Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Mental disorders on online social media through the lens of language and behaviour: Analysis and visualisation

Esteban A. Ríssola [a],[*], Mohammad Aliannejadi [b], Fabio Crestani [a]

[a] *Faculty of Informatics, Università della Svizzera italiana, Lugano, Switzerland*
[b] *IRLab, University of Amsterdam, Amsterdam, The Netherlands*

ARTICLE INFO

ABSTRACT

Due to the worldwide accessibility to the Internet along with the continuous advances in 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. In this paper, we undertake a thorough analysis towards better visualising and understanding the factors that characterise and differentiate social media users affected by mental disorders. We perform different experiments studying multiple dimensions of language, including vocabulary uniqueness, word usage, linguistic style, psychometric attributes, emotions' co-occurrence patterns, and online behavioural traits, including social engagement and posting trends.

Our findings reveal significant differences on the use of function words, such as adverbs and verb tense, and topic-specific vocabulary, such as biological processes. As for emotional expression, we observe that affected users tend to share emotions more regularly than control individuals on average. Overall, the monthly posting variance of the affected groups is higher than the control groups. Moreover, we found evidence suggesting that language use on micro-blogging platforms is less distinguishable for users who have a mental disorder than other less restrictive platforms. In particular, we observe on Twitter less quantifiable differences between affected and control groups compared to Reddit.

1. Introduction

Over the past decade, there has been increasing interest in modelling and detecting mental state changes by analysing online digital footprints (Correia et al., 2020). Among the most relevant reasons for this there is the fact that public health systems are limited in their abilities to deal with the large volume of cases they face daily. Yet online social media platforms are spreading and thus, changing the dynamics of mental health assessment (Chancellor & Choudhury, 2020; Skaik & Inkpen, 2020; Thieme et al., 2020). On a regular basis, individuals use social media platforms to share their interests (Zarrinkalam et al., 2018), personal daily life experiences (Khodabakhsh et al., 2018; Saha et al., 2021), and even their moods and feelings (Prieto et al., 2014). Using such user-generated content, the development of low-cost, unobtrusive prognostic technologies to assess risks and aid at making decisions has excellent potential to allow healthcare practitioners to perform preliminary screenings and raise awareness of mental health outcomes at a large scale (Ramírez-Cifuentes et al., 2020; Yates et al., 2017).

Researchers studying the connection between psychology and language have found many cues about individuals' mental states (including their social and emotional condition), which can be assessed by examining their linguistic expression patterns. In this regard, language use could act as an indicator of the current mental state (Coppersmith, Dredze, Harman, & Hollingshead, 2015;

---

* Corresponding author.
E-mail address: esteban.andres.rissola@usi.ch (E.A. Ríssola).

Ríssola et al., 2019a), personality (Neuman, 2016; Ríssola et al., 2019b) and even personal values (Aliannejadi & Crestani, 2020; Boyd et al., 2015) of individuals. According to Pennebaker et al. (2003), the main reason behind this is that latent mental parameters are encoded in the words that individuals employ to express themselves. Moreover, several studies have shown that individuals' online behavioural traits, such as social engagement, collectively provide significant cues about their mental state. On many occasions, such traits can be considered as proxies, although imperfect, for findings in the medical literature related to positive mental health concerns outcomes (Choudhury et al., 2013; Coppersmith et al., 2014a; Reece et al., 2017). For instance, literature on depression indicates that individuals presenting signs of depression exhibit significant changes in their online activity levels in the evening and before dawn (ten Thij et al., 2020).

Constraints dictated in real-life settings, such as cost and time, hinder the possibility to achieve a timely and effective personal diagnosis. Recently, initiatives such as the Strategic Workshop on Information Retrieval in Lorne (Culpepper et al., 2018) (SWIRL) spotlight the possibility of cross-disciplinary collaborations with several scientific disciplines, including psychology. In particular, they highlight that the core principles of Information Retrieval, and related fields, have been gradually extended and applied in domains that a few years back were not easy to conceive or imagine. In this regard, the Computational Linguistics and Clinical Psychology (CLPsych) (Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015) and the Early Risk Prediction on the Internet (eRisk) (Losada et al., 2018, 2019) experimental frameworks have fostered collective efforts worldwide to leverage social media data to develop models for estimating the occurrence of signs related to mental disorders.

These initial efforts to address the automatic identification of potential cases of mental disorders on social media have mainly modelled the problem as classification (Masood, 2019; Ríssola et al., 2021). Researchers participating in these workshops have examined a wide variety of methods to identify positive cases[1] (Burdisso et al., 2019; Cacheda et al., 2019); 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 (Ríssola et al., 2020). 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. (2020), 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 of great importance to conduct experiments providing insights on how individuals' language attributes and online behaviour are distinctive among users affected by mental health concerns as well as between different mental 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. In this way, forecasting systems could be complemented with preliminary step-by-step directions for practitioners to identify high-risk individuals based on both 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.

In this work, we conduct a thorough study of various dimensions of language and online behavioural traits to characterise users affected by mental disorders on two popular social media platforms, namely, Reddit and Twitter. Moreover, we present several methods for visualising the data to give mental health professionals a deeper understanding of the information extracted. To this aim, we start by comparing users affected by a particular disorder against control individuals in each platform separately. Subsequently, our second aim is to determine whether different mental disorders share the same features or if they are distinct according to the dimensions studied. Finally, we investigate whether the trends and patterns are held 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.

Our main research questions related to the language and behaviour of users of social media platforms are:

- **RQ1**: How different is the 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?

This paper extends our previous work on improving the understanding of mental disorders on online social media (Ríssola et al., 2020). We previously stressed that language technologies have a 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. Driven by this, we conducted an extensive language analysis to improve the understanding and visualisation of the features that characterise and differentiate social media users affected by mental disorders on

---

[1] Throughout the article, we use the term *positive* to refer to individuals who are potentially suffering from mental health concerns such as depression, anorexia, self-harm or post-traumatic stress disorder (PTSD). In contrast, individuals who are in the *control* group do not suffer from any of the previously cited disorders.

Reddit. In particular, we proposed a series of techniques based on statistical and visual analyses which are able to identify significant differences in the language of high-risk individuals on Reddit.

With this in mind, we further explore how mental disorders are manifested on social media as we extend the language analyses from Reddit to Twitter, another popular platform subjected to strong constraints, such as the length of a posting in number of characters. Furthermore, we complement the language dimensions studied on both platforms with a behaviour analysis. In particular, we study a set of attributes that profile the online behavioural traits and trends of affected users, such as the posting regularity. As shown later, the differences observed, mostly between positive and control instances, have a varied magnitude depending on the social media platform. Therefore, given the characteristics and limitations imposed by the platform, there are attributes that provide stronger clues about the manifestation of mental disorders. For instance, many of the language dimensions analysed in this work exhibit larger variance on Reddit than on Twitter. This is an ideal complement to our previous work on improving the understanding of how mental health disorders are manifested on social media as it uncovers the differences in people's language and behaviour based on their participation on social media platforms with different characteristics and constraints. In summary, this paper extends our previous work as follows:

- It extends the language closed-vocabulary analysis adding a complementary dictionary-based tool known as Empath (Fast et al., 2016) and compares the new results with the previous ones.
- It complements language analyses on Reddit by profiling users' online behavioural traits, such as the usage of elements like hashtags, mentions and emoticons.
- It extends Reddit's language and online behaviour analyses to another social media platform, in particular, to Twitter data.
- It presents a thorough comparison between the language and online behaviour of users on both social media platforms to shed light on how distinctive the expression of mental disorders could be on different platforms to understand which of the dimensions analysed are more suitable for each case.

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. This analysis is a valuable step towards developing better risk-assessment and decision-support tools.

