives . . .
- they are talking about biological warfare . yeah it is an interesting topic i must write that down . call minni oh what does she want now ? . oh man .

Figure 6.2. PrimaryCaps self-attention weights visualisation for each of the Big Five personality types.

It is noteworthy, that although we observe clear differences in word usages among the different personality types in the presented examples, strict comparison of the utterances with respect to personality types may not be correct, as each individual (i.e., author of an utterance) can possess more than one personality type. However, looking at various cues expressed in different utterances authored by a certain individual can reveal latent personality types encoded in the individual's language usage.

6.3 Discovering Linguistic Patterns of Personality

As stated in the introduction of this chapter, the assessment of human personality can be useful for a variety of potential applications, ranging from product recommendation to human resources management. Hence, the study of linguistic personality-related attributes in natural language provides a *window* into individuals mind for a better understanding of the reasons behind their behaviour [140].

In this section, we present a novel open-vocabulary approach based on multiword expressions towards improving the understanding of personality manifestation through language. The advantage of open-vocabulary approaches is that they are not limited to pre-defined words lists (Section 2.1.2). Instead, linguistic features are automatically determined from the text (i.e., it is "data-driven"). One appealing advantage of this kind of methods is that they reveal more specific and concrete patterns across a broad range of content domains and are less prone to misinterpretation, suggesting that they are well-suited for capturing the nuances of everyday psychological processes [57]. We put into practice our approach using two real-word datasets which show its potential through quantitative and visual analyses.

6.3.1 Collocations Discovery and Ranking

Collocations, more generally know as multiword expressions, are sets of expressions consisting of two or more words written in close proximity that corresponds to some conventional way of expressing things. Here, we outline the procedure conducted to discover and select a meaningful set of collocations which are intended to capture the most distinctive linguistic patterns of a personality trait. First, we describe how, in an unsupervised manner, we discover a set of candidate collocations from the text produced by individuals with a particular trait. Afterwards, we define a collocation discrimination scoring function taking into account their co-occurrence patterns.

Collocation Discovery: It should be noted that here we are considering a less strict definition of collocation. In essence, we are interested in discovering combinations of words which are in a looser relationship than fixed phrases (sequence of consecutive words), and thus they are variable with respect to intervening material and relative position. For example, in the sentences *she knocked on his door* and *he knocked on the metal front door*, the collocation “knocked door” is accounted despite the variable number of tokens in-between. This allows us to capture combinations of words which, despite the fact that their distance is not constant, show enough regularity in their co-occurrence to reveal some notion of *style*.

An association measure give us a way to quantify how related the words are in a collocation ensuring that phrases we do keep are informative parts of speech and not just accidental juxtapositions. In essence, what we want to know is whether two (or more) words occur together more often than chance. Here, we choose the Pearson’s chi-square test (χ^2) as our association measure. It assumes as the null hypothesis that words comprising a collocation are independent. When the difference between observed and expected frequencies is large enough the null hypothesis is rejected denoting that those words do not occur independently of each other. It should be noted that we are searching for particular patterns in the data. However, we are also taking into account how much data we have seen. Even if there is a remarkable pattern, we will discount it if we have not seen enough data to be certain that it could not be due to chance. One advantage of χ^2 test is that it does not assume normally distributed probabilities, which is usually the case [40]. In the case of two words, w_1 and w_2, the χ^2 test is applied to a 2-by-2 table (like Table 6.3) and is calculated as follows:

$$\chi^2(w_1, w_2) = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

where i iterates over the rows of the table, j iterates over the columns, $O_{i,j}$ is the frequency observed for cell (i, j) and $E_{i,j}$ is the frequency expected (computed from the marginal probabilities).

The χ^2 test can be computed for collocations comprised of n number of words, such as bigrams, trigrams or quadgrams.

Table 6.3. An example of a 2-by-2 table showing the dependence of occurrences of high and school for the computation of the χ^2 test. The words x and y are any two other words in the vocabulary different from w_1 and w_2.

	$w_1 = high$	$w_2 \neq high$
$w_2 = school$	fr(high school)	fr(x school)
$w_2 \neq school$	fr(high x)	fr(x y)

Collocation Ranking: Not all the collocations discovered will have the same importance. Thus, we select a subset of representative collocations to produce a set of candidates. Empirical analyses showed that the association scores follow quite closely a power-law distribution (also known as Zipf's law [248]). A power-law implies that small occurrences are extremely common, whereas large instances are extremely rare. Zipf showed through a series of empirical studies that the distribution of word use was due to tendency to communicate efficiently with least effort. In essence, he showed that when people write, usually, prefer *familiar* words rather than *rare* ones. Lunh [125] who studied words' representative power from their frequency distribution, suggested that the words which better described the content of a collection of texts are located in an area comprised between high-frequency words and rare ones (low-frequency). Inspired by this, we set a cutoff at both extremes of the association scores distribution and select the those closed by this upper and lower bound as candidate collocations.

Next, we estimate a language model, using the maximum likelihood estimate (MLE) [100] considering the collocations in the candidate set. Assume that we have two collections of documents. The first collection, C^+, comprises a set of documents authored by individuals rated as extroverts. While the second one,

C^-, comprises a set documents authored by individuals rated as introverts. Now, suppose that we want to want to create a set of meaningful collocations for C^+. First, we will derive a set of candidate collocations, S^+ and, then, we will estimate a language model L^+. In addition, we will derive a second set of candidate collocations, S^*, by merging C^+ and C^- and compute a language model L^* as well.

Following the work by Jordan et al. [98], we use Kullback-Leibler divergence (KL) a well-known measure from probability theory and information theory used to quantify how much two probability distributions differ. Hence, it can be used for comparing language models. Bearing this in mind, we will rank the collocations in S^+ based on how much they contribute to the KL-divergence between L^+ and L^* as follows:

$$score(c) = P(c|L^+) \log \frac{P(c|L^+)}{P(c|L^*)},$$

where c is a candidate collocation in S^+, $P(c|L^+)$ is the probability of c for L^+ and $P(c|L^*)$ is the probability of c for L^*. Moreover, we use the linear interpolation method, referred to as Jelinek-Mercer smoothing, to remove the zero probabilities assigned to collocations which are in L^* but do not appear in L^+. The same procedure can be applied to derive a meaningful set of collocations from C^-.

6.3.2 Descriptive and Comparative Analysis

Recall that the standard procedure followed when conducting personality analysis tasks is to derive labels for each trait by converting the Big Five scores to nominal classes with a median split. This allows to have a better control on classes distribution and easier-to-interpret tasks [32]. Our study of personality is conducted by taking into account a particular trait and its opposite dimension at the same time.

Here, we focus on Extroversion (EXT) and its corresponding complementary dimension, Introversion (INT), although, the method and analyses could be easily extended to the remaining traits of the Big Five. First, we split the subjects and their corresponding collection of documents in two subsets. One subset groups those individuals who fall somewhere along the continuum with regard to the positive dimension (EXT), and the other contains those in the complementary dimension (INT). To conduct the experiments we use the EAR Conversations dataset (Section 3.3.1) and Stream-of-Consciousness (SoC) Essays dataset (3.3.2). In this way, we can study and compare the manifestation of personality through spoken and written language.

Collocation Discovery: We study three groups of collocations: bigrams, trigrams and quadgrams. For each group, we define a collocational window of up to five tokens, thus allowing for a looser relationship between the words in the collocation than fixed phrases (Section 6.3.1). For instance, in the case of bigrams this means that up to three words could exist between the two words comprising that collocation.

As stated before, when we sort the collocations discovered using the association score we see that the resulting distribution closely follows a power-law. Thus following Lunh's work, we remove a certain percentage of the collocations from the higher and lower ends of the distribution to produce the final set of candidates. The same procedure is repeated for each trait and group of collocation. We empirically studied different cutoff thresholds along with the union, intersection, difference and Jaccard's similarity index [115] of the sets of collocations discovered for EXT and INT (independently). In particular, we observed that when the cutoff increases Jaccard's index decreases. This suggests that more general and common collocations between the two traits are removed while those more specific are kept. We decide to use an cutoff of 5% (removing %5 from each side of the distributions) as we consider it a reasonable trade-off between conserving a specific set of collocations while retaining some intersection.

Table 6.4 shows a quantitative and comparative analysis between the sets of candidate collocations discovered for EXT and INT using χ^2 as the association measure in the different collections. We analyse the union, intersection, difference and Jaccard's similarity index of the candidate collocations selected for each trait. We note that as the number of words in the collocations increases, Jaccard's index between EXT and INT decreases. This suggests that as more words comprised the collocation, e.g. quadgrams, more distinctive patterns are captured since the sets discovered for each trait become more diverse. The difference between the sets shows that there are collocations which exist for one trait but not for the other, denoting a specific personality-related linguistic patterns. As can be observed, the number of candidates collocations selected for SoC Essays is several order of magnitude larger than those found for EAR Convesations. This could be due to the fact conversations are comprised of utterances which are usually much shorter (in number of words) in comparison with essays' sentences. Thus, the number of possible combination of words originating a collocation is smaller. Moreover, as stated before, by applying this method we are looking for particular patterns in the data considering how much data we have observed. There might exist meaningful collocations which will be discounted it there is not enough evidence to be sure that their co-occurrence is not due by chance. In this respect, we also study the collocations discovered from the combination of both datasets. The goal of this experiment is to study whether there are collocations which are common in both contexts but the evidences found in each individual dataset are not enough to consider them as meaningful collocations. We observe that as more data is considered the number of collocations increases, giving the intuition that there are collocations which are use both in spoken and written language.

