The language of mental health problems in social media

George Gkotsis[1,†], **Anika Oellrich**[1,†], **Tim Hubbard**[2], **Richard JB Dobson**[1,3]
Maria Liakata[4], **Sumithra Velupillai**[1,5], **Rina Dutta**[1]
[1]King's College London, IoPPN, London, SE5 8AF, UK; e-mail: george.gkotsis@kcl.ac.uk
[2] King's College London, Guy's Hospital, London, SE1 9RT, UK
[3]Farr Institute of Health Informatics Research, London, WC1E 6BT, UK
[4]University of Warwick, Department of Computer Science, Warwick, CV4 7AL, UK
[5]School of Computer Science and Communication, KTH, Stockholm
[†] Authors contributed equally to this work.

Abstract

Online social media, such as Reddit, has become an important resource to share personal experiences and communicate with others. Among other personal information, some social media users communicate about mental health problems they are experiencing, with the intention of getting advice, support or empathy from other users. Here, we investigate the language of Reddit posts specific to mental health, to define linguistic characteristics that could be helpful for further applications. The latter include attempting to identify posts that need urgent attention due to their nature, e.g. when someone announces their intentions of ending their life by suicide or harming others. Our results show that there are a variety of linguistic features that are discriminative across mental health user communities and that can be further exploited in subsequent classification tasks. Furthermore, while negative sentiment is almost uniformly expressed across the entire data set, we demonstrate that there are also condition-specific vocabularies used in social media to communicate about particular disorders. Source code and related materials are available from: `https://github.com/gkotsis/reddit-mental-health`.

1 Introduction

Mental illnesses are estimated to account for 11% to 27% of the disability burden in Europe (Wykes et al., 2015) and mental and substance use disorders are the leading cause of years lived with disability worldwide (Whiteford et al., 2013). Our knowledge about these mental health problems is still more limited than for many physical conditions, as sufferers may relapse even after successful treatment or exhibit resistance to different treatments. Most mental health conditions begin early, disrupt education (Kessler et al., 1995) and may persist over a lifetime, causing disability when those affected would normally be at their most productive (Kessler and Frank, 1997). For example, Patel and Knapp (1997) estimated the aggregate costs of all mental disorders in the United Kingdom at 32 billion (1996/97 prices), 45% of which was due to lost productivity (Patel and Knapp, 1997). The global burden of mental and substance use disorders increased by 376% between 1990 and 2010 (Whiteford et al., 2013) which means it is an international public health priority to effectively prevent and treat mental health issues.

In the UK, 17% of adults experience a subthreshold common mental disorder (McManus et al., 2009) and up to 30% of individuals with non-psychotic common mental disorders have subthreshold psychotic symptoms (Kelleher et al., 2012) showing that a large proportion of mental illness is unrecognised, but nevertheless has a significant impact upon people's lives. Those people with conditions that meet criteria for diagnosis are treated in primary care or by mental health professionals.

Studies consistently show that between 50-60% of all individuals with a serious mental illness receive treatment for their mental health problem at any given time (Kessler et al., 2001).

However, most of the pathology tracking and improvement assessment is done through questionnaires, e.g. the Personal Health Questionnaire 9 (PHQ9) for depression (Kroenke and Spitzer, 2002), and require a subjective comment by the patient, e.g. "How many days have you been bothered with little interest or pleasure in doing things in the past two weeks?". As with every personal judgement, the responses are influenced by the environment in which the person has been asked, the relationship to the clinician and even the stigma attached to depression (Malpass et al., 2010). While there are aims to integrate real-time reporting into a patient's life (Ibrahim et al., 2015), these are still based on set questionnaires and may not fit with the main concerns of a patient.

Social media, such as Twitter[1], Facebook[2] and Reddit[3], have become an accepted platform to communicate about life circumstances and experiences. A specific example of social media in the context of illness is PatientsLikeMe (Wicks et al., 2010). PatientsLikeMe has been developed to enable people suffering from an illness to exchange information with others with the same condition, e.g. to find alternative treatment opportunities. It has been shown that the support received in such online communities can be empowering by engendering self-respect and a feeling of being in control of the situation (Barak et al., 2008). Hence, social media constitutes a tremendous resource for better understanding diseases from a patient perspective.

