Exploring Online Depression Forums via Text Mining: A Comparison of Reddit and a Curated Online Forum

Luis Moßburger
Media Informatics Group
University of Regensburg
Regensburg, Germany
luis1.mossburger
@stud.uni-regensburg.de

Felix Wende
Media Informatics Group
University of Regensburg
Regensburg, Germany
felix.wende
@stud.uni-regensburg.de

Kay Brinkmann
Media Informatics Group
University of Regensburg
Regensburg, Germany
kay.brinkmann
@stud.uni-regensburg.de

Thomas Schmidt
Media Informatics Group
University of Regensburg
Regensburg, Germany
thomas.schmidt@ur.de

Abstract

We present a study employing various techniques of text mining to explore and compare two different online forums focusing on depression: (1) the subreddit r/depression (over 60 million tokens), a large, open social media platform and (2) Beyond Blue (almost 5 million tokens), a professionally curated and moderated depression forum from Australia. We are interested in how the language and the content on these platforms differ from each other. We scrape both forums for a specific period. Next to general methods of computational text analysis, we focus on sentiment analysis, topic modeling and the distribution of word categories to analyze these forums. Our results indicate that Beyond Blue is generally more positive and that the users are more supportive to each other. Topic modeling shows that Beyond Blue's users talk more about adult topics like finance and work while topics shaped by school or college terms are more prevalent on r/depression. Based on our findings we hypothesize that the professional curation and moderation of a depression forum is beneficial for the discussion in it.

1 Introduction

In recent years, online forums and communities have become an important outlet for growing numbers of people who struggle with depression. For example, the subreddit *r/depression* on Reddit had a growth from around 100,000 members in 2015 to over 570,000 in 2020[1]. While a few decades ago, affected would have had to find support or self-help groups in their local area, they are now able to talk about depression and share their stories with thousands of like minded individuals from the comfort of their own home. Besides r/depression, there are traditional forums focusing on mental health that offer people a place to talk about their condition. Some of these are professionally curated as well, providing support hotlines, ties to experts in psychiatry and further articles on important topics.

In the following paper, we want to investigate how language and content of a platform like r/depression differs from a traditional and professionally accompanied forum and how we can explore these questions with the support of text mining methods. Specifically, we compare content and language through various text mining methods, using word frequencies, topic modeling, word categories as well as sentiment analysis. We regard the project at the moment as descriptive and explorative, therefore we do not want to validate concrete hypothesis but rather explore methods and data to formulate potential hypotheses for future work. While there is research applying computational text analysis on depression forums, we are not aware of similar research focusing on the comparison of a curated and a non-curated forum.

[1] https://subredditstats.com/r/depression

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 70–81
Barcelona, Spain (Online), December 12, 2020.

2 Related Work

First, we shortly describe the main text mining methods we apply in this study: Sentiment analysis is the computational method to analyze the sentiment expressed towards entities, mostly in written text (Liu, 2016). The main application area of sentiment analysis is user generated content on the web like social media (Hutto and Gilbert, 2014; Schmidt et al., 2020b) or movie reviews (Kennedy and Inkpen, 2006). Next to sophisticated machine learning approaches, there are also rule-based approaches working with lexical resources and simple rules (Taboada et al., 2011; Schmidt and Burghardt, 2018). Unsurprisingly, sentiment analysis is a popular method to explore depression and social media. Wang et al. (2013) utilize lexicon-based sentiment analysis to calculate the chance for depression on micro-blogs. IBirjali et al. (2017) predict suicidal ideation in Twitter data via machine learning based sentiment analysis. Davcheva et al. (2019) explore three English-language mental health forums via sentiment analysis. Sentiment scores of users develop depending on several conditions (e.g. how active the user is).

To analyze linguistic and semantic word categories, we use the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2015). It consists of multiple linguistic categories like 1st person singular, pronouns or adjectives as well as rather semantic and psychological categories like anxiety, female/male language or spirituality and a corresponding list of words that have been demonstrated to be connoted with this category. LIWC is an established and trusted resource in psychological investigations (Lee et al., 2015) but is also applied in areas outside of psychology (Schmidt et al., 2020a). Findings with LIWC include that people with depression use first person singular and negatively biased words more frequently than a control group (Lee et al., 2007).