Table 6.4. Quantitative and comparative analysis between the sets of collocations discovered for EXT and INT using EAR Conversations dataset, Stream-of-Consciousness Essays dataset and their combination.

Source		Bigrams	Trigrams	Quadgrams
EAR Conver.	# of Collocations EXT	1,869	442	33
	# of Collocations INT	1,519	248	11
	Jaccard's Index (EXT/INT)	0.32	0.17	0.00
	Difference Size (EXT/INT)	1,082	342	33
	Difference Size (INT/EXT)	580	18	11
SoC Essays	# of Collocations EXT	30,787	3,198	266
	# of Collocations INT	26,923	2,451	168
	Jaccard's Index (EXT/INT)	0.41	0.23	0.10
	Difference Size (EXT/INT)	14,021	2,157	225
	Difference Size (INT/EXT)	10,157	1,410	127
Combined	# of Collocations EXT	33,803	4,391	301
	# of Collocations INT	29,319	3,232	182
	Jaccard's Index (EXT/INT)	0.41	0.22	0.10
	Difference Size (EXT/INT)	15,477	2,712	259
	Difference Size (INT/EXT)	10,993	1,669	140

Collocation Ranking: For EXT and INT we rank the candidate collocations using the discriminating scoring function defined in Section 6.3.1 and keep the top-100 for each trait. Tables 6.5 to 6.7 showcase the top-10 collocations with the largest score selected with our approach. It should be noted that for this study stopwords have been removed as our approach is quite sensitive to their presence. Recall that here we are using a less strict notion of collocations, thus allowing a variable number of tokens in-between the collocation words. With this in mind, several common constructions in natural language can be simply comprised of stopwords as they frequently used together. Thus, this might hinder the discovery of meaningful collocations just because they are not as frequent as collocations comprised of stopwords.

A close inspection to the top-100 ranked collocations for each trait reveals the ability of our approach to capture linguistics patterns of personality. Several collocations found for EXT are correlated with findings confirmed by previous research, like the inclusion of words with social connotations or related with outdoor activities [213] (e.g., fraternity, meeting, concerts) as well as sentimental expression. Furthermore, we discover co-occurrences of words unveiling novel patterns. For EXT we observe a consistent use of swearing words, and the habit of using capitals (denoted as *allcaps*). While in the case of INT, the inspection reveals expressions associated with self-assessment and introspection, such as *eating healthy* or *my thoughts* or the inclusion of the word *think* in the collocation. We also observe that in conversations quadgrams are most of the time the result of repetitions. A fact which is not so common in written language.

Table 6.5. Top-10 Collocations with the highest discriminating score in the EAR Conversations datasets.

	Bigrams		Trigrams		Quadgrams	
	INT	EXT	INT	EXT	INT	EXT
1	like like	um um	hey hey hey	one one one	hey hey hey hey	one one one one
2	oh ok	like know	like like like	yeah yeah yeah	uh hum uh hum	yeah yeah yeah yeah
3	yeah right	kind like	hum uh hum	um um um	look look look records	amen amen amen amen
4	right right	uh um	uh hum uh	oh yeah yeah	choose okay okay okay	um um um um
5	yeah okay	um like	uh hum hum	yeah well like	hum uh hum hum	you know like know
6	yes come	um going	like yeah like	know yeah yeah	okay okay okay paid	go go go go
7	hum uh	like um	yeah like yeah	yeah yeah talk	say look look look	point points point points
8	hey hey	like uh	uh uh hum	like like want	yeah yeah yeah true	oh yeah yeah oh
9	yeah think	oh good	look look look	uh like uh	like like oh gosh	yeah yeah well like
10	ok yeah	yeah want	like think like	like back like	uh huh uh huh	know yeah yeah yeah

Table 6.6. Top-10 Collocations with the highest discriminating score in the SoC dataset.

	Bigrams		Trigrams		Quadgrams	
	INT	EXT	INT	EXT	INT	EXT
1	know write	much time	head head head	allcaps allcaps allcaps	typing typing typing typing	sex sex sex sex
2	want want	like know	let us see	love love love	food food food food	lit lit lit way
3	something something	really glad	want go home	thinking thinking thinking	mom mom mom mom	love love love love
4	know think	really get	minute one minute	feel like time	head head head head	one really like sorority
5	think something	think people	one minute one	really need go	oh well big deal	hour guy seriously think
6	think type	really really	want go something	makes feel like	junior high high school	concert back realized going
7	really annoying	kind like	really like like	feel like life	people people mom mom	really like bother thing
8	thing right	people like	looks like minutes	really looking forward	party party class party	go concert two days
9	kind music	think good	feel like better	one best friends	well let us see	supposed go concert two
10	know hard	think things	twenty minutes guess	get stuff done	really want go back	guy seriously making things

Finally, in Figure 6.3 we visually analyse the top-1000 trigrams produced by our method by plotting the BERT-embeddings [54] for both extroversion and introversion. We observe various groups of semantically related words (clouds

Table 6.7. Top-10 Collocations with the highest discriminating score in the combined datasets.

	Bigrams		Trigrams		Quadgrams	
	INT	EXT	INT	EXT	INT	EXT
1	know write	like know	head head head	one one one	typing typing typing typing	sex sex sex sex
2	want want	kind like	want go home	yeah yeah yeah	food food food food	one one one one
3	something something	much time	let us see	oh yeah yeah	mom mom mom mom	yeah yeah yeah yeah
4	like year	really glad	like think like	oh oh yeah	hey hey hey hey	love love love love
5	think type	really get	one minute one	allcaps allcaps allcaps	head head head head	back realized going different
6	think something	really really	minute one minute	love love love	uh hum uh hum	one really bother sorority
7	really annoying	one one	want go something	yeah well like	oh well big deal	looked like pretty box
8	feel home	like like	looks like minutes	thinking thinking thinking	junior high high school	planned back realized going
9	oh ok	like really	feel like better	feel like time	party party class party	decided plan something guess
10	kind music	think people	feel like home	really need go	people people mom mom	mean like home house

of overlapping points). For each cloud of points a trigram belonging to the group is shown to give an idea of the collocations that belong to that cloud. Moreover, the visualised patterns (i.e., clouds of points) reveal differences in the semantic space between trigrams discovered and selected for EXT and INT. We observe a similar behaviour in the case of bigrams and quadgrams.

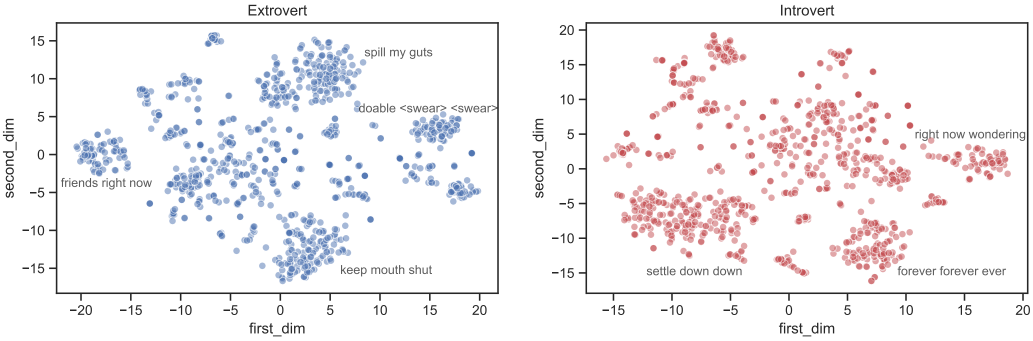


Figure 6.3. Trigrams BERT embeddings visualisation using t-SNE.

6.4 Discussion and Summary

Achieving a better understanding of human personality is becoming increasingly important in our current society. Conversational agents cannot achieve natural and effective conversations by asking naive questions that simply prolong conversations. However, they can incorporate an understanding of various aspects of a dialog with a human, like personality identification, toward providing improved dialogues and interactions. Such technologies will open new avenues to building more empathetic and naturalistic systems. This is particularly beneficial on the Web, where conversational agents (i.e., chat-bots) are increasingly communicating with humans to assist them through every day shopping, hotel reservations, and various other services.