Social media data has recently been recognised as one of the resources to gather knowledge about mental illnesses (Coppersmith et al., 2015a; De Choudhury et al., 2013; Kumar et al., 2015). For example, Twitter data has been used to develop classifiers to recognise depression in users (De Choudhury et al., 2013) and to classify Twitter users who have attempted suicide from those who have not and from those who are clinically depressed (Coppersmith et al., 2015b). Furthermore, data collected from Reddit pertaining to suicidal ideation could demonstrate the existence of the Werther effect (suicide attempts and completions after media depiction of an individual's suicide) (Kumar et al., 2015). Coppersmith and colleagues used Twitter data to determine language features that could be used to classify Twitter users into suffering from mental health problems and unaffected individuals (Coppersmith et al., 2015a). However, while the authors could identify features that allows the classification between healthy and unhealthy Twitter users, they also note that language differences in communicating about the different mental health problem remains an open question. Similarly, Mitchell et al. (2015) used Twitter data to separate users affected by schizophrenia from healthy individuals by automatically identifying characteristic language features for schizophrenia (Mitchell et al., 2015). Both the latter approaches rely on the Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010), but Mitchell et al. also covers features such as Latent Dirichlet Allocation and Brown Clustering. Very recently and concurrently with our own work, De Choudhury and colleagues have shown that linguistic features can be used to predict the likelihood of individuals transitioning from posting about depression and other mental health issues on Reddit to suicidal ideation (De Choudhury et al., 2016). This work showed the ability to make causal inferences on the basis of language usage and employed a small subset of the mental health groups on Reddit.

Following on from the promise that such work holds, our goal was to study language features that are characteristic for individual mental health conditions using **large scale** Reddit data. We anticipate that our findings can be used to assist in separating posts pertaining to different mental health problems and for various language-based applications involving the better understanding of mental health conditions. Reddit is particularly suitable for such research as it has an enormous user base[4], posts and comments are topic-specific and the data is publicly available. We focussed on subreddits (communities where users can post/comment in relation to a specific topic, e.g. suicide ideations)

[1] `https://twitter.com/?lang=en`

[2] `https://en-gb.facebook.com/`

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

[4] According to `https://en.wikipedia.org/wiki/Reddit`, Reddit had 234M unique users with 542M monthly visitors as of 2015

of the Reddit data dump[5] that address the following mental health problems: Addiction, Anxiety, Asperger's, Autism, Bipolar Disorder, Dementia, Depression, Schizophrenia, self harm and suicide ideation. These conditions are commonly encountered by mental health practitioners and contribute significantly to treatment costs. We aimed to identify linguistic characteristics that are specific to any of the mental illnesses covered and can be used for text classification tasks. The investigated characteristics include lexical as well as syntactic features, the uniqueness of vocabularies, and the expression of sentiment and happiness. Our results suggest that there are linguistic features that are discriminative of the user communities used in this study. Furthermore, applying a clustering method on subreddits, we could show that subreddits mostly contain a topic-specific vocabulary. Moreover, we could also highlight that there are differences in the way that sentiment is expressed in each of the subreddits. Source code and related materials are available from: `https://github.com/gkotsis/reddit-mental-health`.

2 Methods and Materials

As our aim was to define linguistic characteristics specific to mental health problems, we downloaded the Reddit data and extracted relevant posts and comments. These were then further investigated with respect to specific linguistic features, e.g. sentence structure or unique vocabularies, to determine characteristics for subsequent classification tasks. The data set as well as the methods employed are described in the following subsections.

2.1 Social media data from Reddit

Reddit is a social media network, where registered users can post requests to a broader community. Posts are hosted in topic-specific fora, so called *subreddits*. Subreddits can be created by users based on the subject they are interested in to communicate. All users can freely join any number of subreddits and participate in discussions. This means that the posts are sent to a community potentially knowledgeable or at least interested in the topic. We used this Reddit feature, to determine subreddits targetting specific mental health problems.

For this purpose, we filtered the entire downloaded data set for subreddits targeting any of the 10 as relevant identified diseases. The entire data set as obtained was separated into posts and comments, and we preserved this separation so that analysis could be executed on either posts, comments or both combined. Posts are initial textual statements that initiate a communication with other users. Comments are replies to posts and are organised in a tree-like structure. Both posts and comments can be written be anyone, and even the Reddit user that wrote the initial post can comment on it. We note here that the number of users, posts and comments varied substantially between subreddits (see Table 1). To refer to sets of both posts and comments (total also in Table 1), we use the term "communication" in the following sections.

Table 1: Numbers of posts, comments, ratio of comments over posts, and the total of posts and comments (called "communications") for each mental health-related subreddit included in this study. Numbers are totalled across all subreddits in the last row of this table. Extrema for each column are highlighted with a purple coloured background.