Topic modeling is a method to create topical categories based on text documents without a priori subject definitions (Jockers, 2013). We apply Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003), which is one of the most established topic modeling approaches (Jockers, 2013). A topic is a list of words that frequently occur with each other in a set of documents. Given a number of expected topics, a LDA model produces lists of such words as a result. LDA models have also already been utilized to examine the use of language of depressive individuals. For example, Resnik et al. (2015) examined the use of supervised topic models in the analysis of linguistic signals for detecting depression.

There is numerous research examining depression and depression communities in social media cf. (Conway and O'Connor, 2016). De Choudhury and De (2014) apply various techniques and examine multiple mental health communities on Reddit. Among other, they characterize mental health social support into the categories "emotional", "informational", "prescriptive" and "instrumental", of which "emotional" and "prescriptive" are more likely to occur. De Choudhury and Kiciman (2017) continued their work by focusing on how the language of comments influences risk to suicidal ideation. Park and Conway (2017) use the LIWC dictionary to analyze r/depression and other health related forums to investigate user based longitudinal changes. They show that users with a long-term participation shifted to the use of more positive language and indicated that this leads to positive effects. Fraga et al. (2018) explore multiple mental health subreddits via discourse pattern analysis and found that the longest discussions are initiated by threads asking for help and that encouragement words are a frequent pattern.

3 Methods and Results

3.1 Data Acquisition and Corpora

For more information and access to parts of the used corpus, visit `https://github.com/lauchblatt/OnlineDepressionForumsTextMining`

3.1.1 Subreddit r/depression

r/depression[2] is an English-speaking subreddit and consist of *submissions*, also called *threads* in other forums. Such submissions include a title and the initial content written by the author. Answers to a submission are called *comments*. Scraping of /r/depression was done using the Pushshift.io API Wrapper

[2]https://www.reddit.com/r/depression/

Metric	Beyond Blue	r/depression
Threads/Submissions	3,922	131,073
Answers/Comments	24,500	715,128
Posts (Threads+ Answers)	28,422	846,201
Tokens	4,982,391	60,632,208
Tokens per post	175.3	71.7
Tokens per initial post	229.2	186.0
Tokens per answers	166.7	50.7
Tokens per sentence	12.6	11.3
Comment/submission ratio	6.2	5.5

Table 1: Corpus Metrics

psaw [3] and the Python Reddit API Wrapper *praw* [4] .

1,007,134 posts (submissions and comments) from all submissions created in 2019 were gathered between 02.04.2020 and 02.20.2020. Due to the much smaller corpus available for Beyond Blue, we decided to limit the extraction to a single year, to keep the corpora comparable to some extent but still decided to include an entire year to avoid influences due to seasonal changes. 2,295 posts were filtered out due to lack of content, like empty posts or ones that only contain a link. Another 158,638 posts were not considered because they were already deleted by the author himself, moderators or spam filter. This leads to 846,201 posts consisting of 131,073 submissions (15.5%) and 715,128 comments (85.5%).

3.1.2 Beyond Blue

"Beyond Blue" is a non-profit organization funded by Australian governments and states (Jorm et al., 2005) and offers an English-speaking forum specifically for depression[5]. Beyond Blue offers support ranging from their website featuring many up to date articles on mental wellbeing and a welcoming atmosphere, to a 24-hour hotline, online chat, several different well-tailored forums and special information on current topics. We acquired permission by the Beyond Blue research team to scrape the content of their forum.

Beyond Blue was scraped using Python with the standard libraries *lxml* and *urllib*. The basic forum architecture is similar to Reddit, however, submissions are called *threads* and *posts* (equivalent to Reddit's comments) and cannot be nested, but may have a note referring to which post they are an answer. The only metadata available in the forum are the threads title and, respectively for every post, text, date and user information. Opposing to r/depression, we gathered all available threads and metadata in Beyond Blue, which still results in a noticeably smaller corpus, that contains posts from April 3rd 2013 until January 12th 2020. This results in 28,422 posts, of which 3,922 are initial posts (13.8%) of threads and 24,500 are answers inside threads (86.2%).