In this chapter, we introduced a novel approach to personality assessment in conversations based on capsule neural networks. To this end, we tackled this problem as a frequency co-variance between different sets of words modelled as capsules. Our experimental evaluation on a real-world dataset of conversations showed that our model outperformed state-of-the-art baselines in personality recognition from text. An overall improvement of more than 10% was achieved by our model in terms of F_1 as compared with baselines. The highest performance is achieved for conscientiousness, replicating previous findings [130]. In addition, we presented a qualitative analysis of the self-attention weights learned by our model providing insights on the words and phrases the network focuses when predicting personality traits from textual records, in particular conversations.

Finally, we introduced a novel open-vocabulary approach based on multiword expressions which revealed new insights on the relationship between personality and language. In particular, we studied the collocation patterns that emerged from both spoken and written sources. We compared the sets discovered and define a collocation discrimination scoring function to produce a ranking of the most distinctive collocations that characterise extroversion and its opposite dimension, introversion. In particular, we observed that that as the number of words in the collocations increases, Jaccard's index between extroversion and introversion sets decreases. This suggests that as more words comprised the collocation, more distinctive patterns are captured since the sets discovered for each trait become increasingly diverse. Furthermore, we visually studied the top-1000 trigrams produced by our method by plotting the BERT-embeddings for each trait. The patterns observed (i.e., clouds of points) revealed differences in the semantic space between trigrams discovered and selected for extroversion and introversion.

Part III

Conclusions and Future Work

Chapter 7

Conclusions

7.1 Summary

Throughout this dissertation we focused on language analysis from two perspectives: *prediction* and *insight*. In particular, we analysed textual records to study two tightly related variables which encompass equally important and integral components of individuals' psychological profiles: mental health state and personality. Below, we revisit the research contributions and findings outlined in each part of this thesis.

Motivated by the ongoing digital transformation in which the proliferation of social platforms have taken a leading role, we addressed the mental health state assessment of individuals by studying online textual digital records. In Part I we focused on developing text mining models which are able to effectively extract clues useful to assess the psychological states from online user-generated content. In addition, we studied how such models could be used to spot in advance variations of individuals' mental state suggesting the onset of a disorder. To this aim, in Chapter 4.2 we use LSA to compute the similarity of a set of words with topical relevance to depression considering every word in the users' chronology of posts and derive various semantic proximity features [200]. Outcomes from our temporal spread of the cues analysis spotlighted that performance could be boost if the assessment algorithm's decision threshold is defined on a user-dependent basis, thus capturing the very subjective behaviour of each user. For instance, users with similar characteristics could be grouped in order to create different *stereotypes* or *profiles*. Our experiments revealed that as model's threshold becomes more conservative and stringent, decisions are taken in the latter chunks. This delay is highly penalised by ERDE and highlights the trade-off between taking *early* decisions at the risk of making more mistakes or waiting to receive more data to take more informed decisions.

In Chapter 4.3 we introduced a methodology to automatically gather post samples of depression by taking advantage of weak-supervision signals [197]. We empirically validated our methodology and showed it can be effectively used to derive large samples of data for the study of depression on online social media settings. As a result, design of data-driven solutions to this problem becomes feasible. In addition, we released the dataset created following the proposed methodology and developed a series of depression post-classifiers to serve as a benchmark [199]. We showcased the potential of post-depression classifiers in identifying latent depression patterns via a case study using time-series analysis.

Even though achieving an effective positive detection performance is important, we consider that tracking and visualising the development of the mental disorder is equally relevant. Gaining insights on the language and online behaviour of individuals affected by mental disorders could lead to identify new predictive markers up to now not accounted in the medical literature; thus, motivating new inquiries into behavioural traits of mental health disorders as observed on social media. As argued by Coppersmith et al. [44] this also poses some interesting questions such as: *"What can we expect to learn about mental health by studying social media?"*. In Chapter 5 we presented a series of analytical studies which aimed at gaining novel insights and extend the current knowledge on the manifestation of mental disorders on online settings.

Outcomes from our thorough analysis showed that psychometric attributes, emotional expression and social engagement provide a quantifiable and significant way to differentiate individuals affected by mental disorders from healthy ones [196]. Across different mental disorders, however, we could not find any significant indicators to be able to distinguish one from the other. In addition, we obtained evidence suggesting that the use of language in micro-blogging platforms, such as Twitter, where users are subjected to constraints influencing their writing style, is less distinguishable between users suffering a mental disorder than other less restrictive platforms, like Reddit. From this study we could conclude that by extracting this information from social media users' posts practitioners could perform a more comprehensive and large-scale assessment of the individuals.

Finally, we close the chapter with a detailed analysis toward improving the understanding of how depression assessment inventories, in particular the BDI, could be automatically estimated based on the evidence available on social media. To this aim, we investigated the relative incidence (i.e., the amount of evidence about each item that can be retrieved from a collection of social media posts) of the 21 BDI items on users' feeds. To measure the incidence of different depression symptoms, we derived queries from the BDI items, retrieved documents from social media collections (Twitter and Reddit) and analysed their relevance scores. Our experiments revealed that certain items yield a consistent incidence (over different search methods and social media platforms), suggesting that the associated symptoms are more prevalent on these collections. Comparing the positive groups of Twitter and Reddit, we observe that, on average, the documents retrieved from Twitter have a higher incidence score than those from Reddit (considered a “very large” effect size). Interestingly, Reddit positive and controls groups showed very small effect sizes and differences are not significant in most cases. Conversely, for Twitter groups effect sizes were larger. Twitter control group presented, on average, a slightly larger incidence score when compared to the positive group. Complementary, we analysed how incidence and other features influence the effectiveness of automatic tools that extract depression symptoms from social media posts. We discovered that systems that attempt to infer the severity of depression symptoms from user data do not fair well on high incidence BDI items. In fact, low incidence items turned out to be the parts of the BDI questionnaire whose answers are easier to estimate using automated means. Moreover, the effectiveness of the automatic systems is also inversely correlated with other features of the BDI items, such as the length (i.e, items whose query representation has fewer words).

Given the emergence of powerful technological platforms supporting personalisation and its expression, achieving a better understanding of human personality is becoming increasingly important in our current society. In particular, this highlights the need for developing conversational agents which can achieve natural and effective conversations by incorporating an understanding of various aspects of a dialog with a human, like personality identification. Inspired by this, in Part II we focused on advancing the state of the art on personality assessment from conversations. Similarly to the first part of the thesis, we also aim at gaining new insights, in this case, on personality expression on conversations.

We started Chapter 6 by presenting a novel approach to personality recognition in conversations based on capsule neural networks. To this aim, we tackled this task as a frequency co-variance between different sets of words modelled as capsules [198]. Our model yielded an overall improvement of more than 10% in terms of F_1 as compared with state-of-the-art baselines in personality recognition from text. The highest effectiveness was achieved when predicting conscientiousness. A qualitative analysis of the self-attention weights learned by our model allowed us to leveraged its inherent interpretability potential to gain insights on the words and phrases the network focuses when predicting personality traits from conversations.

We concluded the chapter by introducing novel open-vocabulary approach based on multiword expressions which revealed new insights on personality and language. In particular, we discover a set of candidates and define a discrimination scoring function to select meaningful expressions that emerge from both spoken and written sources and that are intended to capture the most distinctive linguistic patterns of a personality trait. To put into practise our approach we focus on extroversion and its corresponding complementary dimension, introversion, although, the method and analyses could be easily extended to the remaining traits of the Big Five model. We observed that as the number of words in the expressions increases, Jaccard's index between extroversion and introversion sets decreases. This suggests that as more words comprised the collocation, more distinctive patterns are captured since the sets discovered for each trait become increasingly diverse. Moreover, as conversations are comprised of utterances which are usually much shorter (in number of words) in comparison with essays, we noted that the number of candidates collocations discovered for the essays is several orders of magnitude larger than for conversations. A close inspection to the top-100 ranked collocations obtained with our method for each trait revealed that several collocations found for extroversion are correlated with findings confirmed by previous research, like the inclusion of words with social connotations or related to outdoor activities as well as sentimental expression. Moreover, we discover co-occurrences of words unveiling novel patterns. For extroversion we observe a consistent use of swearing words, and the habit of using capitals. In the case of introversion, on the other hand, we found expressions associated with self-assessment and introspection.

7.2 Ethical Concerns and Discussion

Research involving human beings concerns sensitive topics related to the ethics of the treatment of data and individual's privacy. In general, Computer Science studies do not perform experiments in the same way that, for instance, medicine or psychology does (i.e., direct experimentation with patients). Most of the studies conducted in this area collect and store information that might be associated with individuals. Such information may inconvenience or even threaten the physical and psychological well-being of the human subjects involved as it might be employed for unforeseen purposes or be shared with unintended and unethical recipients.