subreddit	#posts	#comments	#comments/#posts	#total
Anxiety	57,523	289,441	5.03	346,964
BPD	11,880	77,091	6.49	88,971
BipolarReddit	14,954	151,588	10.14	166,542
BipolarSOs	814	4,623	5.68	5,437
OpiatesRecovery	8,651	87,038	10.06	95,689
StopSelfHarm	4,626	24,224	5.24	28,850
addiction	4,360	6,319	1.45	10,679
aspergers	15,053	202,998	13.49	218,051
autism	9,470	52,090	5.50	61,560
bipolar	25,868	198,408	7.67	224,276
cripplingalcoholism	38,241	503,552	13.17	541,793
depression	197,436	902,039	4.57	1,099,475
opiates	56,492	906,780	16.05	963,272
schizophrenia	4,963	31,864	6.42	36,827
selfharm	12,476	68,520	5.49	80,996
SuicideWatch	90,518	619,813	6.85	710,331
Total	462,807	3,506,575	7.58	3,969,382

As shown in Table 1, the *depression* subreddit contains the largest amount of communications (1.1M), while the smallest amount is found in the *BipolarSOs* subreddit (5K). The number of posts is always smaller than the number of comments though the ratio of average number of comments per posts varies. The highest average rate of comments per posts can be seen on subreddit *opiates*, while the smallest number of replies is observed on the *addiction* subreddit.

[5]The data was released by a Reddit user on `https://redd.it/3bxlg7` (comments) and `https://redd.it/3mg812` (posts).

2.2 Determining linguistic features

There are many ways to model communication. Communication in the form of language use can be characterised through a variety of feature types. Our aim is to better understand the nature and depth of the communication that takes place, and one way to do this is by the analysis of linguistic features. These features are particularly relevant in the context of mental health problems, as the abilities of the sufferer to effectively communicate can be affected by such problems (Cohen and Elvevåg, 2014). For example, someone suffering from Bipolar Disorder may suddenly write a lot, but not necessarily in a cohesive manner. In the Iowa Writers' Workshop study (Andreasen, 1987) bipolar sufferers reported that they were unable to work creatively during periods of depression or mania. During depressive episodes, cognitive fluency and energy were decreased, and during manic periods they were too distractible and disorganized to work effectively, so it would be reasonable to expect this to be reflected in their prose. Understanding these features and consequently the nature and content of the posts will allow us to better design useful classification systems and predictive models.

Through discussion, we determined an initial feature set of linguistic characteristics that draws on previously established measures of psychological relevance, such as LIWC and Coh-Metrix (Graesser et al., 2004). However, we note here that in order to not overload our initial feature set, we selected a subset of all the available possibilities. In our feature set, we included linguistic features introduced by Pitler and Nenkova (Pitler and Nenkova, 2008) and partially overlapping with those used in Coh-Metrix for predicting text quality. More specifically, we adopt features that aim at assessing the *readability* of textual content. Readability is a measurement that aims to assess the required education level for a reader to fully appreciate a certain content. The task of understanding textual content and assessing its quality encompasses various factors that are captured through the features that we also propose here (see supplemental material for more information about the implementation). A subset of these features have been used successfully to predict the answers to be marked as accepted in online Community-based Question Answering websites (Gkotsis et al., 2014).

Our first set of features pertains to the usage of specific words in documents. For instance, we look at the usage of definite articles, since we believe that definite articles are used for specific and personal communications. Similarly, we keep track of pronouns, first-person pronouns, and the ratio between them, as indicators of the degree of first-person content.

Additional features in this initial set aimed at examining text complexity. In our approach, text complexity can scale both horizontally (length, topic cohesion) and vertically (clauses, composite meanings). For the horizontal assessment, we count the number of sentences. Another set of features, which target the understanding of topic continuity and cohesion across sentences, is word overlap between adjacent sentences, either by taking into account all words, or just nouns and pronouns. For the vertical assessment, we employ the following features: a) we count the noun chunks and verb phrases (sequences of nouns and verbs, respectively) and the number of words contained within them, b) we construct the parse tree of each sentence and measure its height, and c) we count the number of subordinate conjunctions (e.g. "although", "because" etc.). A parse tree represents the syntactic structure of a sentence, and tools such as dependency or constituency parsers are readily available for utilisation, e.g. as implemented in the Python module spaCy[6].

Finally, we found that a few posts do not contain any text in their body, apart from their title. This was typically the case for posts that contained a Uniform Resource Locator (URL) to a web page of interest to the community. We believe that the ratio of the number of these posts over the total number of posts is associated with the degree of information dissemination[7], as opposed to the personal story-telling that might occur in other cases, and thus included this as an additional feature.

[6] https://spacy.io/

[7] For instance, we found that most URLs posted in *addiction* link to YouTube

2.3 Word-based classification to assess subreddit uniqueness

For this classification-based approach, we employed a representation based on individual words, as well as information on words that frequently co-occurred together. The aim of this task was to examine how closely aligned the vocabularies of each subreddit were, assessed via a pairwise comparison. As highlighted in Table 1, the data volume (in terms of posts and comments) differed significantly for the different subreddits. In order to compensate for the difference in size, we utilised a randomisation process by repeating the same experiment 10 times with a set of 5000 randomly drawn posts for each repeat and individually for each of the two subreddits that were compared with each other.