3.2 General Corpus Analysis

We used *SpaCy*[6] as central tool for the general corpus analysis. Table 1 illustrates general corpus statistics.

The Beyond Blue corpus consists of 4,982,391 tokens divided among 395,544 sentences. Considering the Beyond Blue corpus contains a total of 28,422 posts, this results in 175.3 tokens per post on average. There is a notable difference between the average token counts of initial posts of a thread (229.2) and answers (166.7). The average token count of a sentence is 12.6. Combining these results with the distribution of all posts into submissions and comments, the submissions (13.8% of all posts) are responsible for 18.0% of the tokens. r/depression contains a total of 60,632,208 tokens within 5,369,000 sentences.

[3] https://github.com/dmarx/psaw
[4] https://github.com/praw-dev/praw
[5] https://www.beyondblue.org.au/get-support/online-forums/depression
[6] https://spacy.io/

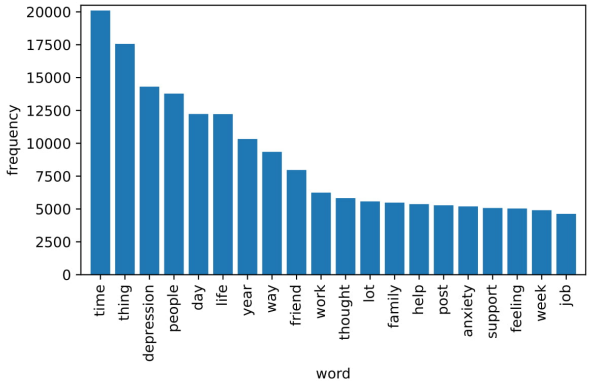

Figure 1: Twenty most frequent words of Beyond Blue

A post on r/depression contains 71.7 tokens on average. Splitting the posts into submissions and comments, the average token count is 186.0 for a submission and 50.7 for a comment. An average sentence contains 11.3 tokens. The submissions of r/depression, which make up 15.5% of all posts, contain 40.2% of all tokens.

Looking at the amount of submissions and comments, the comment/submission-ratios of 6.2 (Beyond Blue) and 5.5 (r/depression) are similar. The length of the posts makes the difference. By average, a post on Beyond Blue contains 144.5% more relevant tokens than a post on r/depression. The reason for this is that there are a lot of very short comments on r/depression as one can see by the very large difference concerning the tokens per answers. In addition, comparing the average token count of submissions, the token count on Beyond Blue is still a notable 20% larger than that of r/depression. These results lead to the assumption that users and moderators engage in much more length into the problems of initial posts.

3.3 Word Frequencies

We analyzed the most frequent words to gain a first impression of the topics and sentiment of the forums. Therefore, all stop words and tokens, which were not tagged as a noun, were removed from the corpora. We have chosen only nouns instead of all words, as these provide a better first overview of the corpora. The tokens were lemmatized and then counted to calculate the word frequency.

The most used word of the Beyond Blue corpus (figure 1) is ”time” with 20,101 mentions. This is due to the users writing regarding to a specific time in their or other people’s life. It is striking that the majority of the most frequent nouns are indeed other time related terms: ”day” (0.245%), ”year” (0.207%) and ”week” (0.098%). ”Depression”, which is the main topic of the forum, is unsurprisingly the third most used word with 14,307 mentions (0.287%). Other frequent words like ”friend” and ”family”, ”work” and ”job”, ”feeling” and ”thought” or ”help” and ”support” possibly indicate some topics, which are explored in detail in the topic modeling section. Looking at the r/depression corpus (Figure 2), ”time” is also the most used word (0.376%). The time-related terms are very common as well with 0.262% (day), 0.228% (year), 0.084% (month) and 0.080% (week), which could indicate that much of the conversation relies on narrating own experiences. In direct comparison, the top twenty most frequent words of both forums are quite similar. Four of the twenty most frequent words of Beyond Blue are not included in the ones of r/depression (”lot”, ”post”, ”anxiety”, ”support”) and vice versa (”person”, ”school”, ”shit”, ”month”). The occurrence of ”school” (0.107%) among the most frequent nouns might indicate a younger user base for r/depression.