In the context of OMSA, the use of social media data raises two major ethical concerns [167]: (a) users' data, even if public, are employed in ways the users may not have intended, and (b) health data is particularly sensitive information. The usual response to these concerns is that these data are public, and users agree to share their public data under the terms of service of social media platforms. As long as the data are public and there is no interaction with individuals, social data research can generally be conducted following the guidelines of Institutional Review Boards (IRBs) or equivalent ethics boards (if applicable). Nonetheless, users might not be aware that their data are public, or might not want their public data to be used for such purposes. Social media platforms' terms of service are often difficult to understand [61; 124; 191] and various social platforms have been criticised for having unclear privacy management systems [117]. Bearing this in mind, the boundaries between public and private data might be confusing. Furthermore, there are countries like the United States where the use of publicly available data (e.g., tweets) might not meet the criteria of research involving human subjects according to the Code of Federal Regulations[1] [25]. In this respect, there is a great deal of disagreement among universities' IRBs when it comes to social computing research. Data collected from online sources, such as Reddit or Twitter, in most cases do not constitute research that requires their purview or informed consent practices [25]. The majority of researchers who deal with datasets from social media platforms do not gain consent from individual users whose posts are collected, nor are those users usually informed by the researcher. To shed light on this matter, Fiesler et al. [62] conducted an exploratory survey of the users' perceptions of the use of tweets in academic research. In particular, they probed a series of contextual factors (e.g, how the research is conducted or disseminated, who the researchers are, what the study is about) that impact on whether Twitter users found analysing their content acceptable. They discovered that a few users were previously aware of the fact that their public posts could be used by researchers, and most of them felt that researchers should not be able to use tweets without their consent. However, such stances are highly contextual.

Ethical issues involving social media research have been addressed in various studies, for example McKee [138] and Conway [42] have presented surveys on the topic. Additionally, Benton et al. [16] complemented these surveys by providing several practical suggestions organised as a set of guidelines targeted to researchers who deal with social media data in health-related areas. In particular, these recommendations address topics related to research on human subjects, which actually encompass research with data from living individuals. The differences between the custom procedures and ethics followed in each area pose new challenges for the development of research protocols. Furthermore, the back-

[1] See `https://bit.ly/3AeA2iy`

ground of the researchers conducting these studies might contribute to make these challenges even more complex. Typically, research in the health domain is conducted by individuals who have been trained in medical fields, and therefore are assumed to have an understanding of human subject research protocols and concerns related to IRBs. Social media research instead, might be conducted by computer scientists, who are not so familiar with such guidelines.

Researchers conducting OMSA and, thus, working with social media data must take the necessary precautions to protect the privacy of individuals and their ethical rights to avoid further psychological distress when sharing the data. For instance, the use of whitelist approaches for anonymising data are usually effective. Although screen names or URLs can be anonymised using hash functions, the possibility of cross-referencing the text against the social platforms archives still exists, and it could lead to breach of user privacy. For this reason, often researchers are asked to sign a confidentiality agreement so as to guarantee the privacy of the data.

Another important issue refers to the practical use of these technologies (e.g., to define interventions). As argued by De Choudhury [50], design considerations in this space need to ensure that the benefits obtained by intervening exceed the risks. This can be achieved by exposing risks to either the individuals or to trusted social contacts or clinicians. Some social media sites offer basic intervention services to support vulnerable people. For instance, Facebook counts with a suicide intervention tool which encourages individuals to contact a close friend or an assistance hotline when other uses state their publications shows suicidal tendencies.

Finally, we would like to conclude the current discussion on the ethical implications of the work conducted in this thesis by highlighting the tensions and implications that the development and use of empathetic and naturalistic conversational systems pose on the matter. As stated in the introduction, we emphasise the need for conversational agents which are able to *detect* and *mimic* the personality style of a user toward more engaging and effective dialogues.

Language is an integral component of human identity and considered one of the cornerstone of human life in society. Conversational agents are thus naturally compared to human beings, whether or not the individuals are aware of their artificial nature. Humans tend to anthropomorphise machines [111]. This kind of behaviour is exacerbated when users can interact conversationally with a system and especially if the system has been embodied with a personality [206]. Such scenario could have psychological or legal consequences, or give rise to varying degrees of manipulation. Moreover, the use of conversational agents in sensitive domains, such as healthcare (e.g. providing treatment and counselling services [113; 232]), creates new ethical tensions like the impossibility of explaining in natural language the chain of decisions leading to a particular medical recommendation or the potential risk of bias [126]. Although conversational agents are becoming ubiquitous in our daily lives, their large-scale deployment is still in its infancy, and therefore there is still not enough experimental data to assess their long-term effects on human beings. Several ethical questions and concerns are still open for further investigation.

7.3 Future Research Directions

This thesis has resulted in several findings and lessons learnt on identifying and analysing mental disorders on online social media; in particular, toward improving the understanding of how such disorders eventually are manifested through language use. Further, we have studied the computational assessment of personality in conversations providing novel techniques and perspectives. The trajectory of the thesis has stemmed specific future research directions on different topics. Below, we describe some of the possible future directions.

Algorithmic Fairness: As ML becomes more pervasive in sensitive domains, such as healthcare, special care should be paid to a recent issue that has drawn scholars attention: *algorithmic bias*. As it happens with people, algorithms are vulnerable to biases that render their decisions *unfair*. In the context of risk-assessment and decision-making systems, fairness is defined as the "*absence of any prejudice or favouritism toward an individual or a group based on their inherent or acquired characteristics*" [143]. Hence, an algorithm whose decisions are skewed towards a particular group of people, often a certain minority, is considered *unfair*. For instance, a recent study [19] showed that certain word embeddings methods embodied implicit gender bias. One possible reason could reside in the data employed to train these algorithms. In general, the different learning methods attempt to recognise and leverage statistical patterns in the data. Therefore, if the ground truth data contain some bias or historical discrimination, the algorithm will likely incorporate it into its future predictions [162]. Since OMSA uses human-related data, an interesting avenue to explore would be to analyse the extent to which bias is affecting the different algorithms when they are applied to predict the onset of a mental disorder. It is important to understand whether training data are representative of the population. For instance, the strategy developed by Coppersmith et al. [43] enables to automatically collect considerable large amounts of data on the basis of self-disclosed diagnoses. Yet, that might not really represent all the attributes of the population as a whole.

Signals Crossover: As shown in Chapter 5 we could not find any significant indicators across different mental disorders to be able to distinguish one from the other, suggesting that in most of the cases determining the exact disorder is a more difficult task and still remains an open question. Different disorders are manifested through users' writing with different features. An interesting lesson to take here and which opens new avenues for further research is that, overall, designing and implementing risk-assessment and decision-making technologies that can equally perform for different detection tasks is still a complex endeavour. While there might be overlapping features associated with the identification of different disorders, such as self-harm and depression, the way individuals express has subtle variations that depend on the disorder under analysis. Although these nuances in the data allow systems to effectively discover positive cases, they prevent them from being applicable to detect other disorders than the one they were meant to identify. Furthermore, it could happen that an individual is suffering from more than one disorder at the same time, despite being diagnosed with one. This situation is commonly known as *comorbidity*. For instance, according to the DSM, a Major Depressive Disorder is a very common comorbid disorder, and it frequently coexists with other disorders, such as PTSD. This issue highlights the fact that computational tools should not be expected to serve as standalone diagnostic instruments but as unobtrusive mechanisms for early identification of potential mental problems.

Alternatives to Supervised Learning: As shown in Chapter 2.1, most of the approaches proposed in the literature use some form of supervised learning. These algorithms heavily depend on the amount and quality of training data available. Collecting large amounts of labelled data is usually a complex and time-consuming endeavour and in the case of deep learning architectures, this might even prevent their application as they usually require several thousands of examples to converge. Hence, it would be interesting to further explore how *unsupervised learning* algorithms can be applied in the context of OMSA. There are some initial attempts toward this direction with few works proposing different rules-based approaches [169; 45]. Also, the application of *transfer learning and domain adaptation* could contribute to mitigate the lack of massive sets of hand-labelled training data.

Effectiveness vs. Delay Trade-off: In chapter 4 we presented an early risk-assessment system based on the concept of semantic proximity. The different experiments we performed have shown the effect of the decision threshold on its effectiveness and the temporal spread of the cues that indicate the onset of depression. As any early risk-assessment systems in a real-life setting, one would have to decide which metric should be the focus when optimising the various parameters of a specific model. This decision could be taken with the assistance or advice of professionals in the health domain. For instance, under certain circumstances, producing a considerable amount of false alarms (*false positives*) can be tolerated at the benefit of discovering most of the *real* cases. This could help prevent further consequences (e.g, suicide). Therefore, it becomes essential to understand the practical implications of using such early risk-assessment systems in the clinical practice. For instance, one could consider the difference that the two evaluation measures of ERDE_5 and ERDE_{50} could have in a real-life scenario. A mental health agency might decide to set penalty costs on the consequences of late detection. This leads to a natural trade-off depending on whether the goal is to be more conservative and raise a handful of timely alerts or to maximise the recall at the cost of issuing false alarms. Early risk detection is challenging because multiple objectives are involved (ERDE, Precision, Recall, Latency). Thus, the exploration of the *most suitable* trade-off for each situation remains still an open question. Moreover, systems developed for this purpose in most of the cases incorporated some form of offline processing. This is an interesting point to consider given that in a real-life scenario time is also an essential variable. It would also pose new challenges to think how different early risk-assessment systems would perform or should be designed to incorporate real-time processing.