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. Conversely, comparing the mental disorders analysed, on Reddit and Twitter respectively, we were not able to find significant attributes which could differentiate one from the other, though such disorders are identified in the language and behaviour of the individuals. In addition, we observe less quantifiable differences between positive and control groups on Twitter than on Reddit, suggesting that the use of language on Twitter is less distinguishable between users potentially suffering mental disorders. This could be due to the platform constraints, most importantly the character limit. Nonetheless, the usage of social engagements elements, such as hashtags, mentions and emojis, revealed meaningful differences between the various groups on Twitter.

The remainder of the paper is organised as follows. Section 2 summarises the related work; Section 3 details the approach followed to answer the research questions; Section 4 describes the specific data used in this work; Section 5 presents the corresponding results and analyses; Section 6 concludes the work.

2. Related work

Social media has been extensively studied for its potential to automatically identify individuals suffering from mental disorders. Here, we summarise a number of relevant studies that have attempted to examine the connection between language, behaviour and mental disorders on social media from an analytical perspective.

Using online social media data as a gauge of depression, Park et al. (2012) performed a preliminary study to identify whether this data was truly representative of users' clinical depressive symptoms. The data analysis conducted revealed that depression was mentioned most frequently for describing one's depressed state and, to less degree, for sharing general information about depression.

An early work on automatic depression detection was undertook by Choudhury et al. (2013) who gathered ground truth data on the presence of depression using crowd-sourcing. 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, Dredze, Harman, Hollingshead, and Mitchell (2015) investigated Twitter data to determine which language features could be helpful to distinguish between mental disorder sufferers and healthy individuals. It was found that, though a set of useful features could be identified, there are still language differences in communicating about mental health concerns which worth investigate.

Reece et al. (2017) conducted a state-space temporal analysis whose goal was to track the evolution of depression and PTSD. They developed time series models which, exploiting posts' textual content, presented distinctive patterns between positive and controls users without any supervision signal. Furthermore, such models exhibit their potential to outline a plausible timeline for the disorders development (onset and recovery).

Choudhury et al. (2016) studied several mental health and suicide support communities on Reddit[2] whose members mostly participate looking for help and support. The authors identified that transitions from mental health disclosure to an expression of

---

² Denominated *subreddits* on the platform.

suicidal ideation were associated with various markers, such as heightened self-attentional focus, poor linguistic coherence, reduced social engagement, and expression of negative affects like as anxiety.

Gkotsis et al. (2016) analysed various mental health subreddits to discover discriminating language features between the users in the different communities. Their study concluded that, in general, non-topically related subreddits had distinctive syntactic and lexical features and condition-specific vocabularies. Similarly, Gaur et al. (2018) presented an unsupervised approach to map the content of various mental health-related subreddits to the best matching DSM-5[3] categories. They developed a domain-specific lexicon alongside a drug abuse ontology and used them to quantify the relationship between subreddits' content and DSM-5 categories by automatically assigning the corresponding labels. Also on Reddit, Yoo et al. (2019) studied how users perceived bipolar and depressive disorders and shared medical information about them on the respective mental health support communities on the platform. Relying on semantic network analysis, they observed that, overall, users on both communities often expressed negative emotions, and showed great interest in writing about their past episodes and sleeping problems.

Although insightful, such studies of Reddit communities might not provide accurate information about the distinction between mental health sufferers and healthy users as several members of such subreddits are persons who search for advise or to share their experience about close friends or relatives who could be experiencing a mental health concern. In this work, we focus on the analysis of users' language attributes irrespective of the topic discussed.

Overall, much work is focused on automatically identifying whether online social media users are deemed as positive mental health concerns cases. On very few occasions, the authors of such works attempt to obtain some explanations from the outcomes of the classification models to better understand why certain decisions were taken, as happened with some eRisk participants (Amini & Kosseim, 2020; Aragon et al., 2021; Uban et al., 2021). However, research devoted to understanding, measuring, visualising and providing insight on the attributes that characterise such users and differentiate them both from healthy individuals and between diverse disorders has been noticeably scarce. Moreover, the language dimensions and behavioural traits studied in this work have not yet been contrasted on social media platforms of different characteristics and constraints.

3. Experimental design

Language and behaviour analysis usually focus on one of two applications, *prediction* (typically, involving supervised machine learning algorithms) or *insight*. In prediction, the ultimate goal is to accurately map language and behaviour into a single or few given outcomes. Insight is conceptually focused on the opposite, finding the language and behaviour that is most characteristic of an outcome, i.e., identifying the patterns that provide a window into thoughts, attitudes, psychology or health of people. In this work, we sought to gain *insights* on the manifestation of mental disorders on online settings, particularly social media platforms. With this in mind, in the following section we outline the experimental design we defined to analyse social media posts from the perspective of language and behaviour. We describe the main outcome of each experiment and state how the differences between the classes under analysis can be quantified.

3.1. Open-vocabulary

Vocabulary Uniqueness: To address **RQ1** and **RQ6**, we analyse the similarity (or diversity) of the unique sets of words that comprise the vocabulary of the classes (i.e., positive and control). With this analysis we can investigate to which extent the two groups share a common vocabulary and whether specific words (if any) could be only employed by the members of a certain class.

Considering each vocabulary as a set, we inspect the relative size of their intersection by computing Jaccard's index. This metric provides a way to quantify 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 calculate Jaccard's index as follows:

$$J(P, C) = |P \cap C| / |P \cup C| . \tag{1}$$

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. As the value approaches 1 the sets share an increasing overlap of elements. Conversely, an index close to 0 indicates that the sets are more diverse among themselves.

Word Usage: Complementary to analysing vocabularies overlapping and differences, we also study word usage patterns as a mean to inspect the language of the different classes. Here, we sought to address **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.

Based on both current and historical data, language models are probability distributions over the words in a collection of documents which intend to capture the regularities of language. Thus, when considering a unigram language model each of the terms in the vocabulary will receive a probability based on its occurrence patterns. An estimate of the likelihood that a word will appear in a document D in a collection of documents S is done using:

$$P(w_i|D) = (1 - \alpha_D)P(w_i|D) + \alpha_D P(w_i|S) , \tag{2}$$

---

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

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, known 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 work, 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.

After calculating the language models for each class, we graphically analyse the probability distributions computed and study how much the distributions diverge. To support our visual observations, we use Kullback–Leibler divergence (KL). This measure from probability and information theories allows us to quantitatively measure the extent to which the probabilities distributions under observation differ. When the distributions compared are identical, their KL-divergence reaches 0. Thus, the divergence is always positive 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:

$$KL(P \parallel C) = \sum_x P(x) log \frac{P(x)}{C(x)} \, . \tag{3}$$

3.2. Linguistic style and psychometric attributes

A usual approach in which language can be associated to psychological variables consists in counting words that are included within different categories of language manually defined (Choudhury et al., 2013; Chung & Pennebaker, 2007; Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015). As opposed to the experiment outlined in the previous section, this type of analysis is called "close-vocabulary" (Schwartz et al., 2013). By studying *function words*[4] and topic-specific vocabulary, we address **RQ1**, **RQ3**, **RQ4**, and **RQ6**. This kind of analysis is aimed at quantifying specific stylistic patterns that could distinguish positive cases of mental disorders from control cases. For instance, when people are depressed, they tend to focus on themselves, and therefore, it is expected that they would employ personal pronouns like "I" more frequently (Association, 2013). Additionally, individuals suffering from certain mental health disorders may use more words related to specific topics. We showcase this later by comparing depression and self-harm to cases of anorexia.

It worth noting that we decide to keep the stop-words since many words such as pronouns, articles and prepositions disclose individuals' way of thinking, emotional state, personality, and relationship with others (Chung & Pennebaker, 2007). Indeed, such words, called function words, represent less than 0.1% of people's vocabulary, but nearly 60% of the words they employ (Chung & Pennebaker, 2007).