3.4 Word Categories

To analyze the language used in both forums we split their content into language categories with the help of the LIWC dictionary. Results are received by assigning words of a given text to the linguistic and semantic word categories of the LIWC dictionary and yield the occurrence of each category in percentages as a result. Table 2 illustrates the results.

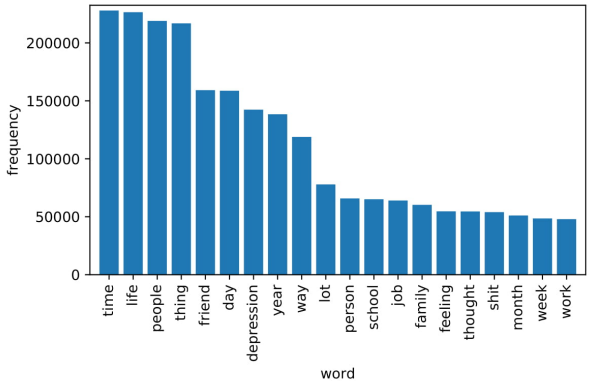

Figure 2: Twenty most frequent words of r/depression

LIWC Category	**Examples for words in this category**	**Reddit Value**	**BB Value**	**Difference**
Personal Pronoun	e.g. he, she, you	15.77	16.21	0.44
1st Person Singular	e.g. I, me, myself	15.01	12.78	2.23
1st Person Plural	e.g. our, we, us	0.99	1.45	0.46
2nd Person	e.g. y'all, you, yourself	4.34	5.37	1.03
3rd Person Singular	e.g. he, she, oneself	1.52	1.67	0.15
3rd Person Plural	e.g. them, they, they'll	0.78	0.66	0.12
Affective Processes	e.g. emotion, hopes, ugh and all terms of positive/negative emotion	14.30	14.72	0.42
Positive Emotion	e.g. better, fabulous, joy	4.43	4.81	0.38
Negative Emotion	e.g. bad, rotten, upset	7.12	6.54	0.58
Past Focus	e.g. ago, previous, remembered and other verbs in past tense	3.90	3.66	0.24
Present Focus	e.g. current, nowadays, understands and other verbs in present tense	17.70	17.95	0.25
Future Focus	e.g. expect, hopeful, wishing	1.60	1.65	0.05
Death	e.g. bury, die, kill	0.38	0.15	0.23
Swear Words	e.g. damn, hell, moron	1.33	0.74	0.59

Table 2: Percentages for LIWC word categories occurrence in r/depression and Beyond Blue

Sentiment Class	Beyond Blue	r/depression
Positive	40.20%	34.49%
Neutral	34.42%	34.20%
Negative	25.38%	31.32%

Table 3: Overall ratio of sentences classified with a polarity class

Sentiment Class	Beyond Blue		r/depression	
	initial	answer	initial	answer
Positive	29.12%	42.54%	28.44%	38.23%
Neutral	33.45%	34.62%	34.33%	34.12%
Negative	37.43%	22.84%	37.23%	27.65%

Table 4: Ratio of sentences classified with a polarity class, split by initial posts and answers

For clearer visualization, we only included a few categories, which either have large differences or are important for our context of depression forums. First, users of r/depression make more frequent use of negative language than those of Beyond Blue as shown by the *positive emotions* and *negative emotions* categories. This validates similar findings for sentiment analysis. Differences can also be seen in the rather negative categories *swear words* and *death*, where r/depression has nearly twice or three times the amount of Beyond Blue. Especially the difference considering swear words might be another indicator for a younger user base using more informal language. r/depression has a stronger *past focus* while Beyond Blue has a slightly stronger *present focus*. While both forums use about the same amount of pronouns, there is a notable difference in which pronouns they use. Within r/depression, the first person singular is more frequently used. Contrary, inside Beyond Blue first person plural and second person are more common. This might point to an important difference in communication since users in r/depression tend to talk more about themselves while the user base in Beyond Blue engages more in discussion directed towards each other. This is to some extent also in line with the finding that answers to initial posts are longer in Beyond Blue than in r/depression. Nevertheless, the results overall underline related findings that language of depressive individuals is negatively charged and includes frequent use of first person singular (Lee et al., 2007). Beyond Blue's smaller percentages in those two categories therefore imply that their user base utilizes less depressive language.