In order to compare the vocabularies of two subreddits with each other, we built dictionaries for each pair of subreddits, by retrieving all words and sequences of words (of length 2) occurring in one or both subreddits. We then used this list of words and frequently co-occurring words to classify posts into belonging to one of the two subreddits that are being compared and recorded the performance for each of the 10 cycles for each subreddit pair. The classification performance was then averaged across all 10 cycles to obtain a representative score for each pair of subreddits. Using this classification approach, high performance scores indicate a distinctive vocabulary while low performance scores suggest a shared vocabulary across both the subreddits. The results of this pairwise comparison are illustrated in Figure 1. More details are provided in the supplementary materials, covering the algorithm and randomisation steps.

2.4 Detecting sentiment and happiness in posts

One additional aspect that can be assessed when looking at the linguistic aspect of communications on social media is the expression of sentiment. Sentiment has been noted as a crucial indicator of how much involved someone is in a specific event (Tausczik and Pennebaker, 2010; Murphy et al., 2015), and therefore can also play a role in the expression of mental illness. Some of the conditions investigated here may have characteristic mood patterns, e.g. it is likely that someone suffering from depression will use negative sentiment and express unhappiness, while someone suffering from Bipolar Disorder may change between positive and negative mood expressions over time. However, by assessing sentiment and happiness for a large population of individuals, novel patterns for individual mental health problems may evolve.

As part of our investigation, we used two different methods, one to detect sentiment (Nielsen, 2011) and another to detect happiness (Dodds et al., 2011). Both methods, which were developed for social media studies, rely on a topic-specific dictionary. For each *post* in our subreddits, we determined the sentiment and happiness score by matching words against the dictionaries. We accumulated these scores on a per post basis and normalised it by the square root of the number of words in the post that were identified in the respective dictionary. Scores for happiness were further normalised to assign them to the same range as the values for sentiment: negative values are expressions of negative sentiment/unhappiness, positive values are expressions of positive sentiment and happiness, and a value of 0 can be seen as neutral.

We note here that while our aim is to classify both posts and comments, we limited ourselves in this task to posts only. Comments could be considered to be a source of noise, which may mask potential sentiment and happiness coming from posts, given that our data set contains a lot more comments than posts. In future work we would like to experiment with more sophisticated linguistic methods for identifying sentiment and emotion.

3 Results

After identifying the subreddits relevant to the mental health problems we were interested in, we determined linguistic features (related to content of communications) for each of the subreddits. The results of our investigations are presented in the following subsections.

3.1 Subreddits exhibit differences in linguistic features

Table 2 provides a summary of all the linguistic features for each of the 16 subreddits, that were assessed as part of this study. From this table, we

see that two subreddits stand out in a number of the assessed criteria: *BiPolarSOs* and *cripplingalcoholism*. *BiPolarSOs* is a subreddit that provides support and advice to people in a relationship where either one or both partners are affected by Bipolar Disorder. Note that this means that users on this subreddit may not be affected by the disorder themselves and may result in different communications from a subreddit where only people with Bipolar Disorder are communicating. In our data set it was the smallest subreddit in terms of total number of communications (see Table 1). The subreddit *cripplingalcoholism* aims to facilitate communication between people addicted to alcohol. In the description of the subreddit, there is no emphasis on supporting each other and people can also share what they may consider positive experiences regarding their condition (e.g. "On day 8 of a bender that was supposed to end today because my boss was supposed to send me a bunch of work on Monday. She just emailed me and said she won't be sending it until Wednesday! Sweet chocolate Jesus on a bicycle, I did a jig in my jammies, cracked open a new handle of rye, and am about to take the dog on a nice drunken walk. Sobriety, I'll see you Wednesday. Maybe").

From Table 2, we see that the *BipolarSOs* subreddit not only has a higher number of first-person pronouns and a larger number of definite articles, but also that the average sentence seems to be more complex due to a high average height of the sentence parse trees, long verb clauses and a high number of subordinating conjunctions, while the average number of sentences per communication is comparable to those of the other subreddits. This suggests that people posting on this subreddit explain in detail their experience or advice. On the contrary, the *cripplingalcoholism* subreddit possesses shorter communications characterised by the lowest number of sentences per communication, the smallest maximum height of sentence trees, a low number of subordinate conjunctions and short verb clauses. Using word frequency occurrences, we also observed that here the language seems stronger than on other subreddits with the most frequently occurring word being "fuck" (details of results not provided here[8]).