By analysing the writing style of individuals, *Linguistic Inquiry and Word Count* (Tausczik & Pennebaker, 2010) (LIWC)[5] provides mental health practitioners with quantitative data regarding those individuals mental health. LIWC consists of an extensive set of dictionaries compiled by psychologists, and including psychologically meaningful categories, which are useful for analysing the linguistic style patterns of an individual's writing style. In this research, we determine the proportion of documents authored by each individual that has at least one word on specific LIWC categories and produce a distribution. Specifically, we select a subset of the psychometric categories where significant differences were found between positive and control group members. Afterwards, we depict and compare the resulting distributions employing box-plots.

Furthermore, we extend this analysis using Empath[6] (Fast et al., 2016). 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 use box-plots to visualise the distributions obtained from a subset of categories where we observed significant differences between control and positive users and contrast them.

3.3. Emotional expression

Typically, people express their feelings, attitudes and emotions through their words. As an example, *gloomy* and *cry* evoke sadness, while *delightful* and *yummy* denote joy. In this section, we focus on **RQ1**, **RQ3**, **RQ4**, and **RQ6** by examining how people with mental disorders convey their emotions in social media posts. In addition, we study whether emotional expression could be a differential factor between users affected by mental health concerns and those who are not. To this aim, we followed Plutchik's model of emotions (Plutchik, 1980) which proposes the existence of eight basic and prototypical emotions, these include joy, sadness, anger, fear, disgust, surprise, trust and anticipation. In addition, we also considered the polarity (positive or negative).

To capture word-emotion connotations, we apply the emotion lexicons that (Mohammad, 2018; Mohammad & Turney, 2013) developed, in which each word is associated with the emotion it conveys. These lexicons include commonly used words from social media platforms along with common English terms. Furthermore, they contain words that not necessarily evoke a certain emotion, but which co-occur with those that do. As an example, the words *death* and *failure* refer to notions which are often associated to sadness, and, hence, suggest some kind of sadness.

---

[4] Function words contribute to the structure of the sentence, rather than to its meaning.
[5] See http://liwc.wpengine.com/.
[6] See http://empath.stanford.edu/.

3.4. Social engagement and posting trends

Coppersmith et al. (2014b) suggested that users' online behavioural traits could be profiled by studying various attributes related to users' documents, such as creation time and length, as well as users' level of participation on the social media platforms. Such features constitute proxies, albeit imperfect, for significant findings in the mental health literature that may be manifested and quantified on online settings. Furthermore, on some occasions they might provide a starting point for further research into behavioural traits as observed on social media. As an example, Reece et al. (2017) found a relationship between higher average document length and both depression and PTSD. Yet, in the traditional medical literature few studies link verbosity to mental health concerns. With this in mind, we address **RQ2** and **RQ6** by profiling the online behaviour of users considering the following elements:

- Mentions: A document containing another account's username, preceded by the "@" symbol. For instance, `Hello @earis-sola!`.
- 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: A smiley or an ideogram inserted into an e-mail or web page. Emojis are similar to emoticons, though they are pictures rather than typographical representations.
- Other annotated elements: It includes words with emphasis (such as `I don't *think* I...` and `a *great* time`), censored words (such as `f**k, s**t`) and repeated words.
- Retweets (RT): On Twitter, usually "RT" indicates that someone is re-posting someone else's content, i.e., quoting another 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.

We examine each user's documents to determine what percentage scores positively on any of these elements (i.e., have at least one occurrence of that element). Using box-plots, we visualise and contrast the distributions obtained.

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 respective 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:

$$\frac{1}{n-1} \sum_{t=2}^{n-1} t_{n-1} - t_n, \tag{4}$$

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.

4. Data

In this work, we hypothesise that users with a mental disorder behave differently on different social media platforms. One approach, data type, and data modality does not reveal the same information on every social media platform. For example, language use can be a very discriminating feature on a platform, while performing poorly on another. For this reason, we study various mental disorder indicators and visualisation techniques on two mainstream social media platforms, namely, Reddit and Twitter. Our choice is motivated by the fact that the two platforms have very different environment and features, leading to a very different range of behaviours by their users (Roberts et al., 2012).

Reddit "is a social news aggregation, web content rating, and discussion website, recently including livestream content through Reddit Public Access Network".[7] It consists of various *subreddits*, each of which focuses on a different topic. Users can post pictures, web links, or other types of content. The platform does not enforce any extreme restrictions in terms of the length of posts. **Twitter** "is an American microblogging and social networking service on which users post and interact with messages known as tweets".[8] As the description suggests, users on Twitter post short messages. The structure and features of Twitter are very different from Reddit. For instance, sharing a post (i.e., retweet) is a common means of interaction with a post. Limiting each post to 280 characters has various effects on the way people express themselves and their feelings. It can also affect the topic and frequency of the posts on Twitter instead of other platforms (Gligorić et al., 2018).

With this in mind, as a test bed for our experiments we use the datasets developed at CLPsych (Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015) (Twitter) and eRisk (Losada et al., 2018, 2019) (Reddit) experimental frameworks. These initiatives

---

[7] See https://www.reddit.com.

[8] See https://twitter.com.

Table 1

Summary of eRisk 2018 and 2019 collections (Reddit) and CLPysch 2015 (Twitter). The activity period represents the number of days passed from the first to last the document collected for each user.

(a) Reddit (eRisk)

	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

(b) Twitter (CLPsych)

	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

sough to develop a common evaluation framework for researchers interested in advancing the state of the art on the assessment of mental disorders, such as depression or PTSD, on social media platforms. Each dataset is comprised of submissions in the form of posts created by users of social media platforms, namely Reddit and Twitter. Two groups of users are included in each dataset: (a) cases of users potentially suffering from mental health concerns, such as depression (positive group); and (b) control individuals. On both experimental frameworks individuals in the positive group were collected by searching for self-expressions of diagnoses (e.g., the sentence "I was diagnosed with depression") and manually verifying that the posts retrieved truly contained a genuine statement of diagnosis (Coppersmith et al., 2014a). Conversely, members of the control group were randomly sampled from the extensive pool of users available on the platform. Given the API limits imposed by each social media platform, the datasets creators were able to collect up to 2000 submissions per user for eRisk and 3000 for CLPysch.

We decide to conduct our various analysis using these collections since they have been carefully curated, validated and used through the various experimental frameworks' 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. It should be noted that a direct comparison at the level of classes between eRisk and CLPsych is limited by the availability of mental disorders that were included in each dataset. The only overlapping label between both platforms is depression. Nonetheless, we consider both datasets as valuable and valid assets to perform the analyses proposed in this work as we are interested in determining whether the general trends and patterns we observed on users affected by mental disorders are held on both platforms, despite the particularities that each one has. Especially when users are subject to various types of constraints (e.g. character limit), which could influence their writing style and expression (Gligorić et al., 2019). Table 1 presents a summary of the datasets employed in this research.

5. Results and analyses

In this section, we outline the results reached after analysing the various collections presented in Section 4 which were used for the experiments described in Section 3 aiming at addressing the research questions posed in the introduction.

5.1. Open-vocabulary

Vocabulary Uniqueness: In Table 2 we analyse the vocabulary used by different groups of individuals suffering from mental health concerns (i.e, depression, anorexia and self-harm for Reddit and depression and PTSD for Twitter) in comparison to the respective control groups. In particular, we study vocabularies' difference, intersection and union. Our analysis of Reddit data indicates that positive cases of depression and anorexia are highly similar to the relevant control groups, with Jaccard indexes of 59% and 65%, respectively. The self-harm cases, however, tend to use a broader range of words in their documents as compared to the control group (44%). One possible explanation for such difference might be related to the way in which the collections where built. When creating depression and anorexia collections, eRisk's organisers also included in the respective control groups users who were active on Reddit's communities devoted to discussing about depression or anorexia, but were not suffering from any disorder. Examples of this are a mental health practitioner giving support to other forum members or people who have some relative suffering from depression or anorexia. This type of control users made the task (i.e., determining early signs of depression or anorexia) more challenging because their topics of interest are highly related to those of the positive group. Conversely, none of these users were included in self-harm's control group.