Multidisciplinary Research: By definition, OMSA involves many different areas across Computer Science, Social Science and Medicine. In general, research initiatives in each field have taken separate avenues and, because of this, the full potential of research on OMSA is still unexplored. Interdisciplinary work should be further promoted so as to improve the understanding of the problem and subsequently design more effective screening methods. A joint collaboration between the cited areas could foster the development of better methods that combine medical foundations with the power of Computer Science. Furthermore, these collaborative efforts could provide better explanations of the computational models used for each task.

Multimodality: In this thesis we have focused on the computational analysis of textual records, and therefore the language use, to conduct OMSA on individuals. The increasing popularity of image-and-video sharing social media platforms, such as Instagram, TikTok and YouTube, suggest that different modalities of information (e.g., text, images, audio, video) could be combined to extract complementary signals enhancing the effectiveness and understanding of the task [71]. Multimodal OMSA is an emerging field at the intersection of NLP, IR, image processing and speech processing. Mood traces, such as facial and prosodic expressions can be captured from videos, and be leveraged in addition to the textual content. In this respect, there have been initial attempts by combining posts' textual content with users' profile avatars [215] and posted images [186]. It should be noted that the availability of data to conduct any form of multimodal analysis represents a challenging endeavour; mainly because collecting data from the same users over different platforms, though feasible, is not a trivial task.

Personality-conditioned Language Generation: As stated by Zhang et al. [247] communication between a human and a machine is still in its infancy and one of its common issues includes the lack of a consistent personality in the agent. In this thesis, we believe we improved the state of the art of automatic personality recognition in conversations. Once the conversational agent has inferred the personality of its human interlocutor, it would be desirable to adapt its way of expressing to become more "engaging"[2] and more "human"[3] [214]. In Chapter 6.3 we have presented an open-vocabulary approach to discover and select a meaningful set of collocations which are intended to capture the most distinctive linguistic patterns of a personality trait. As collocations are important for various applications, such as natural language generation (e.g., to make sure that the output sounds natural) and corpus linguistic research [133], an interesting avenue for further research is to investigate how the most meaningful collocations that were discovered and selected with our method could be used for personality-conditioned natural language generation. For instance, one possible way could be to retrieve the original pieces of text in which such collocations are used to create a collection. This can be later used to fine-tune recently developed large-scale unsupervised language models which are able to generate coherent synthetic text samples, such as GPT-2 [182]. Afterwards, the quality of the text generated along the Big Five traits could be evaluated by human assessors. Moreover, the focus could be on both generating personality-conditioned language and assessing whether this can affect users' perception in a conversation and how.

[2] *Engagingness* is concern with to the general question: "How much did you enjoy talking to this user?"

[3] *Humanness* is concern with to the general question: "Do you think this user is a bot or a human?"

Bibliography

[1] S. Abdullah, M. Matthews, E. Frank, G. J. Doherty, G. Gay, and T. Choudhury. Automatic detection of social rhythms in bipolar disorder. *J. Am. Medical Informatics Assoc.*, 23(3):538–543, 2016.

[2] L. Achilles, M. Kisselew, J. Schäfer, and R. Kölle. Using surface and semantic features for detecting early signs of self-harm in social media postings. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[3] E. C. Ageitos, J. Martínez-Romo, and L. Araujo. NLP-UNED at erisk 2020: Self-harm early risk detection with sentiment analysis and linguistic features. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[4] M. Al-Mosaiwi and T. Johnstone. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542, 2018.

[5] M. Aliannejadi and F. Crestani. Venue suggestion using social-centric scores. In *Proccedings of ECIR Workshop on Social Aspects in Personalization and Search*, 2018.

[6] M. Aliannejadi, M. Harvey, L. Costa, M. Pointon, and F. Crestani. Understanding mobile search task relevance and user behaviour in context. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10-14, 2019*, pages 143–151, 2019.

[7] H. Almeida, A. Briand, and M. Meurs. Detecting early risk of depression from social media user-generated content. In *Working Notes of CLEF 2017*

- Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017., 2017.

[8] A. P. Association. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Publishing, Washington, 5th edition, 2013.

[9] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20*, pages 3256–3274, 2020.

[10] S. A. Bahrainian and F. Crestani. Towards the next generation of personal assistants: Systems that know when you forget. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands*, pages 169–176, 2017.

[11] S. A. Bahrainian, M. Liwicki, and A. Dengel. Fuzzy subjective sentiment phrases: A context sensitive and self-maintaining sentiment lexicon. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland*, pages 361–368, 2014.

[12] A. T. Beck, R. A. Steer, and M. G. Carbin. Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8(1):77 – 100, 1988.

[13] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6):561–571, 1961.

[14] G. Bedi, G. A. Cecchi, D. F. Slezak, F. Carrillo, M. Sigman, and H. de Wit. A window into the intoxicated mind? Speech as an index of psychoactive drug effects. *Neuropsychopharmacology*, 39(10), 2014.

[15] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, and A. T. Campbell. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal*, 38(3):218–226, Sep 2015.

[16] A. Benton, G. Coppersmith, and M. Dredze. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL, Valencia, Spain, April 4, 2017*, pages 94–102, 2017.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[19] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357, 2016.

[20] R. L. Boyd and J. W. Pennebaker. Language-based personality: a new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, 18:63 – 68, 2017. Big data in the behavioural sciences.

[21] R. L. Boyd and H. A. Schwartz. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41, Jan 2021.

[22] R. L. Boyd, S. R. Wilson, J. W. Pennebaker, M. Kosinski, D. J. Stillwell, and R. Mihalcea. Values in words: Using language to evaluate and understand personal values. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, Oxford, UK*, pages 31–40, 2015.

[23] C. R. Brewin, S. Rose, B. Andrews, J. Green, P. Tata, C. McEvedy, S. Turner, and E. B. Foa. Brief screening instrument for post-traumatic stress disorder. *British Journal of Psychiatry*, 181(2):158–162, 2002.

[24] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff. Digital disease detection - harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157, 2009.

[25] A. Bruckman. *Research Ethics and HCI*, pages 449–468. Springer New York, New York, NY, 2014.

[26] S. G. Burdisso, M. Errecalde, and M. M. y Gómez. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197, 2019.

[27] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro. Early detection of depression: Social network analysis and random forest techniques. *J Med Internet Res*, 21(6):e12554, Jun 2019.

[28] F. Cacheda, D. F. Iglesias, F. J. Nóvoa, and V. Carneiro. Analysis and experiments on early detection of depression. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.

[29] L. Canzian and M. Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015, Osaka, Japan, September 7-11, 2015*, pages 1293–1304, 2015.

[30] J. Cassell and T. Bickmore. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13(1-2):89–132, 2003.

[31] P. A. Cavazos-Rehg, M. J. Krauss, S. Sowles, S. Connolly, C. Rosas, M. Bharadwaj, and L. J. Bierut. A content analysis of depression-related tweets. *Computers in Human Behavior*, 54:351–357, 2016.

[32] F. Celli and B. Lepri. Is big five better than mbti? A personality computing challenge using twitter data. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy*, 2018.

[33] S. Chancellor and M. D. Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Medicine*, 3(1), 2020.

[34] S. Chancellor, Y. Kalantidis, J. A. Pater, M. De Choudhury, and D. A. Shamma. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3213–3226, New York, NY, USA, 2017. ACM.

[35] P. G. F. Cheng, R. M. Ramos, J. Á. Bitsch, S. M. Jonas, T. Ix, P. L. Q. See, and K. Wehrle. Psychologist in a pocket: Lexicon development and content validation of a mobile-based app for depression screening. *JMIR MHealth UHealth*, 4(3), 2016.

[36] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734, 2014.

[37] J. Choi and W. B. Croft. Temporal models for microblogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 2491–2494, 2012.

[38] M. D. Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, USA*, 2013.

[39] C. Chung and J. Pennebaker. The psychological functions of function words. *Frontiers of social psychology. Social communication*, 2007.

[40] K. W. Church and R. L. Mercer. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24, 1993.

[41] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In E. M. Bender, L. Derczynski, and P. Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, USA, August 20-26, 2018*, pages 1485–1497. Association for Computational Linguistics, 2018.

[42] M. Conway. Ethical issues in using twitter for public health surveillance and research: Developing a taxonomy of ethical concepts from the research literature. *J Med Internet Res*, 16(12):e290, Dec 2014.

[43] G. Coppersmith, M. Dredze, and C. Harman. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Baltimore, USA, 2014.

[44] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational*

Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Association for Computational Linguistics, 2015.

[45] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell. Clpsych 2015 shared task: Depression and PTSD on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, USA*, 2015.

[46] G. Coppersmith, C. Harman, and M. Dredze. Measuring post traumatic stress disorder in twitter. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 2014.