3.5 Sentiment Analysis

Sentiment analysis was conducted by using VADER (Hutto and Gilbert, 2014). VADER is a lexicon-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media and is therefore also used for Reddit (Schmidt et al., 2020b). VADER shows good evaluation results for thie context of social media (Hutto and Gilbert, 2014). Therefore, it is a fitting selection for our sentiment approach. VADER provides sentiment analysis on a sentence-level. VADER classifies the given text with one of three polarities: positive, neutral and negative.

There is nearly the same ratio of neutral sentences in both forums, but positive and negative sentence ratios differ quite notably (see table 3). Beyond Blue consists of 40.20% positive sentences and 25.38% negative ones, whereas r/depression shows 34.49% positive and 31.32% negative sentences. Accordingly, r/depression generally has a more negative sentiment than Beyond Blue. The results of sentiment analysis when splitting initial posts and answers into sentences, are shown in Table 4.

The sentiment of initial posts is nearly identical for both forums. The answers are more positive compared to the initial posts overall, but answers on Beyond Blue are more positive than the ones on r/depression. According to the sentiment, the content that users post in one of these forums is approximately the same. The difference are the replies to those posts, which are more positive on Beyond Blue. This implies that the community in Beyond Blue reacts more supportive to people talking about their depression.

3.6 Topic Modeling

For our analysis, we created a Latent Dirichlet Allocation (LDA) model with the Python library *gensim* (Rehurek and Sojka, 2010) for each forum. We decided on creating one text document per post as it led to better results than using an entire thread. Tokenization and lemmatization were done for each post and stop words were removed. Posts which then had less than five tokens left were excluded from the topic modeling process, as they most likely will not contain any useful information. Originally, we created models with thirty topics, each represented by thirty keywords for both forums. Manual decrease for the number of topics to fifteen and the number of words to twenty lead to clearer visualization. The topics removed were mostly duplicates or topics with unclear results, the filtered keywords were duplicates and those which did not provide additional information. After testing with different variables, we decided on a chunk size of 2500 posts and 15 passes for each model, which generated meaningful topics.

Please refer to the appendix for the reduced and filtered 15 topics per forum and the corresponding 20 key terms for each topic (Appendix: Figure 3-4). Please note that the names for the topics are added by us based on our interpretation of the term list as LDA topic modeling does not produce names for topics. We will use these names in the following sub chapter.

While r/depression has entire topics revolving around school and college life (topics 2 [”School”] and 10 [”College”]) pointing towards user groups consisting of teenagers and young adults, Beyond Blue has several topics dealing with subjects like finances or insurance (e.g. topics 12 [”Social Care”] and 14 [”Work”]). This becomes even clearer when looking at the topic ”Family” for each forum (topic 4 for r/depression and topic 6 for Beyond Blue). While the r/depression topic seems to be mainly from the viewpoint of someone growing up, with keywords like *mom* or *dad*, the Beyond Blue topic seems to be from the viewpoint of a parent with keywords like *boy* or *girl*.

Other differences can be found in the user's interests and hobbies of each forum. r/depression contains topics revolving around ”Entertainment Media” (topic 12) or ”Social Life” with friends (topic 9), whereas Beyond Blue has topics about general ”Lifestyle” (topic 13, keywords e.g. *exercise, book, walk*) and ”Local Life” (topic 5, keywords e.g. *dog, shop, town, car*). Similar topics, that occur for both forums include ”Relationships” (r/depression topic 8; Beyond Blue topic 4), medical aspects (r/depression topics 14 [”Therapy”] and 15 [”Treatments”]; Beyond Blue topic 2 [”Mental Condition”], 10 [”Treatments”] and 11 [”Therapy”]), ”Work” (r/depression topic 13, Beyond Blue topic 14) or ”Family”. Topics about relationships and medication are important in both forums and seem to be a consistent discussion point.