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. In spite of this, it should be noted that such noticeable differences

Table 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

may be due in part to the informal nature that characterises microblogging platforms (Choi & Croft, 2012; Metzler et al., 2012; Nepomnyachiy et al., 2014) along with Twitter's limit of maximum characters per post. 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 (Ríssola & Tolosa, 2016). In fact, the number of unique terms on Twitter is substantially larger than on Reddit, possibly causing a smaller overlap between the sets compared.

Furthermore, such comparisons between the different vocabularies allow us to outline which words are used by positive groups and not by control users. For instance, on Reddit we found the following terms to be distinctive for positive groups: self-harm, trazodone[9] (Depression); anorexics, depersonalisation,[10] emetrol, peptol[11] (Anorexia).

Word Usage: In Fig. 1 we examine and contrast the language models derived from the various collections and classes. As previously observed, there are cases of out-of-vocabulary words between the classes. That is, there are terms which only appear in the positive class vocabulary and not in the control one (and vice-versa). In that case, smoothing becomes necessary to control the probability assigned to such out-of-vocabulary words. To illustrate how different the language use between positive and controls users could be, we turn to Fig. 1(a) in which we showcase the language models of depression and control classes on Reddit. As observed, word usage between positive and control depicts very different trends. In contrast, as Figs. 1(c) and 1(d) reveal, on Twitter such differences are much smoother as there is a notable overlap between the distributions.

Table 2 presents the KL-divergence computed between the positive and control groups' language models on both platforms. Since the KL-divergence lacks an upper bound, we also compute it among the different control groups to serve as point of reference. We observe that KL-divergence calculation reinforces the trends observed in the plots, specially on Reddit. As we contrast the distribution of different control groups on Reddit, we note smaller KL-divergence values (≈ 0), denoting that these distributions are very similar. Conversely, this value is considerable larger when comparing positive and control classes. Moreover, when compared to Reddit, the positive groups on Twitter exhibit almost half of the KL-divergence with respect to controls. Interestingly, the divergence observed on Twitter between positive and controls groups and vice-versa is practically equal. Recall that KL-divergence is a distribution-wise asymmetric measure that allows 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, as observed in Fig. 1(b), it is difficult to spotlight prominent differences between the language models of the three positive classes on Reddit. Anorexia and depression exhibit similar word usage distributions. In comparison with either depression or anorexia, self-harm shows the most noticeable difference. On Twitter, positive group language models present an even larger overlap.[12] This strengthen the fact that the word probability distribution between the various positive groups is very similar and how they use words. Notice that language models for depression in Figs. 1(b) and 1(b) exhibit different shapes. The reason for this is related to they way language models are computed. As stated in Section 3 an estimate of the likelihood for each word is calculated

---

[9] *Trazodone* is a medication that helps treat depression.

[10] People suffering from depersonalisation feel detached or disconnected from their bodies and thoughts.

[11] *Peptol* and *Emetrol* are pharmacological drugs that relieve stomach discomfort.

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

(a) Reddit: Depression vs. control

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

(c) Twitter: Depression vs. control
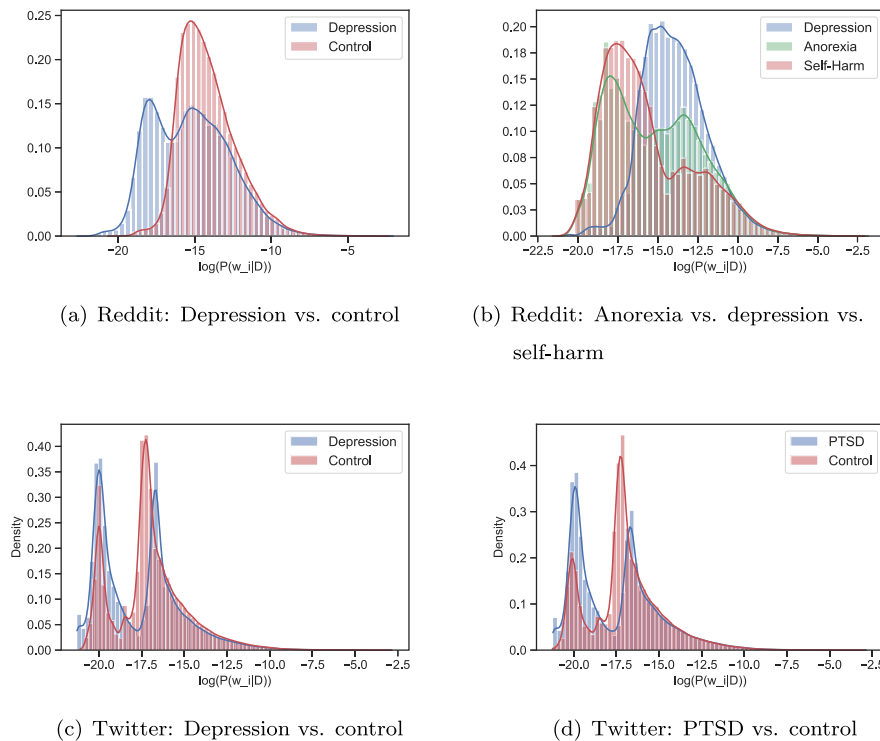
(d) Twitter: PTSD vs. control

Fig. 1. Language models probability distribution comparison on Reddit and Twitter. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

based on its frequency on both the concatenation of a class of documents and the whole collection. Thus, depending on the classes involved in the computation, the second term in formula (2) will vary. Recall that this term allows to deal with out-of-vocabulary words.

Practitioners could used the open-vocabulary analysis presented in this section as supporting material to paint a bigger picture about users mental health state. For instance, the language model of a potentially affected user could be contrasted against reference language models for positive and control classes to determine the which one it resembles the most. we have randomly sampled 10% of users with anorexia and 10% of control users and calculated the KL-divergence of each individual language model against reference language models for positive and control, and report the average values of each class. We observe that users with anorexia consistently exhibit a lower divergence with respect to positive reference language model than to the control one (0.41 vs. 0.49). Conversely, control users showed a higher similarity with the control reference language model than with the positive one (0.82 vs. 0.90).

More precisely, we have randomly sampled 10% of users with anorexia and 10% of control users and calculated the KL-divergence of each individual language model against reference language models for positive and control, and report the average values of each class. We observe that users with anorexia consistently exhibit a lower divergence with respect to positive reference language model than to the control one (0.41 vs. 0.49). Conversely, control users showed a higher similarity with the control reference language model than with the positive one (0.82 vs. 0.90). We have described this proof of concept on Section 5.1.

5.2. Linguistic style and psychometric attributes

By comparing a selected set of categories from LIWC and Empath,[13] we show that language use has significant differences between positive and control users on Reddit and Twitter. Fig. 2 depicts the proportion of documents authored by the different users that have at least one word on selected LIWC categories. The bars in the plots are coloured with respect to the class they represent. 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).

Interestingly, we observe that on both social media platforms there is a notorious difference between positive and control users in the use of the personal pronoun "I". This replicates previous findings for depression (Choudhury et al., 2013; Coppersmith, Dredze,

---

[13] Due to space constraints and since trends on Empath support those of observed on LIWC, we do not include the corresponding figures.