[47] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition, 2009.

[48] J. S. Culpepper, F. Diaz, and M. D. Smucker. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). *SIGIR Forum*, 52(1):34–90, Aug. 2018.

[49] T. Davies. Abc of mental health. mental health assessment. *BMJ*, 314(7093):1536–1539, May 1997.

[50] M. De Choudhury. Anorexia on tumblr: A characterization study. In *Proceedings of the 5th International Conference on Digital Health 2015, DH '15*, pages 43–50, Florence, Italy, 2015.

[51] M. De Choudhury, S. Counts, and E. Horvitz. Major life changes and behavioral markers in social media: Case of childbirth. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, New York, NY, USA, 2013. ACM.

[52] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 2098–2110, 2016.

[53] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 1990.

[54] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[55] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLOS ONE*, 2011.

[56] M. Donnellan, R. D. Conger, and C. M. Bryant. The big five and enduring marriages. *Journal of Research in Personality*, 38(5):481 – 504, 2004.

[57] J. C. Eichstaedt and A. C. Weidman. Tracking fluctuations in psychological states using social media language: A case study of weekly emotion. *European Journal of Personality*, 34(5):845–858, 2020.

[58] B. Elvevåg, P. W. Foltz, D. R. Weinberger, and T. E. Goldberg. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93(1), 2007.

[59] A. A. Farias-Anzaldua, M. Montes-y-Gómez, A. P. López-Monroy, and L. C. González-Gurrola. UACH-INAOE participation at erisk2017. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.

[60] E. Fast, B. Chen, and M. S. Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4647–4657, 2016.

[61] C. Fiesler, C. Lampe, and A. S. Bruckman. Reality and perception of copyright terms of service for online content creation. CSCW '16, page 1450–1461, 2016.

[62] C. Fiesler and N. Proferes. "participant" perceptions of twitter research ethics. *Social Media + Society*, 4(1), 2018.

[63] N. Fuhr. Some common mistakes in ir evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, Feb. 2018.

[64] D. C. Funder and C. D. Sneed. Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology*, 64(3):479–490, 1993.

[65] D. G. Funez, M. J. G. Ucelay, M. P. Villegas, S. Burdisso, L. C. Cagnina, M. Montes-y-Gómez, and M. Errecalde. Unsl's participation at erisk 2018 lab. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.

[66] M. Gaur, U. Kursuncu, A. Alambo, A. Sheth, R. Daniulaityte, K. Thirunarayan, and J. Pathak. "let me tell you about your mental health!": Contextualized classification of reddit posts to dsm-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 753–762, 2018.

[67] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, pages 291–304, 2007.

[68] A. J. Gill and J. Oberlander. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368, 2002.

[69] M. Gjurković and J. Šnajder. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97, June 2018.

[70] G. Gkotsis, A. Oellrich, T. J. Hubbard, R. J. Dobson, M. Liakata, S. Velupillai, and R. Dutta. The language of mental health problems in social media. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2016.

[71] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. Song. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197, 2021.

[72] K. Gligoric, A. Anderson, and R. West. How constraints affect content: The case of twitter's switch from 140 to 280 characters. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 596–599, 2018.

[73] K. Gligorić, A. Anderson, and R. West. Causal effects of brevity on style and success in social media. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), Nov. 2019.

[74] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011*, pages 149–156. IEEE Computer Society, 2011.

[75] B. E. Goldstein, editor. *Cognitive psychology: Connecting mind, research and everyday experience*. New york: Cengage learning, 2015.

[76] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, 2016.

[77] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

[78] W. G. Graziano and N. Eisenberg. Chapter 30 - agreeableness: A dimension of personality. In R. Hogan, J. Johnson, and S. Briggs, editors, *Handbook of Personality Psychology*, pages 795–824. Academic Press, San Diego, 1997.

[79] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 2004.

[80] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Ohler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J. Biomed. Health Informatics*, 19(1):140–148, 2015.

[81] A. Grünerbl, V. Osmani, G. Bahle, J. C. Carrasco, S. Oehler, O. Mayora, C. Haring, and P. Lukowicz. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *Proceedings of the 5th Augmented Human International Conference*, AH '14, 2014.

[82] Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 417–420, 2007.

[83] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt. Detecting depression and mental illness on social media: an integrative

review. *Current Opinion in Behavioral Sciences*, 18(Supplement C):43 – 49, 2017. SI: 18: Big data in the behavioural sciences (2017).

[84] N. Hayes and S. Joseph. Big 5 correlates of three measures of subjective well-being. *Personality and Individual Differences*, 34(4):723–727, 2003.

[85] D. R. Heine, Steven J.and Lehman, K. Peng, and J. Greenholtz. What's wrong with cross-cultural comparisons of subjective likert scales? the reference-group effect. *Journal of Personality and Social Psychology*, 82(6):903–918, 2002.

[86] G. Hertel, J. Schroer, B. Batinic, and S. Naumann. Do shy people prefer to send e-mail? personality effects on communication media preferences in threatening and nonthreatening situations. *Social Psychology*, 39(4):231–243, 2008.

[87] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[88] G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. In *Proceedings of the 21th International Conference on Artificial Neural Networks*, ICANN'11, pages 44–51, Espoo, Finland, 2011.

[89] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[90] D. J. Hughes, M. Rowe, M. Batey, and A. Lee. A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2):561–569, 2012.

[91] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh, editors, *ICWSM*. The AAAI Press, 2014.

[92] J. D. Hwang and K. Hollingshead. Crazy mad nutters: The language of mental health. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2016.

[93] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander. Large scale personality classification of bloggers. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, ACII'11, pages 568–577, Memphis, USA, 2011.

[94] J. U. Islam, Z. Rahman, and L. Hollebeek. Personality factors as predictors of online consumer engagement: an empirical investigation. *Marketing Intelligence Planning*, 35:510–528, 2017.

[95] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White. Monitoring tweets for depression to detect at-risk users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology - From Linguistic Signal to Clinical Reality, Vancouver, Canada*, 2017.

[96] O. P. John, A. Angleitner, and F. Ostendorf. The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, 2(3):171–203, 1988.

[97] O. P. John and S. Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research*, pages 102–138. Guilford Press, New York, USA, 1999.

[98] C. Jordan, C. Watters, and Q. Gao. Using controlled query generation to evaluate blind relevance feedback algorithms. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '06, page 286–295, 2006.

[99] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, 2017.

[100] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, USA, 1st edition, 2000.

[101] P. Kanerva, J. Kristoferson, and A. Holst. Random indexing of text samples for latent semantic analysis. In *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 103–6. Erlbaum, 2000.

[102] R. Katikalapudi, S. Chellappan, F. Montgomery, D. Wunsch, and K. Lutzen. Associating internet usage with depressive behavior among college students. *IEEE Technology and Society Magazine*, 31(4):73–80, 2012.

[103] G. Kazai, P. Thomas, and N. Craswell. The emotion profile of web search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Paris, France,* page 1097–1100, 2019.

[104] M. Khodabakhsh, H. Fani, F. Zarrinkalam, and E. Bagheri. Predicting personal life events from streaming social content. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1751–1754, 2018.

[105] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, USA, 2015.

[106] I. M. Kloumann, C. M. Danforth, K. D. Harris, C. A. Bliss, and P. S. Dodds. Positivity of the english language. *PLOS ONE*, 7(1):1–7, 01 2012.

[107] K. Konstabel, J.-E. Lönnqvist, G. Walkowitz, K. Konstabel, and M. Verkasalo. The "short five" (s5): Measuring personality traits using comprehensive single items. *European Journal of Personality*, 26(1):13–29, 2012.

[108] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[109] K. Kroenke, R. L. Spitzer, J. B. Williams, and B. Löwe. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *General Hospital Psychiatry*, 32(4):345 – 359, 2010.

[110] I. Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, 47(4):2025–2047, Jun 2013.

[111] B. Kuipers, J. McCarthy, and J. Weizenbaum. Computer power and human reason. *SIGART Bull.*, (58):4–13, June 1976.

[112] A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on twitter. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 789–795, 2013.

[113] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association : JAMIA*, 25(9):1248–1258, Sep 2018.

[114] J. L. Lastovicka and E. A. Joachimsthaler. Improving the detection of personality-behavior relationships in consumer research. *Journal of Consumer Research*, 14(4):583–587, 03 1988.

[115] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2nd edition, 2014.

[116] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017*, Toulon, France, 2017.

[117] Y. Liu, Q. Mei, D. A. Hanauer, K. Zheng, and J. M. Lee. Use of social media in the diabetes community: An exploratory analysis of diabetes-related tweets. *JMIR Diabetes*, 1(2):e4, Nov 2016.

[118] W.-Y. Loh. Classification and regression trees. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 1:14–23, 2011.

[119] D. E. Losada and F. Crestani. A test collection for research on depression and language use. In *Conference Labs of the Evaluation Forum*, pages 28–39. Springer, 2016.