The two features relating to lexical cohesion (by means of adjacent sentences using similar words, LF10 and LF11 in Table 2), show little variation across all the 16 different subreddits. Though cohesion when only taking nouns and pronouns into consideration improves, the best value obtained is 0.22, indicating a mostly low lexical cohesion across communications on each of the subreddits. One of our longer term goals is to be able to classify posts to individual subreddits, and these scores would not be sufficiently informative for this goal due to their low variation.

3.2 Subreddit vocabulary uniqueness through classification

The word occurrence-based comparison of the subreddits was performed to better determine whether subreddits can be distinguished based on their lexical content (see supplementary material for more information). The results obtained are shown in Figure 1.

Figure 1: Heatmap of pairwise classification of posts (*only*) between subreddits. High values denote high accuracy in classification and therefore represent high discriminability in language. Low values represent low score in classification and therefore high language proximity between subreddits.

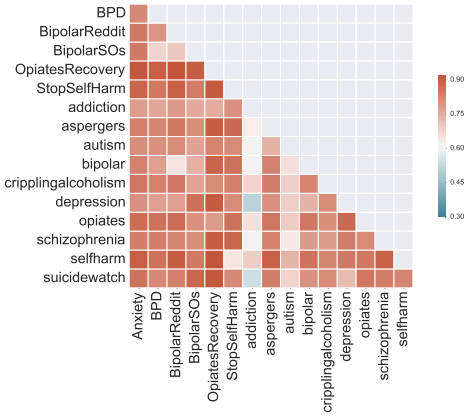


Figure 1 shows that apart from a small number of exceptions, the language of individual subreddits is discriminable, which can be further exploited for classification purposes in later stages. For example, the subreddit *OpiateRecovery* shows mostly high values, which means that the language used (based on frequency of words and word pairs) on this

[8] Available as a wordcloud visualisation of all subreddits at `https://github.com/gkotsis/reddit-mental-health/tree/master/wordclouds`

Table 2: Obtained results for each of the language features per subreddit. Language features investigated were: LF1 – Average number of definite article "the" in each communication; LF2 – Average number of first person pronouns; LF3 – Average number of pronouns in each communication; LF4 – Average number of noun chunks; LF5 – Average number of length of maximum verb phrase in each communication; LF6 – Average number of subordinate conjunctions; LF7 – Average value of maximum height of sentences' parse trees; LF8 – Average number of sentences in each communication; LF9 – Average ratio of number of first person pronouns over total number of pronouns; LF10 – Similarity between adjacent sentences over nouns or pronouns *only* (lexical cohesion); LF11 – Similarity between adjacent sentences over *all* words (lexical cohesion); LF12 – Ratio of posts without any body text (containing only a title and a URL) over total number of posts.

	Not normalised features								Normalised			
Subreddit	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LF11	LF12
Anxiety	2.24	7.67	13.13	1.85	25.68	1.43	5.95	6.51	0.54	0.19	0.18	0.10
BPD	2.23	7.98	14.28	1.83	28.32	1.48	6.14	6.63	0.54	0.20	0.18	0.09
BipolarReddit	2.14	6.49	11.54	1.84	23.91	1.44	5.98	6.15	0.54	0.18	0.18	0.01
BipolarSOs	3.53	9.68	23.06	1.99	40.52	1.49	6.67	10.12	0.40	0.17	0.17	0.04
OpiatesRecovery	2.24	6.48	11.93	1.80	23.49	1.28	5.66	6.69	0.49	0.17	0.17	0.01
StopSelfHarm	1.60	5.90	11.71	1.77	20.60	1.25	5.43	5.52	0.46	0.22	0.19	0.12
addiction	1.95	5.54	10.66	1.35	21.50	0.98	4.30	5.93	0.44	0.17	0.18	0.63
aspergers	1.94	5.12	9.69	1.72	20.76	1.53	5.88	5.06	0.53	0.17	0.18	0.12
autism	2.07	3.62	8.65	1.61	19.89	1.37	5.50	5.01	0.41	0.13	0.17	0.61
bipolar	1.86	6.16	10.62	1.73	20.82	1.31	5.53	5.67	0.56	0.18	0.17	0.15
cripplingalcoholism	0.92	2.28	4.07	1.36	8.76	0.95	4.10	3.02	0.54	0.12	0.16	0.16
depression	2.25	8.71	14.75	1.84	29.04	1.37	5.89	7.17	0.52	0.21	0.19	0.02
opiates	1.14	2.60	5.11	1.48	10.96	1.13	4.52	3.25	0.49	0.14	0.16	0.21
schizophrenia	2.13	5.96	11.15	1.80	23.74	1.45	5.82	5.85	0.50	0.18	0.18	0.13
selfharm	1.39	5.41	9.73	1.70	17.57	1.19	5.11	4.94	0.52	0.21	0.18	0.01
suicidewatch	1.96	7.10	13.44	1.85	27.85	1.29	5.74	6.73	0.46	0.22	0.19	0.03

subreddit is mostly unique. *OpiateRecovery* shows some vocabulary overlap with the *opiates* and *addiction* subreddits, which suggests that there are some shared topics on these subreddits. One of the exceptions is the subreddit *addiction*. As illustrated in the heatmap the *addiction* subreddit shows particularly low values with other subreddits such as *depression* and *suicidewatch*. This finding is not surprising as substance addiction can lead to depression and suicidal thoughts, which is expected to be also expressed in the nature of the communication. Note that the diagonal of the matrix is suppressed to reduce the matrix dimension.