Harman, Hollingshead, & Mitchell, 2015). Moreover, the proportion of messages using words related to positive emotions (named *posemo*) is larger than negative ones (named *negemo*), even for the positive classes, as measured by LIWC on Reddit and Twitter. As literature suggests, a reason for this might be that English words used in everyday language are biased towards positivity (Kloumann et al., 2012). Furthermore, this phenomena known as Pollyanna effect (Dodds et al., 2015) has also been observed in similar studies (Bathina et al., 2021). The differences identified on Reddit using LIWC and Empath between positive and control groups are statistically significant when applying a Welch two-sample t-test (p-value < 0.001). The only exceptions in LIWC are categories *we* and *she/he* as observed on Fig. 2(a). On Empath categories affect, nervousness, shame, optimism, love, pain, suffering, violence, death and healing yield significant differences between positive classes and their respective control groups. In contrast, between depression, anorexia and self-harm we were not able to find any significant differences among these categories and those in Fig. 2(a). Similarly, in most of the categories analysed with LIWC on Twitter illustrated in Fig. 2(c) statistically significant differences are observed between at least one positive group and its respective control. Using Empath we observe this on categories posemo, negemo, violence, shame, feminine, swearing, communication and death. Moreover, we also see significant differences between individuals suffering from depression and PTSD, including cognitive and social processes and drives for LIWC and violence, feminine, death, communication and swearing for Empath.

In Fig. 2(b) we also showcase LIWC categories relative to *biological process*. This group of categories is comprised of words which are connected to the body and its essential functions. We observe that users with anorexia exhibit a particular use of these words when they are contrasted against depression and self-harm classes. Intuitively, this outcome is expected since anorexia is characterised by an intense fear of gaining weight and a distorted perception of the body image. In general, individuals affected by anorexia are highly likely to retain under control their weight and physical shape, which can pose a strong negative effect on their lives. Such individuals may, therefore, discuss themes about their bodies and its functioning more frequently. In this regard, we found that categories *ingest*, *body* and *health* have statistically significant differences (p-value < 0.001) between anorexia and depression. In the case of self-harm, only *ingest* category is significantly different from anorexia. Furthermore, LIWC observations are complemented and supported by Empath's categories body, eating, exercise, and ugliness.

Finally, we extend LIWC built-in categories with other domain-specific lexicons. In particular, we consider (Choudhury et al., 2013) depression lexicon[14] which consists of words closely related to texts authored by people sharing their experience on mental disorders or its symptoms on online settings and also include words associated to names of medications. We find that, for both Reddit and Twitter, the use of such terms does not significantly differ between users with depression and the rest of the mental disorders studied. This evidences that users suffering from either self-harm, anorexia or PTSD often employ such words as well. Similarly, we were not able to find any significant differences with the set of absolutist terms[15] retrieved from the research conducted by Al-Mosaiwi and Johnstone (2018).

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 constraints disproportionately preserve negative emotions and articles, adverbs, and conjunctions have the highest probability of being omitted (Gligorić et al., 2019). 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 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. Using two different though complementary dictionary-based tools, we have shown that this close-vocabulary analysis provides a reliable way to collect quantitative data from user-generated content about the mental state of individuals on social media.

5.3. Emotional expression

Based on the average number of documents that contain at least one word associated to a particular emotion (including the polarity), Fig. 3 provides a pairwise comparison between the different positive classes and their respective controls. On Reddit, there are considerable differences between the emotional expressions of the classes under analysis. These results suggest 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 who tend to share slightly more emotions on average. 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 expression 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.

Additionally, Figs. 4 and 5 depict a frequency correlation analysis of the various emotions comparing each class against its relevant control. Due to space constraints, we only include depression. However, similar trends are observed from the remaining

---

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

[15] Examples of absolutist terms: *absolutely*, *constantly*, *definitely*, *never*.
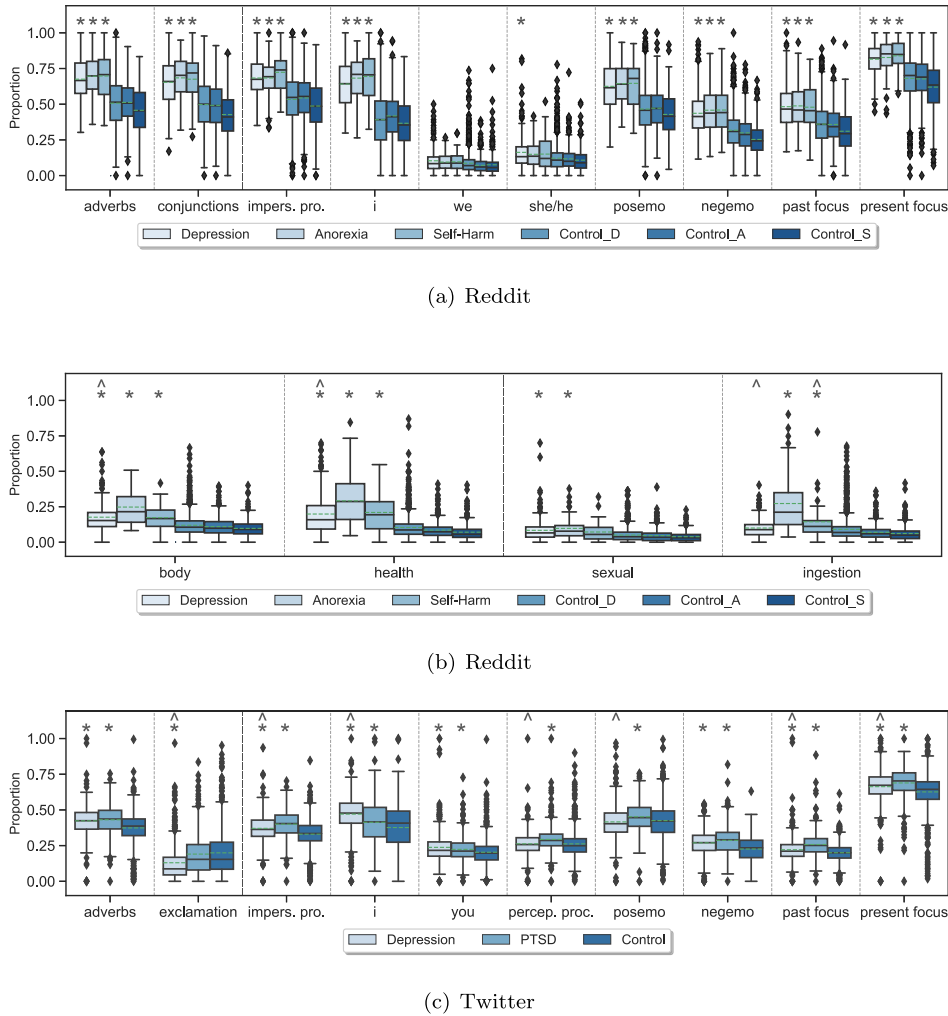
(a) Reddit

(b) Reddit

(c) Twitter

Fig. 2. Box and whiskers plot of the proportion of documents that users have on Reddit and Twitter (*y*-axis) matching selected LIWC categories. 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 on Reddit and Depression/PTSD on Twitter are denoted by ˆ (*p*-value < 0.001).

positive classes. We observe that different classes exhibit different correlations with certain emotions. Interestingly, sadness shows a negative correlation with joy and positive polarity for depression (Fig. 4). However, such correlation is not observed for the respective control group. As another example, we see that for self-harm's respective control class surprise presents a larger positive correlation with joy, trust, as well as with positive and negative polarities when compared to self-harm class. The opposite trend is observed between surprise and disgust as well as between fear and disgust. Interestingly, when we compare the collective correlation of surprise with all other emotions, we see a considerable difference between Control and Depression groups. While the Control group exhibits a positive correlation with all other emotions, the Depression group exhibits more fluctuation. We believe that such observations are worthy and useful to the community. Through studies of such nature, we can better understand how certain emotional patterns arise as a result of emotional expression.