While topics concerning self-harm and suicide can also be found in both forums (topic 1 [”Emotional Pain”] and 3 [”Self-Harm”] for r/depression and topic 1 [”Self-Harm”] for Beyond Blue), they are represented very differently. In r/depression with keywords like *hate, cry, tired, emptiness, suffering, meaningless* and *torture*, the topics seem to be about venting or describing one's feelings and emotions that are connected to thoughts about self-harm. Even though this can be found in some words of the Beyond Blue topic as well (e.g. *anger, invisible*), it seems the underlying motive is about supporting other members of the community who have suicidal thoughts, shown by words like *encourage, concern, support* or *relate*. Beyond Blue contains more words like *reach, seek, offer* and *help* (topic 8 [”BB Community”]) or *suggestion, advice* and *sharing* (topic 3 [”BB Forum”]) paired with terms describing the forum itself like *BB, community* or *reply* and *post* which indicates that people offer help or reach out for help from other members of the forum. While the r/depression topics also include some words like *encourage* (topic 14 [”Therapy”]) or *recommend* (topic 15 [”Treatments”]), their occurrences are rare. All in all, similar to the ones about self-harm, most of the r/depression topics rather seem to be about describing one's own feelings and experiences (e.g. topics 9 [”Social Life”], 10 [”College”] and 11 [”Future Expectations”]).

4 Limitations

We want to highlight some of the major limitations of our project. The striking problem is that both corpora differ largely from each other, predominantly considering the size. We tried to control this limitation to some extent by (1) focusing on normalized metrics for our methods and (2) including at least a year for r/depression to avoid seasonal changes and to reduce the size of r/depression. The demographic of

the user base might differ as well. Beyond Blue is a predominantly Australian forum while r/depression is constituted by a more international user base. While statistics show that Reddit's user base is primarly male and young[7], certain results point to a more adult user base for Beyond Blue (althoug we have no concrete statistics). r/depression has also no moderation by trained professionals. We address further differences between the user base in the upcoming chapter. Another important limitation lies in the application of some of the methods. The results of the sentiment analysis have to be taken with caution. While VADER is optimized for social media content, we did not precisely evaluate the performance on these specific corpora. While the LIWC dictionary is an established resource for lexical analysis, it is also not specifically designed for social media content and might therefore lack some important terms or categories.

5 Discussion

Please not that due to the fact that we gathered publicly available data and we only deal with anonymized data, we did not need permission by the ethics commission of our institution at the time of the writing.

First, we want to highlight similarities between both forums, which might point to general attributes of depression forums and their users: In both forums time-related terms are among the most frequent ones, showing that the narration of past and present events is a consistent discussion in depression forums. Both forums show a tendency to positive sentiment, thus illustrating that support and the recollection of positive aspects are an important part of depression forums in general. Topics about relationships, family, work and medication are highly important in both forums and structured rather similar. Especially the consistency considering the topics about relationships and family show the importance of these aspects for depressive illnesses.

Considering differences of the two forums, our findings correlate well and support assumptions over all applied methods. General corpus analysis and analysis about word categories conducted with LIWC show that the posts in general, as well as the language in Beyond Blue revolve around support and positivity, whereas negative speech and posts with little information value are more common in r/depression. This is in line with the finding that the answers to posts are much longer on average in Beyond Blue than in r/depression pointing to a more seriously engaged user and moderator base. The sentiment scores support this impression with Beyond Blue being noticeably more positive, mainly because answers to initial submissions in r/depression have a more negative sentiment than the rather positive responses on Beyond Blue. The findings produced by the topic modeling approach include the impression that the user base of Beyond Blue is older, as some topics have a focus on financial and organizational themes, and indicate a "parent view" while r/depression's users speak about school, college and their parents. Most interestingly, and complementing other findings, topics dealing with self-harm and suicide can be found in both forums but are considerably more support oriented in Beyond Blue and more emotional and negatively biased in r/depression.

Concluding, Beyond Blue and r/depression seem to have different focal points. Where Beyond Blue's support oriented language and content seems to indicate conversations where concrete problems are solved, r/depression's emotional posts, past tense and rather negative sentiment imply the focus lies on sharing experiences and converse about emotions. While our findings are exploratory at the moment, we hypothesize that a professionally curated and smaller forum might be more beneficial to be used by persons affected by depression since it is more supportive and positive in its content. This conclusion has to be taken with caution because (1) many differences might rely on the different user demographic, which seems to be younger in r/depression and (2) the open and more peer-oriented culture of Reddit might still be helpful and beneficial in some situations.