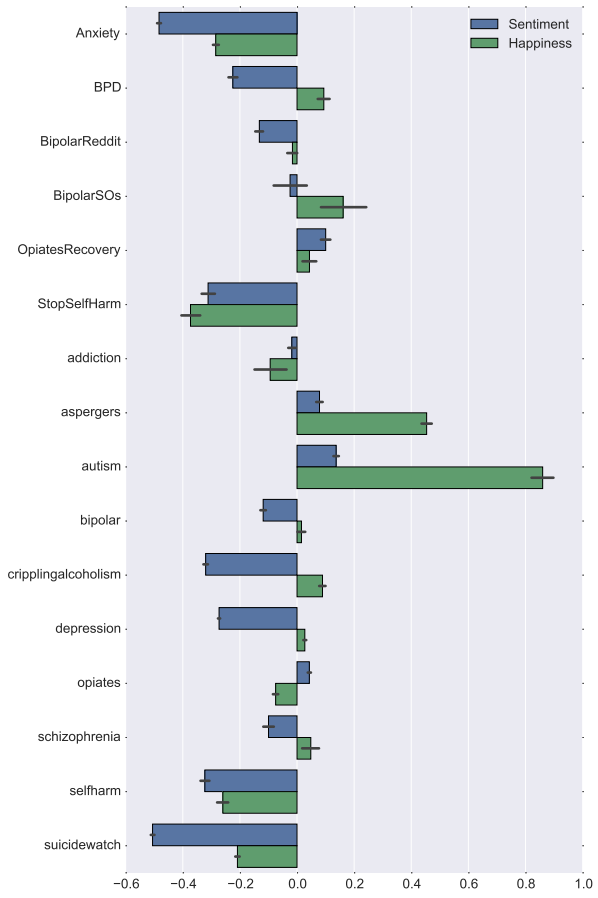
Among our 16 subreddits, there are some subreddits that allude to the same mental health condition, e.g. *BipolarReddit* and *BipolarSOs* both aim to foster a community to facilitate exchange about Bipolar Disorder. While the subreddit *BipolarSOs* invites participation from users that are affected themselves or are in a relationship with someone affected by Bipolar Disorder, *BipolarReddit* is solely focussed on people suffering from this disorder. In Figure 1, we can also see that vocabularies seem to be partially shared (indicated by a lighter colour) across those subreddits addressing the same mental health problem. For example, all three subreddits relating to Bipolar Disorder (*bipolar*, *BipolarReddit* and *BipolarSOs*) show a pairwise score of ~ 0.6 as opposed to ~ 0.9 with other subreddits. Similarly, both the self-harm subreddits also share a pairwise vocabulary of ~ 0.6. Interestingly, the subreddits *autism* and *schizophrenia* also indicate a proximity of the vocabularies and further investigations are required to assess the shared vocabularies.

3.3 Sentiment/happiness expressions on subreddits

In order to assess the emotions that Reddit users express on subreddits related to mental health problems, we used two different methods: (i) to assess sentiment and (ii) to specifically assess happiness. The results obtained by both methods are shown in Figure 2. This figure illustrates that, on average, a lot of negative sentiment is expressed across the different subreddits relating to mental health problems. We can see that posts from the subreddit *SuicideWatch* express the highest rate of negative sentiment, followed by posts from the *Anxiety* and self-harm-related subreddits.

While in the majority of cases both sentiment and

Figure 2: A small number of subreddits show a majority of positive sentiments while a large number of subreddits show predominantly negative sentiments. Positive values in this bar plot correspond to positive sentiment and happiness, while negative values indicate negative sentiments or unhappiness.



happiness expression possess the same direction (i.e. either positive or negative), in a number of subreddits this is not the case. For example, the subreddit *cripplingalcoholism* shows expressions of happiness as well as the expression of negative sentiment. As alluded to earlier, this particular subreddit includes people that see alcoholism as a lifestyle choice. Though there may be happiness expressions related to overcoming alcoholism, there are also happiness expressions relating to the glorification of alcoholism, e.g. "[...] At the bottom of this pile of clothes is a full pint! How it came to rest there I don't know, but thank you Taaka gods for your gift on this day. [...]".