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
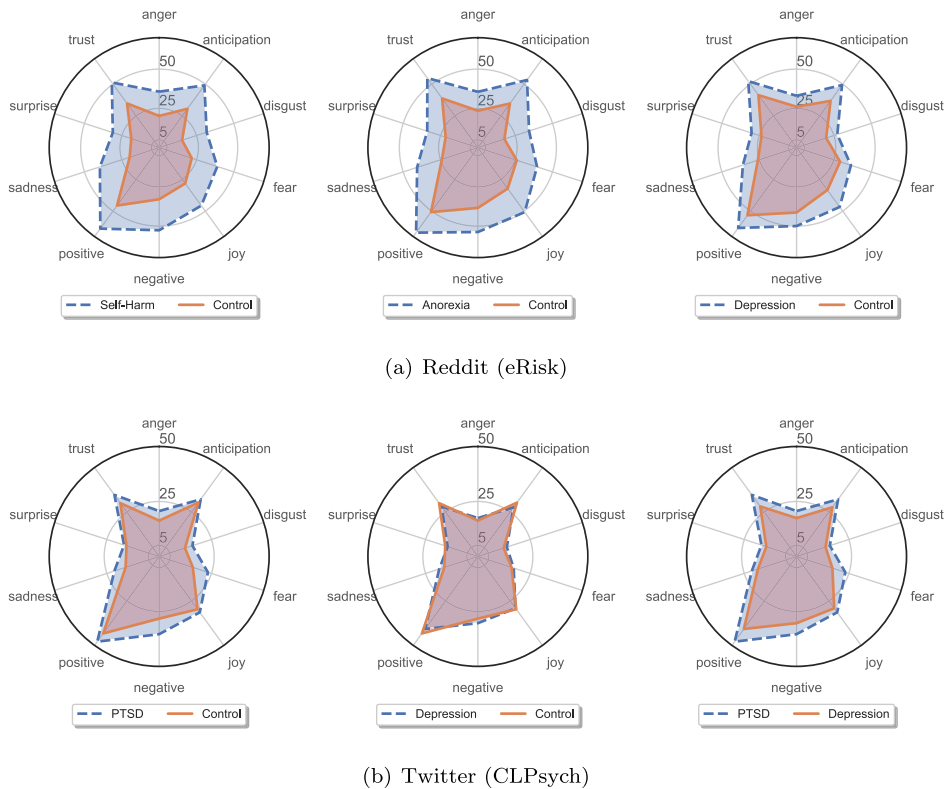
(a) Reddit (eRisk)

(b) Twitter (CLPsych)

Fig. 3. Graph depicting the average number of documents containing one or more words associated with a particular emotion, as well as the polarity (positive vs. negative) for each positive group and its control.
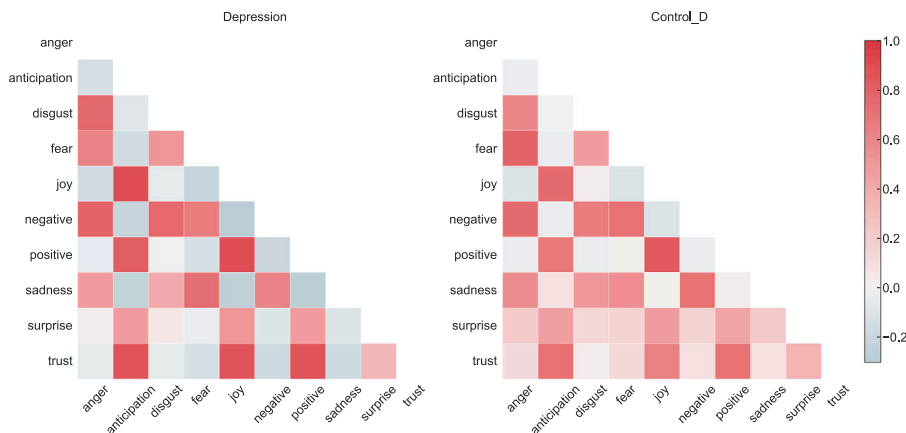

Fig. 4. Heatmap depicting the frequency correlation of the different emotions for positive (left) and control (right) groups on Reddit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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.

5.4. Social engagement and posting trends

In Table 1 we present various statistics about the collections studied in this research. Overall, we observe that positive users on Reddit create on average fewer documents, though longer, than control users. The case of users with anorexia is particularly interesting. Their documents are even longer than those written by users who are affected by self-harm or depression. It is good to
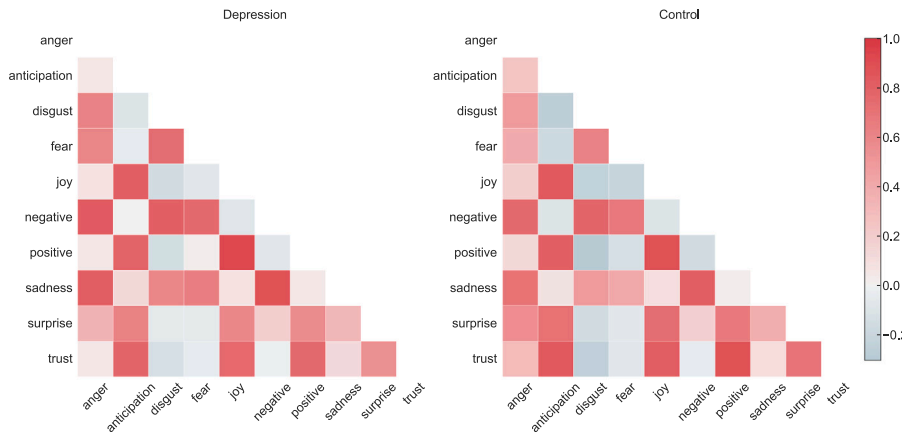
Fig. 5. Heatmap depicting the frequency correlation of the different emotions for positive (left) and control (right) groups on Twitter. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

out-stand that we did not identify meaningful differences among the various control groups. This trend is consistent over various experiments and it is expected as each control group entails a random sample of Reddit posts at different temporal periods. Hence, they constitute a representative sample of Reddit users.

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.

Fig. 6 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 ($\approx 1/\%$), meaning that out of ten words one is a hashtag. While for PTSD and control users out of ten words two are hashtags ($\approx 2\%$). 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 ($\approx 7\%$). Compared to depressed and control individuals, on average PTSD users tend to include more mentions in their posts, but the frequency of mentions for the three groups in each post is practically equal.

An appealing feature of the collections studied in this work 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. Fig. 7 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 Fig. 7(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

---

[16] Inflicting some sort of pain can bring a momentary sense of calm to the negative feelings experienced by the individual. Although, 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.
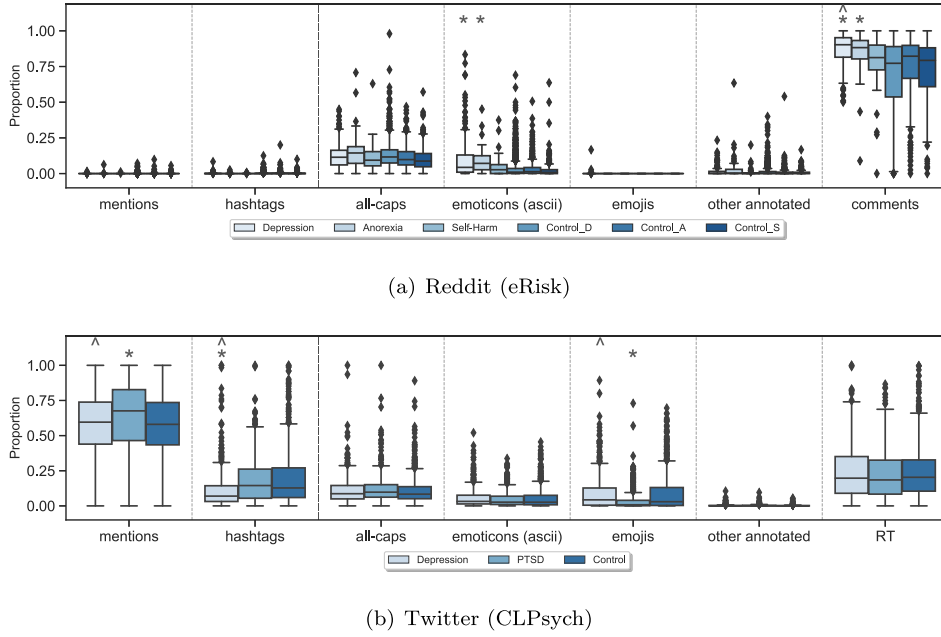
(a) Reddit (eRisk)

(b) Twitter (CLPsych)

Fig. 6. 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).