There are certain approaches we want to pursue to further investigate our results. We want to create a more balanced corpus by acquiring a collection of similar forums like Beyond Blue instead of just one. It is also necessary to optimize several of our text mining techniques for our context. Lastly, we also plan to work closely with professional psychologists and therapists to integrate their expertise into our project.

[7] https://www.techjunkie.com/demographics-reddit/

References

Marouane Birjali, Abderrahim Beni-Hssane, and Mohammed Erritali. 2017. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science*, 113:65–72.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Mike Conway and Daniel O'Connor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology*, 9:77–82.

Elena Davcheva, Martin Adam, and Alexander Benlian. 2019. User dynamics in mental health forums–a sentiment analysis perspective.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Eleventh International AAAI Conference on Web and Social Media*.

Barbara Silveira Fraga, Ana Paula Couto da Silva, and Fabricio Murai. 2018. Online social networks in health care: A study of mental disorders on reddit. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 568–573. IEEE.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Matthew L Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Anthony F Jorm, Anthony F Jorm, Helen Christensen, Kathleen M Griffiths, Anthony F Jorm, Helen Christensen, and Kathleen M Griffiths. 2005. The impact of beyondblue: the national depression initiative on the australian public's recognition of depression and beliefs about treatments. *Australian & New Zealand Journal of Psychiatry*, 39(4):248–254.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125.

Chang H Lee, Myungju Lee, Sungwoo Ahn, and Kyungil Kim. 2007. Preliminary analysis of language styles in a sample of schizophrenics. *Psychological reports*, 101(2):392–394.

Changhwan Lee, Kyungil Kim, Jeongsub Lim, and Yoonhyoung Lee. 2015. Psychological research using linguistic inquiry and word count (liwc) and korean linguistic inquiry and word count (kliwc) language analysis methodologies. *Journal of Cognitive Science*, 16(2):132–49.

Bing Liu. 2016. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Albert Park and Mike Conway. 2017. Longitudinal changes in psychological states in online health community members: understanding the long-term effects of participating in an online depression community. *Journal of medical Internet research*, 19(3):e71.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107.

Thomas Schmidt and Manuel Burghardt. 2018. An evaluation of lexicon-based sentiment analysis techniques for the plays of gotthold ephraim lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149.

Thomas Schmidt, Isabella Engl, Juliane Herzog, and Lisa Judisch. 2020a. Towards an analysis of gender in video game culture: Exploring gender-specific vocabulary in video game magazines. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020*, pages 333–341.

Thomas Schmidt, Florian Kaindl, and Christian Wolff. 2020b. Distant reading of religious online communities: A case study for three religious forums on reddit. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020*, pages 157–172.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 201–213. Springer.

A Appendix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Emotional Pain	School	Self-Harm	Family	Money	Physical Condition	Mental Condition	Relationships	Social Life	College	Future Expectations	Entertainment Media	Work	Therapy	Treatments
1	love	school	live	parent	money	brain	mental	relationship	friend	college	future	enjoy	work	help	doctor
2	shit	teacher	die	mom	pay	effect	issue	guy	family	fail	happiness	watch	job	need	medication
3	hate	bully	suicide	dad	buy	healthy	health	girl	close	class	dream	game	quit	therapist	med
4	fuck	counselor	death	mother	save	weight	illness	girlfriend	social	study	goal	fun	career	therapy	psychiatrist
5	fucking	black	alive	brother	afford	exercise	physical	date	meet	grade	reality	music	company	listen	hospital
6	hurt	winter	attempt	sister	poor	smoke	panic	boyfriend	lonely	graduate	society	video	apply	advice	antidepressant
7	kill	ghost	option	father	bill	weed	chemical	woman	group	test	purpose	interest	position	support	treatment
8	cry	white	decision	cat	cost	gym	patient	ex	conversation	student	joy	hobby	fire	reach	pill
9	pain	lunch	choice	son	debt	gain	chronic	wife	online	university	create	movie	area	professional	recommend
10	tired	senior	survive	animal	rent	habit	ADHD	abuse	message	semester	opportunity	medium	project	offer	episode
11	cut	bus	existence	argument	law	mechanism	Wellbutrin	young	friendship	exam	achieve	art	program	suggest	appointment
12	lie	military	suffering	sadly	education	addiction	imbalance	partner	party	final	path	tv	business	helpful	medical
13	stupid	summer	meaningless	replace	motivate	lift	address	sex	loneliness	disappoint	success	favorite	financial	willing	switch
14	hell	classmate	cruel	threaten	expensive	unhealthy	OCD	adult	connection	highschool	desire	boring	boss	psychologist	prescribe
15	heart	popular	torture	aunt	interview	meditation	psychiatric	baby	isolate	homework	ahead	YouTube	store	community	risk
16	emptiness	beer	workout	ease	earn	body	psychological	husband	join	major	successful	bored	office	provide	insurance
17	inside	transfer	fighting	couch	dollar	motivated	eating	divorce	invite	academic	expectation	computer	shift	session	psych
18	sick	September	inevitable	grandma	insurance	active	disagree	marry	distance	course	direction	distract	field	resource	available
19	exist	ditch	empathize	insult	car	balance	apathy	raise	interaction	score	succeed	series	coworker	encourage	dangerous
20	dead	withdraw	sorrow	adopt	unmotivated	lifestyle	conflict	cheat	circle	freshman	pursue	passionate	team	counsellor	withdrawal