Furthermore, Figure 2 shows a small number of subreddits, where posts seem to express positive sentiment. For example, the posts extracted from the subreddit *OpiatesRecovery* seem to express not only positive sentiment but also happiness. This particular subreddit aims to foster a comunity that focusses on helping each other get through opiate withdrawals and users can post their progress. While there are posts that discuss relapses, there are statements such as "[...] I'm happy to say the shivers/flashes/heebeegeebees are a lot, lot better. Not 100% gone, but gone enough. I can deal with flashes every 4-6 hours, cant deal with them every 15 minutes. [...]" to share the successes made during withdrawal. The results shown in this figure are average values, which means that subreddits that show an overall tendency to happiness and positive sentiment, may contain some posts including words of negative sentiment and unhappiness, e.g. "[...] Buying garbage from some ignorant thug to put into my fucking blood knowing how lethal it can be, but oh it couldn't happen to me. It's bizarre that after all this time of staying away I still can't fully grasp how fucking close to death I was every day. [...]" from *OpiateRecovery*.

4 Discussion

In our study, we analysed 16 different subreddits covering a range of mental health problems (see supplementary material for more details). In our selection, there are subreddits with overlapping content, e.g. *StopSelfHarm* and *selfharm*. We conducted an analysis based on a selection of linguistic features and found that most of the subreddits that are topic-unrelated, possess a unique vocabulary (in terms of words/word-pairs and the frequencies thereof) and discriminating lexical and syntactic features. We also observed differences in sentiment and happiness expressions, which can give further clues about the nature of a post.

As symptoms are shared across conditions and more so, some of the mental health problems are co-occurring (e.g. anxiety and depression), medications and treatment strategies are shared across the different illnesses, too. This, in consequence, means that part of the vocabulary and thoughts across the different subreddits are shared, making it harder to distin-

guish between the different subreddits and, consequently, the condition in question. Given the latter, it is even more surprising that the similarity matrix shown in Figure 1 shows a good separation of topic-specific vocabularies on subreddits.

With respect to the expression of sentiment and emotions, further work is needed. The methods applied here were developed based on Twitter data and further investigations are necessary to find the parts of the dictionary that are overlapping and an expert-guided assessment as to whether the recognised expressions are representative and meaningful in the context of mental health problems. A previous study has investigated how support is expressed in social media (Wang et al., 2015) and can be leveraged in future work to see whether similar support models hold true for the subreddits concerning mental health conditions. Moreover, the methods we have used so far are based on lexica, which lack contextual information. In future work, we plan to add more contextualised semantic methods for determining sentiment and emotions.

One limitation of the work presented here is that we did not include any subreddits that are unrelated to mental health. For example, we could have included a subreddit such as *Showerthoughts* into our subset to assess which of the features are unique to mental health problems only. However, this would require the definition of what is a truly unrelated subreddit and variety of topics so that the control set is not biased in itself. Furthermore, as our primary aim was to build a classifier that distinguishes several mental health problems based on the findings reported here, an implicit assumption is that a post is by default relevant to mental health conditions and does not need to be classified as such. Nevertheless, we plan to address this limitation in future work.

5 Conclusions

After extracting data from several subreddits pertaining to mental health problems, we investigated a subset of language features to determine discriminatory characteristics for each of the subreddits. Our results suggest that there are discriminatory linguistic features among subreddits, such as sentence complexity or vocabulary usage. We could also show that while mostly all subreddits relating to mental health problems possess highly negative sentiment, there are a number of subreddits, where positive sentiment and happiness can be observed in posts. However, in order to determine the most discriminative features between different mental health conditions, additional work is required continuing from the results shown here. In conclusion, these results pave the way for future work on classification of posts and comments concerning a mental health condition, which in turn could allow the assignment of urgency markers to address a specific communication.

Acknowledgments

RD is supported by a Clinician Scientist Fellowship from the Health Foundation in partnership with the Academy of Medical Sciences and RD and GG are funded by the e-HOST-IT research programme. AO and RJBD would like to acknowledge NIHR Biomedical Research Centre for Mental Health, the Biomedical Research Unit for Dementia at the South London, the Maudsley NHS Foundation Trust and Kings College London. RJBD's work is also supported by awards to establish the Farr Institute of Health Informatics Research, London, from the Medical Research Council, Arthritis Research UK, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Economic and Social Research Council, Engineering and Physical Sciences Research Council, National Institute for Health Research, National Institute for Social Care and Health Research, and Wellcome Trust (grant MR/K006584/1). SV is supported by the Swedish Research Council (2015-00359), Marie Sklodowska Curie Actions, Cofund, Project INCA 600398. ML would like to acknowledge the PHEME FP7 project (grant No. 611233).