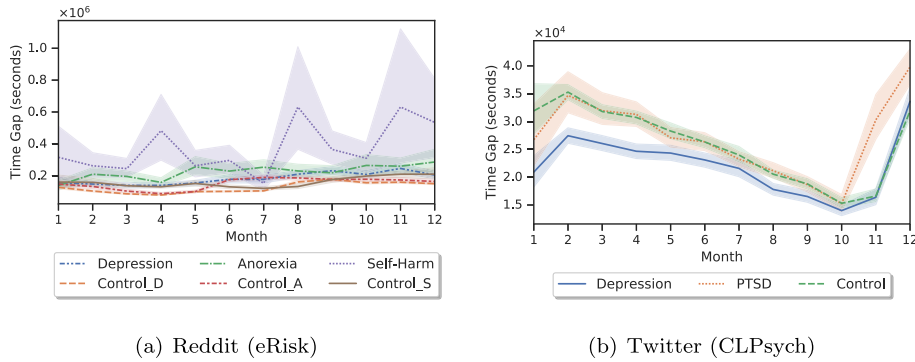(a) Reddit (eRisk)

(b) Twitter (CLPsych)

Fig. 7. Time-gap analysis by month.

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 Fig. 7(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.

As observed on the time-gap analysis conducted certain behavioural patterns related to specific periods or events that take place during the year, e.g., seasons change or Christmas and New year's eve emerged. Moreover, we also have an intuition that there could be some individuals who experience seasonal affective disorder (SAD) throughout the year. Many people go through short periods of time where they feel sad or not like their usual selves. Sometimes, these mood changes begin and end when the seasons change. This can be more acute for individuals who are already suffering from depression or other mental disorders. However, as posts are not geo-located, we cannot directly identify such cases yet some patterns still manifest. The fact that there are certain periods of the year in which different trends emerge, such as a higher posting volume at specific intervals, suggests that some event/s negatively affected users during those periods. We believe that this information could be useful for practitioners and medical agencies to launch prevention campaigns and encourage individuals to visit a professional to talk or discuss about their mood, feelings and

mental condition during that time. This could foster interventions at early stages and avoid any further consequences. In addition, it could prevent the emergence of new cases. In fact, as observed on the plot, control users on Twitter exhibit a similar behaviour to depressed users towards the end of the year. This could be the result of family events or even the lack of them and hence it could negatively affect individuals mental health. On some users, this could trigger over time other episodes. Finally, if automatic systems are used to identify potential cases, their thresholds and decision rules could become more *sensitive* during these periods and, thus, raise a larger volume of alarms.

6. Implications, conclusions, and future work

Continuously generated content on social media yields new opportunities for innovative analysis on mental health concerns. Social media data is produced in a natural way by users' own willingness and therefore, has become a complement of great value for traditional assessment instruments (like self-assessment inventories) which are employed to establish the likeliness of occurrence of mental health concerns. Unlike most of the work in the area that aims at predicting the mental state and solving such a complex problem without the help of a practitioner, we follow our previous work (Ríssola et al., 2020), focusing on providing useful means of diagnosing mental health patients by the practitioners. In particular, we proposed novel ways of consuming and visualising the vast amount of textual and behavioural data to extract indicative cues to the practitioners. In this work, we extended our initial findings (Ríssola et al., 2020) on the Reddit platform to Twitter and studied how our analysis can be generalised to other social media platforms. Moreover, we focused not only on the textual data, but also behavioural data on both social media platforms.

Outcomes from our thorough analysis revealed significant differences in writing and behavioural patterns on social media platforms between individuals with mental disorders and control users. The visualisation and study of certain probabilistic attributes enabled us to detect trends in affected users' writing style, emotional expression, and online behaviour on social media. In particular, we observed and researched the vocabulary, psychometric and emotional indicators, and activity of different users on two social media platforms of diverse characteristics. While analysing and visualising these attributes, we found interesting differences that could be helpful for the development of diagnosis systems, and, specially, to provide health professionals with a more comprehensive way to assess individuals' mental state. However, we were not able to find any significant and distinctive indicators among the different mental disorders to differentiate from one another. We obtained evidence suggesting that the use of language in micro-blogging platforms, like Twitter, in which users are subject to constraints influencing their writing style, is less distinguishable between users suffering a mental disorder than other less restrictive platforms, like Reddit. As a result, we come to the conclusion that social media posts could be analysed to automatically detect individuals who hold a higher probability of being diagnosed with a mental disorder. Yet, specifying the disorder entails higher complexity and that is the reason why the judgement of an expert may be required.

Implications resulting from our work serve as a starting point for future research dedicated to understanding how mental health concerns are manifested on social media. The various analysis techniques presented in this work and the information they provide could be integrated with more conventional direct and indirect assessment instruments used by mental health practitioners to create a more comprehensive cross-sectional characterisation of the persons' mental condition. As shown throughout the article, when the mental disorder of users develop the way they write and express detaches from the *standard* behaviour. Once potentially affected users are identified, a report could be generated to estimate how much their writing differs from healthy individuals in an attempt to characterise their condition. For instance, one dimension to consider in such report could be the close-vocabulary analysis presented in Section 5.2. If the users consistently show that they have increased the usage of words in the selected LIWC and Empath categories, this could provide stronger evidences that they are affected by a mental disorder. Similar scenarios could be described considering the rest of the dimensions studied in this work.

We plan to investigate various directions as future work. Firstly, we aim to expand our study to multiple datasets collected from different social media platforms to provide more insights into the behaviour of users with various mental issues on different social media platforms. Moreover, this would allow us to conduct a thorough analysis on the impact of different datasets on the results that we obtain. To do so, we aim to characterise each dataset with a number of features such as having length limitations and see how such limitations or differences across social media platforms would affect users' behaviour. Secondly, while in this work we have mainly studied the lexical nature of social media posts, in the future we plan to extend to semantic models such as word2vec (Mikolov et al., 2013) and contextual embeddings (Devlin et al., 2018), as well as incorporating multimodal information in our analysis such as multimedia context (e.g. images and videos) (Gkoumas et al., 2021). Finally, we plan to study the effectiveness of the analyses we performed in this work on machine learning approaches by using them as features to predict the individuals with mental health issues.

CRediT authorship contribution statement

Esteban A. Ríssola: Conceptualization, Methodology, Software, Writing – original draft. **Mohammad Aliannejadi:** Conceptualization, Methodology, Writing – review & editing. **Fabio Crestani:** Supervision, Validation, Writing – review & editing.

Acknowledgements

References

Al-Mosaiwi, M., & Johnstone, T. (2018). 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.

Aliannejadi, M., & Crestani, F. (2020). Venue suggestion using social-centric scores. In *Bias and social aspects in search and recommendation - first international workshop* (pp. 127–142).

Amini, H., & Kosseim, L. (2020). Towards explainability in using deep learning for the detection of anorexia in social media. In *Natural language processing and information systems - 25th international conference on applications of natural language to information systems* (pp. 225–235).