Figure 3: Cleaned LDA topics for r/depression

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Self-Harm	Mental Condition	BB Forum	Relationships	Local Life	Family	Alcoholism	BB Community	Physical Health/ Food	Treatments	Therapy	Social Care	Lifestyle	Work	Emotional/Weather
1	suicide	mental	thank	love	dog	child	alcohol	help	eat	medication	psych	team	find	job	cry
2	harm	health	reply	life	group	mum	drinking	support	weight	mood	psychologist	claim	work	work	tear
3	self	illness	post	relationship	cat	parent	drug	forum	food	bipolar	appointment	ps	exercise	money	emotion
4	phone	depression	thread	believe	town	husband	alcoholic	welcome	ocd	anti	gp	private	mind	pay	song
5	comment	anxiety	read	abuse	live	daughter	drunk	professional	healthy	depressant	psychiatrist	Psychologist	book	study	cold
6	encourage	symptom	response	woman	car	son	sober	find	binge	effect	session	charge	good	uni	hug
7	community	issue	appreciate	hurt	city	mother	bottle	seek	gain	diagnosis	doctor	disability	enjoy	boss	weather
8	concern	suffer	share	self	house	old	loathing	post	eating	doctor	see	centrelink	walk	financial	light
9	service	condition	advice	soul	pet	young	mate	gp	disorder	psychiatrist	visit	clinic	positive	apply	crying
10	moderator	treatment	story	beautiful	country	family	booze	counsellor	body	episode	refer	fee	help	career	rain
11	general	experience	kind	forgive	local	sister	stop	advice	kg	disorder	plan	hr	read	stress	sun
12	supportive	professional	respond	human	animal	dad	frighten	BB	habit	symptom	therapy	cover	thought	employment	flow
13	invisible	cause	comment	happiness	club	wife	Alcohol	reach	intrusive	prescribe	treatment	safety	day	quit	burst
14	issue	chronic	helpful	pain	meet	home	meeting	offer	loss	depressive	discuss	department	goal	interview	shine
15	relate	medical	forum	special	move	brother	sincerely	helpful	genetic	ad	hospital	insurance	activity	position	warm
16	support	test	lovely	way	home	baby	tragedy	hope	label	take	patient	worker	learn	bill	emotional
17	clinically	Depression	encouragement	live	shop	father	licence	good	perfectionist	therapy	therapist	system	time	debt	dark
18	anger	admission	suggestion	laugh	walk	love	merry	chat	meal	dosage	Medicare	emergency	focus	workplace	inside
19	suicidal	PTSD	supportive	strength	volunteer	girl	poison	community	appetite	change	medical	payment	achieve	company	ocean
20	attempt	recover	sharing	deserve	area	boy	drink	suggest	product	BPD	recommend	policy	play	income	bright

Figure 4: Cleaned LDA topics for Beyond Blue