[120] D. E. Losada, F. Crestani, and J. Parapar. CLEF 2017 erisk overview: Early risk prediction on the internet: Experimental foundations. In *Conference and Labs of the Evaluation Forum*, 2017.

[121] D. E. Losada, F. Crestani, and J. Parapar. Overview of erisk early risk prediction on the internet. In *Conference and Labs of the Evaluation Forum*. CEUR-WS.org, 2018.

[122] D. E. Losada, F. Crestani, and J. Parapar. Overview of erisk 2019 early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 340–357, 2019.

[123] D. E. Losada, F. Crestani, and J. Parapar. Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece*, pages 272–287, 2020.

[124] E. Luger, S. Moran, and T. Rodden. *Consent for All: Revealing the Hidden Complexity of Terms and Conditions*, page 2687–2696. 2013.

[125] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[126] D. D. Luxton. Ethical implications of conversational agents in global public health. *Bulletin of the World Health Organization*, 98(4):285 – 287, 2020-4-01.

[127] K. Luyckx and W. Daelemans. Personae: a corpus for author and personality prediction from text. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco, 2008.

[128] X. Ma, E. Yang, and P. Fung. Exploring perceived emotional intelligence of personality-driven virtual agents in handling user challenges. In *The World Wide Web Conference*, WWW '19, page 1222–1233, 2019.

[129] C. Maigrot, S. Bringay, and J. Azé. Concept drift vs suicide: How one can help prevent the other? *International Journal of Computational Linguistics and Applications*, 8(1), Oct 2017.

[130] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.*, 30(1):457–500, Nov. 2007.

[131] I. A. Malam, M. Arziki, M. N. Bellazrak, F. Benamara, A. E. Kaidi, B. Es-Saghir, Z. He, M. Housni, V. Moriceau, J. Mothe, and F. Ramiandrisoa. IRIT at e-risk. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.

[132] L. Manikonda, G. Beigi, H. Liu, and S. Kambhampati. Twitter for sparking a movement, reddit for sharing the moment: #metoo through the lens of social media. *CoRR*, abs/1803.08022, 2018.

[133] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[134] D. C. Martin. The mental status examination. In H. J. Walker HK, Hall WD, editor, *Clinical Methods: The History, Physical, and Laboratory Examinations*, chapter 207, pages 924–924. Boston: Butterworths, 3rd edition, 1990.

[135] R. Masood. Adapting models for the case of early risk prediction on the internet. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, editors, *Advances in Information Retrieval*, pages 353–358. Springer International Publishing, 2019.

[136] D. Maupomé and M. Meurs. Using topic extraction on social media content for the early detection of depression. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.

[137] R. R. McCrae and O. P. John. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215, 1992.

[138] R. McKee. Ethical issues in using social media for health and health care research. *Health Policy*, 110(2):298 – 301, 2013.

[139] A. R. McLarney-Vesotski, F. Bernieri, and D. Rempala. Personality perception: A developmental study. *Journal of Research in Personality*, 40(5):652 – 674, 2006.

[140] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(5):862–877, 2006.

[141] M. R. Mehl and J. W. Pennebaker. The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, 84(4):857–870, 2003.

[142] M. R. Mehl, J. W. Pennebaker, D. M. Crow, J. Dabbs, and J. H. Price. The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, 33(4):517–523, Nov 2001.

[143] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[144] D. Metzler, C. Cai, and E. Hovy. Structured event retrieval over microblog archives. NAACL HLT '12, page 646–655, 2012.

[145] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, Lake Tahoe, USA, 2013.

[146] A. Milton and M. S. Pera. What snippets feel: Depression, search, and snippets. In *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Samatan, France*, 2020.

[147] K. S. Minor, K. A. Bonfils, L. Luther, R. L. Firmin, M. Kukla, V. R. MacLain, B. Buck, P. H. Lysaker, and M. P. Salyers. Lexical analysis in schizophrenia: How emotion and social word use informs our understanding of clinical presentation. *Journal of Psychiatric Research*, 64(Supplement C), 2015.

[148] S. A. Mirlohi Falavarjani, J. Jovanovic, H. Fani, A. A. Ghorbani, Z. Noorian, and E. Bagheri. On the causal relation between real world activities and emotional expressions of social media users. *Journal of the Association for Information Science and Technology*, 72(6):723–743, 2021.

[149] W. Mischel, Y. Shoda, and O. Ayduk. *Introduction to Personality: Toward an Integrative Science of the Person*. Wiley, 8th edition, 2007.

[150] S. Mohammad. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan*, 2018.

[151] S. Mohammad and P. D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, pages 436–465, 2013.

[152] S. M. Mohammad and S. Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Comput. Intell.*, 31(2):301–326, May 2015.

[153] M. A. Moreno, L. A. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, and T. Becker. Feeling bad on facebook: depression disclosures by college students on a social networking site. *Depression and Anxiety*, pages 447–455, 2011.

[154] C. M. Morrison and H. Gore. The relationship between excessive internet use and depression: A questionnaire-based study of 1,319 young people and adults. *Psychopathology*, 43(2):121–126, 2010.

[155] I. B. Myers and P. B. Myers. *Gifts differing: Understanding personality type*. Davies-Black Publishing, 2010.

[156] S. Nepomnyachiy, B. Gelley, W. Jiang, and T. Minkus. What, where, and when: Keyword search with spatio-temporal ranges. GIR '14, 2014.

[157] Y. Neuman. *Computational Personality Analysis: introduction, practical applications and novel directions*. Springer, 2016.

[158] Y. Neuman and Y. Cohen. A vectorial semantics approach to personality assessment. *Scientific Reports*, 4(1), Apr 2014.

[159] Y. Neuman, Y. Cohen, D. Assaf, and G. Kedma. Proactive screening for depression through metaphorical and automatic text analysis. *Artif. Intell. Med.*, 56(1):19–25, 2012.

[160] M. Obuchi, J. F. Huckins, W. Wang, A. daSilva, C. Rogers, E. Murphy, E. Hedlund, P. Holtzheimer, S. Mirjafari, and A. Campbell. Predicting brain functional connectivity using mobile sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(1), Mar. 2020.

[161] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183 – 188, 2015.

[162] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

[163] W. H. Organization. *International Classification of Diseases (ICD-10)*. 1990.

[164] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, pages 1–135, 2008.

[165] M. Park, C. Cha, and M. Cha. Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop On Healthcare Informatics*, 2012.

[166] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[167] M. J. Paul and M. Dredze. *Social Monitoring for Public Health*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2017.

[168] S. Paul, S. K. Jandhyala, and T. Basu. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.

[169] T. Pedersen. Screening twitter users for depression and ptsd with lexical decision lists. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015.

[170] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. The development and psychometric properties of liwc2015. Ut faculty/researcher works, 2015.

[171] J. W. Pennebaker and L. A. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312, 1999.

[172] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1), 2003.

[173] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar*, pages 1532–1543, 2014.

[174] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[175] R. Plutchik. *The emotions: Facts, theories and a new model.* Crown Publishing Group/Random House, New York, USA, 1962.

[176] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia*, pages 275–281, 1998.

[177] D. Preoţiuc-Pietro, J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. A. Schwartz, and L. Ungar. The role of personality, age and gender in tweeting about mental illnesses. In *Proceedings of the 2nd Workshop on*

Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2015.

[178] D. Preotiuc-Pietro, J. Carpenter, S. Giorgi, and L. Ungar. Studying the dark triad of personality through twitter behavior. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 761–770, 2016.

[179] D. Preoţiuc-Pietro, M. Sap, H. A. Schwartz, and L. Ungar. Mental illness detection at the world well-being project for the clpsych 2015 shared task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015.

[180] V. M. Prieto, S. Matos, M. Alvarez, F. Cacheda, and J. L. Oliveira. Twitter: A good place to detect health conditions. *PLOS ONE*, 9(1):1–11, 01 2014.

[181] L. Qiu, H. Lin, J. Ramsay, and F. Yang. You are what you tweet: Personality expression and perception on twitter.

[182] A. Radford, R. C. Jeffrey Wu, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. Technical report, Technical report, OpenAi, 2019.

[183] L. S. Radloff. The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401, 1977.

[184] F. Ramiandrisoa, J. Mothe, F. Benamara, and V. Moriceau. IRIT at e-risk 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.

[185] D. Ramírez-Cifuentes and A. Freire. Upf's participation at the CLEF erisk 2018: Early risk prediction on the internet. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.

[186] D. Ramírez-Cifuentes, A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, D. A. Velazquez, J. M. Gonfaus, and J. Gonzàlez. Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis. *J Med Internet Res*, 22(7):e17758, Jul 2020.

[187] N. Ramírez-Esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *Proceedings of the Second International Conference on Weblogs and Social Media, ICWSM 2008, Seattle, Washington, USA, March 30 - April 2, 2008*, 2008.

[188] A. G. Reece and C. M. Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):15, Aug 2017.

[189] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer. Forecasting the onset and course of mental illness with twitter data. *Scientific Reports*, 7(1), 2017.