Aragon, M. E., Lopez-Monroy, A. P., Gonzalez-Gurrola, L.-C. G., & Montes, M. (2021). Detecting mental disorders in social media through emotional patterns - the case of anorexia and depression. *IEEE Transactions on Affective Computing*.

Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington: American Psychiatric Publishing.

Bathina, K. C., ten Thij, M., Lorenzo-Luaces, L., Rutter, L. A., & Bollen, J. (2021). Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*, 5(4), 458–466.

Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. In *Proceedings of the ninth international conference on web and social media* (pp. 31–40).

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

Cacheda, F., Fernandez, D., Novoa, F. J., & Carneiro, V. (2019). Early detection of depression: Social network analysis and random forest techniques. *Journal of Medical Internet Research*, 21(6), Article e12554.

Chancellor, S., & Choudhury, M. D. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *Npj Digital Medicine*, 3(1), 43.

Choi, J., & Croft, W. B. (2012). Temporal models for microblogs. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 2491–2494).

Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the seventh international conference on weblogs and social media*.

Choudhury, M. D., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). 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* (pp. 2098–2110).

Chung, C., & Pennebaker, J. (2007). The psychological functions of function words. In *Frontiers of social psychology*, (pp. 343–359). New York, NY, USA: Psychology Press.

Coppersmith, G., Dredze, M., & Harman, C. (2014a). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality* (pp. 51–60).

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). 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* (pp. 1–10).

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). 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* (pp. 31–39).

Coppersmith, G., Harman, C., & Dredze, M. (2014b). Measuring post traumatic stress disorder in Twitter. In *Proceedings of the 8th international conference on weblogs and social media*.

Correia, R. B., Wood, I. B., Bollen, J., & Rocha, L. M. (2020). Mining social media data for biomedical signals and health-related behavior. *Annual Review of Biomedical Data Science*, 3(1), 433–458.

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

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Mitchell, L., Harris, K. D., Kloumann, I. M., Bagrow, J. P., Megerdoomian, K., McMahon, M. T., Tivnan, B. F., & Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8), 2389–2394.

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 conference on human factors in computing systems* (pp. 4647–4657).

Gaur, M., Kursuncu, U., Alambo, A., Sheth, A., Daniulaityte, R., Thirunarayan, K., & Pathak, J. (2018). "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* (pp. 753–762).

Gkotsis, G., Oellrich, A., Hubbard, T. J., Dobson, R. J., Liakata, M., Velupillai, S., & Dutta, R. (2016). 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* (pp. 63–73).

Gkoumas, D., Li, Q., Lioma, C., Yu, Y., & Song, D. (2021). What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66, 184–197.

Gligorić, K., Anderson, A., & West, R. (2018). 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* (pp. 596–599).

Gligorić, K., Anderson, A., & West, R. (2019). Causal effects of brevity on style and success in social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–23.

Khodabakhsh, M., Fani, H., Zarrinkalam, F., & Bagheri, E. (2018). Predicting personal life events from streaming social content. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 1751–1754).

Kloumann, I. M., Danforth, C. M., Harris, K. D., Bliss, C. A., & Dodds, P. S. (2012). Positivity of the english language. *PLoS One*, 7(1), 1–7.

Losada, D. E., Crestani, F., & Parapar, J. (2018). Overview of eRisk: Early risk prediction on the internet. In *Experimental IR meets multilinguality, multimodality, and interaction - 9th international conference of the CLEF association* (pp. 343–361).

Losada, D. E., Crestani, F., & Parapar, J. (2019). Overview of eRisk 2019 early risk prediction on the internet. In *Experimental IR meets multilinguality, multimodality, and interaction - 10th international conference of the CLEF association* (pp. 340–357).

Masood, R. (2019). Adapting models for the case of early risk prediction on the internet. In *Advances in information retrieval - 41st European conference on IR research* (pp. 353–358).

Metzler, D., Cai, C., & Hovy, E. (2012). Structured event retrieval over microblog archives. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 646–655).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th international conference on neural information processing systems* (pp. 3111–3119). Lake Tahoe, USA.

Mohammad, S. (2018). Word affect intensities. In *Proceedings of the eleventh international conference on language resources and evaluation*.

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

Nepomnyachiy, S., Gelley, B., Jiang, W., & Minkus, T. (2014). What, where, and when: Keyword search with spatio-temporal ranges. In *Proceedings of the 8th Workshop on Geographic Information Retrieval*.

Neuman, Y. (2016). *Computational personality analysis: introduction, practical applications and novel directions*. Springer.

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in Twitter. In *Proceedings of the ACM SIGKDD workshop on healthcare informatics*.

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

Plutchik, R. (1980). Emotion: Theory, research, and experience: Vol. 1. Theories of emotion. In R. Plutchik, & H. Kellerman (Eds.), *Approaches to emotion* (pp. 3–33). Academic Press.

Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: A good place to detect health conditions. *PLoS One*, *9*(1), 1–11.

Ramírez-Cifuentes, D., Freire, A., Baeza-Yates, R., Puntí, J., Medina-Bravo, P., Velazquez, D. A., Gonfaus, J. M., & Gonzàlez, J. (2020). Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis. *Journal of Medical Internet Research*, *22*(7), Article e17758.

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports*, *7*(1), 1–11.

Ríssola, E. A., Aliannejadi, M., & Crestani, F. (2020). Beyond modelling: Understanding mental disorders in online social media. In *Advances in information retrieval - 42nd European conference on IR research* (pp. 296–310).

Ríssola, E. A., Bahrainian, S. A., & Crestani, F. (2019a). Anticipating depression based on online social media behaviour. In *Flexible query answering systems - 13th international conference* (pp. 278–290).

Ríssola, E. A., Bahrainian, S. A., & Crestani, F. (2019b). Personality recognition in conversations using capsule neural networks. In *2019 IEEE/WIC/ACM international conference on web intelligence* (pp. 180–187).

Ríssola, E. A., Losada, D. E., & Crestani, F. (2021). 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.

Ríssola, E. A., & Tolosa, G. H. (2016). Improving real time search performance using inverted index entries invalidation strategies. *Journal of Computer Science and Technology*, *16*(1), p. 6–13.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012). EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of the eighth international conference on language resources and evaluation* (pp. 3806–3813).

Saha, K., Seybolt, J., Mattingly, S. M., Aledavood, T., Konjeti, C., Martinez, G. J., Grover, T., Mark, G., & De Choudhury, M. (2021). What life events are disclosed on social media, how, when, and by whom? In *Proceedings of the 2021 CHI conference on human factors in computing systems*.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One*, *8*(9), Article e73791.

Skaik, R., & Inkpen, D. (2020). Using social media for mental health surveillance: A review. *ACM Computing Surveys*, *53*(6), 1–31.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction*, *27*(5), 1–53.

ten Thij, M., Bathina, K., Rutter, L. A., Lorenzo-Luaces, L., van de Leemput, I. A., Scheffer, M., & Bollen, J. (2020). Depression alters the circadian pattern of online activity. *Scientific Reports*, *10*(1), 1–10.

Uban, A.-S., Chulvi, B., & Rosso, P. (2021). An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, *124*, 480–494.

Walsh, C. G., Chaudhry, B., Dua, P., Goodman, K. W., Kaplan, B., Kavuluru, R., Solomonides, A., & Subbian, V. (2020). Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. *JAMIA Open*, *3*(1), 9–15.

Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2968–2978).

Yoo, M., Lee, S., & Ha, T. (2019). Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit. *Information Processing & Management*, *56*(4), 1565–1575.

Zarrinkalam, F., Kahani, M., & Bagheri, E. (2018). Mining user interests over active topics on social networks. *Information Processing & Management*, *54*(2), 339–357.