The authors acknowledge infrastructure support from the National Institute for Health Research (NIHR). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

References

Nancy C Andreasen. 1987. Creativity and mental illness. *American Journal of Psychiatry*, 144(10):1288–1292.

Azy Barak, Meyran Boniel-Nissim, and John Suler. 2008. Fostering empowerment in online support groups. *Computers in Human Behavior*, 24(5):1867–1883.

Alex S Cohen and Brita Elvevåg. 2014. Automated computerized analysis of speech in psychiatric disorders. *Current Opinion in Psychiatry*, 27(3):203–209.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015a. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. *In Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, page 1.

Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. 2015b. Quantifying Suicidal Ideation via Language Usage on Social Media.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *In Proceedings of the 2013 International AAAI Conference on Weblogs and Social Media (ICWSM)*, page 2.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. *In Proceedings of 2016 Special Interest Group on Computer-Human Interaction (SIGCHI)*.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLOS ONE*, 6(12):e26752.

George Gkotsis, Karen Stepanyan, Carlos Pedrinaci, John Domingue, and Maria Liakata. 2014. It's all in the content: state of the art best answer prediction based on discretisation of shallow linguistic features. In *In Proceedings of the 2014 ACM conference on Web science*, pages 202–210. ACM.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Zina M Ibrahim, Lorena Fernández de la Cruz, Argyris Stringaris, Robert Goodman, Michael Luck, and Richard JB Dobson. 2015. A Multi-Agent Platform for Automating the Collection of Patient-Provided Clinical Feedback. In *In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 831–839. International Foundation for Autonomous Agents and Multiagent Systems.

Ian Kelleher, Helen Keeley, Paul Corcoran, Fionnuala Lynch, Carol Fitzpatrick, Nina Devlin, Charlene Molloy, Sarah Roddy, Mary C Clarke, Michelle Harley, et al. 2012. Clinicopathological significance of psychotic experiences in non-psychotic young people: evidence from four population-based studies. *The British Journal of Psychiatry*, 201(1):26–32.

Ronald C Kessler and Richard G Frank. 1997. The impact of psychiatric disorders on work loss days. *Psychological Medicine*, 27(04):861–873.

Ronald C Kessler, Cindy L Foster, William B Saunders, and Paul E Stang. 1995. Social consequences of psychiatric disorders, I: Educational attainment. *American Journal of Psychiatry*, 152(7):1026–1032.

Ronald C Kessler, Patricia A Berglund, Martha L Bruce, J Randy Koch, Eugene M Laska, Philip J Leaf, Ronald W Manderscheid, Robert A Rosenheck, Ellen E Walters, and Philip S Wang. 2001. The prevalence and correlates of untreated serious mental illness. *Health Services Research*, 36(6 Pt 1):987.

Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515.

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *In Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 85–94. ACM.

Alice Malpass, Alison Shaw, David Kessler, and Deborah Sharp. 2010. Concordance between PHQ-9 scores and patients experiences of depression: a mixed methods study. *British Journal of General Practice*, 60(575):e231–e238.

Sally McManus, Howard Meltzer, TS Brugha, PE Bebbington, and Rachel Jenkins. 2009. Adult psychiatric morbidity in England, 2007: results of a household survey.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. *In Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, page 11.

Sean M Murphy, Bernard Maskit, and Wilma Bucci. 2015. Putting Feelings into Words: Cross-Linguistic Markers of the Referential Process. *In Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, page 80.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

A Patel and M Knapp. 1997. The cost of mental health: report to the Health Education Authority. In *Centre for Economics of Mental Health, Institute of Psychiatry London, UK Working paper*.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 186–195. Association for Computational Linguistics.

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

Yi-Chia Wang, Robert E Kraut, and John M Levine. 2015. Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. *Journal of Medical Internet Research*, 17(4).

Harvey A Whiteford, Louisa Degenhardt, Jürgen Rehm, Amanda J Baxter, Alize J Ferrari, Holly E Erskine, Fiona J Charlson, Rosana E Norman, Abraham D Flaxman, Nicole Johns, et al. 2013. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*, 382(9904):1575–1586.

Paul Wicks, Michael Massagli, Jeana Frost, Catherine Brownstein, Sally Okun, Timothy Vaughan, Richard Bradley, and James Heywood. 2010. Sharing health data for better outcomes on PatientsLikeMe. *Journal of Medical Internet Research*, 12(2):e19.

Til Wykes, Josep Maria Haro, Stefano R Belli, Carla Obradors-Tarragó, Celso Arango, José Luis Ayuso-Mateos, István Bitter, Matthias Brunn, Karine Chevreul, Jacques Demotes-Mainard, et al. 2015. Mental health research priorities for Europe. *The Lancet Psychiatry*, 2(11):1036–1042.