[190] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer. Forecasting the onset and course of mental illness with twitter data. *Scientific Reports*, 7(1), 2017.

[191] J. R. Reidenberg, T. Breaux, L. F. Cranor, B. French, A. Grannis, J. T. Graves, F. Liu, A. McDonald, T. B. Norton, R. Ramanath, N. C. Russell, N. Sadeh, and F. Schaub. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*, 30(1):39–88, 2015.

[192] P. Resnik, W. Armstrong, L. Claudino, and T. Nguyen. The university of maryland clpsych 2015 shared task system. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015.

[193] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber. Beyond lda: Exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015.

[194] P. Resnik, A. Garron, and R. Resnik. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, USA, 2013.

[195] S. Rhim, U. Lee, and K. Han. Tracking and modeling subjective well-being using smartphone-based digital phenotype. In *Proceedings of the 28th ACM*

Conference on User Modeling, Adaptation and Personalization, UMAP ’20, page 211–220, 2020.

[196] E. A. Ríssola, M. Aliannejadi, and F. Crestani. Beyond modelling: Understanding mental disorders in online social media. In *Proceedings of the 42th European Conference on Advances in Information Retrieval*, ECIR ’20, pages 296–310, Lisbon, Portugal, April 14–17, 2020.

[197] E. A. Ríssola, S. A. Bahrainian, and F. Crestani. Anticipating depression based on online social media behaviour. In *Flexible Query Answering Systems - 13th International Conference, FQAS 2019, Amantea, Italy, July 2-5, 2019, Proceedings*, pages 278–290, 2019.

[198] E. A. Ríssola, S. A. Bahrainian, and F. Crestani. Personality recognition in conversations using capsule neural networks. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019*, pages 180–187, 2019.

[199] E. A. Ríssola, S. A. Bahrainian, and F. Crestani. A dataset for research on depression in social media. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP’20, page 338–342, Genoa, Italy, July 14-17, 2020.

[200] E. A. Ríssola, D. E. Losada, and F. Crestani. Discovering latent depression patterns in online social media. In *Proceedings of the 10th Italian Information Retrieval Workshop*, IIR 2019, pages 13–16, Padova, Italy, September 16-18, 2019.

[201] E. A. Ríssola, D. E. Losada, and F. Crestani. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare*, 2(2), Mar. 2021.

[202] E. A. Ríssola and G. H. Tolosa. Improving real time search performance using inverted index entries invalidation strategies. *Journal of Computer Science and Technology*, 16(1):p. 6–13, Apr. 2016.

[203] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu. EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 3806–3813, may 2012.

[204] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA*, pages 109–126, 1994.

[205] A. Roshchina, J. Cardiff, and P. Rosso. A comparative evaluation of personality estimation algorithms for the twin recommender system. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 11–18, Glasgow, Scotland, UK, 2011.

[206] E. Ruane, A. Birhane, and A. Ventresque. Conversational AI: social and ethical considerations. In E. Curry, M. T. Keane, A. Ojo, and D. Salwala, editors, *Proceedings for the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland, December 5-6, 2019*, volume 2563 of *CEUR Workshop Proceedings*, pages 104–115. CEUR-WS.org, 2019.

[207] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems 30*, pages 3856–3866. Long Beach, USA, 2017.

[208] F. Sadeque, D. Xu, and S. Bethard. Uarizona at the CLEF erisk 2017 pilot task: Linear and recurrent models for early depression detection. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.

[209] A. Sadilek, H. A. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*, 2012.

[210] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *J Med Internet Res*, 17(7):e175, Jul 2015.

[211] J. Savoy. Authorship attribution based on specific vocabulary. *ACM Trans. Inf. Syst.*, 30(2), May 2012.

[212] H. A. Schwartz, J. Eichstaedt, M. L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, , and L. Ungar. Towards assessing changes in degree of depression through facebook. In *Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014.

[213] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9), 2013.

[214] A. See, S. Roller, D. Kiela, and J. Weston. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1702–1723, 2019.

[215] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3838–3844, 2017.

[216] R. Skaik and D. Inkpen. Using social media for mental health surveillance: A review. *ACM Comput. Surv.*, 53(6), dec 2020.

[217] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[218] C. Stachl, F. Pargent, S. Hilbert, G. M. Harari, R. Schoedel, S. Vaid, S. D. Gosling, and M. Bühner. Personality research and assessment in the era of machine learning. *European Journal of Personality*, 34(5):613–631, Sep 2020.

[219] S. Stajner and S. Yenikent. A survey of automatic personality detection from texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6284–6295, Barcelona, Spain, December, 2020. International Committee on Computational Linguistics.

[220] D. Stefanescu, R. Banjade, and V. Rus. Latent semantic analysis models on wikipedia and tasa. In *LREC*, 2014.

[221] J. Sun, H. A. Schwartz, Y. Son, M. L. Kern, and S. Vazire. The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, 118(2):364–387, 2020.

[222] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 2009.

[223] R. P. Tett, D. N. Jackson, and M. Rothstein. Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, 44(4):703–742, 1991.

[224] A. Thieme, D. Belgrave, and G. Doherty. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Trans. Comput.-Hum. Interact.*, 27(5), Aug. 2020.

[225] B. Thompson. *Foundations of Behavioral Statistics: An Insight-Based Approach*. The Guilford Press, 2006.

[226] M. Tkalcic, B. D. Carolis, M. de Gemmis, A. Odic, and A. Kosir, editors. *Emotions and Personality in Personalized Services - Models, Evaluation and Applications*. Human-Computer Interaction Series. Springer, 2016.

[227] M. Tkalcic and L. Chen. Personality and recommender systems. In *Recommender Systems Handbook*, Berlin, Heidelberg, 2010. Springer-Verlag.

[228] M. Trotzek, S. Koitka, and C. M. Friedrich. Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.

[229] L. Tudor Car, D. A. Dhinagaran, B. M. Kyaw, T. Kowatsch, S. Joty, Y.-L. Theng, and R. Atun. Conversational agents in health care: Scoping review and conceptual analysis. *J Med Internet Res*, 22(8):e17158, Aug 2020.

[230] A. S. Uban and P. Rosso. Deep learning architectures and strategies for early detection of self-harm and depression level prediction. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[231] S. Urbina and A. Anastasi. *Psychological testing*. Upper Saddle River, N.J.: Prentice Hall, 7th edition, 1997.

[232] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *Canadian journal of psychiatry. Revue canadienne de psychiatrie*, 64(7):456–464, Jul 2019.

[233] E. Villatoro-Tello, G. Ramírez-de-la-Rosa, and H. Jiménez-Salazar. Uam's participation at CLEF erisk 2017 task: Towards modelling depressed blogers. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.

[234] A. Vinciarelli and G. Mohammadi. A survey of personality computing. *IEEE Trans. Affect. Comput.*, 5(3):273–291, 2014.

[235] C. G. Walsh, B. Chaudhry, P. Dua, K. W. Goodman, B. Kaplan, R. Kavuluru, A. Solomonides, and V. Subbian. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. *JAMIA Open*, 01 2020.

[236] R. Wang, W. Wang, A. daSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1), 2018.

[237] J. K. White, S. S. Hendrick, and C. Hendrick. Big five personality variables and relationship constructs. *Personality and Individual Differences*, 37(7):1519–1530, 2004.

[238] M. Wolf, F. Theis, and H. Kordy. Language use in eating disorder blogs: Psychological implications of social online activity. *Journal of Language and Social Psychology*, 32(2):212–226, 2013.

[239] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. Yu. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099, Brussels, Belgium, 2018. Association for Computational Linguistics.

[240] H.-C. Yang and Z.-R. Huang. Mining personality traits from social messages for game recommender systems. *Knowledge-Based Systems*, 165:157 – 168, 2019.

[241] M. Yang, W. Zhao, J. Ye, Z. Lei, Z. Zhao, and S. Zhang. Investigating capsule networks with dynamic routing for text classification. In *Proceedings*

of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3110–3119, Brussels, Belgium, 2018.

[242] T. Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363 – 373, 2010.

[243] A. Yates, A. Cohan, and N. Goharian. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark*, pages 2968–2978, 2017.

[244] E. Yom-Tov, L. Fernandez-Luque, I. Weber, and S. P. Crain. Pro-anorexia and pro-recovery photo sharing: A tale of two warring tribes. *J Med Internet Res*, 14(6):e151, Nov 2012.

[245] W. Youyou, D. Stillwell, H. A. Schwartz, and M. Kosinski. Birds of a feather do flock together: Behavior-based personality-assessment method reveals personality similarity among couples and friends. *Psychological Science*, 28(3):276–284, Mar 2017.

[246] F. Zarrinkalam, M. Kahani, and E. Bagheri. Mining user interests over active topics on social networks. *Information Processing Management*, 54(2):339–357, 2018.

[247] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Melbourne, Australia, 2018.

[248] G. K. Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, Oxford, England, 1949.