Proceedings of the RaPID-4 @LREC 2022, pages 80–85
Marseille, 25 June 2022

CorEDs: a Corpus on Eating Disorders

Melissa Donati, Carlo Strapparava
University of Trento, FBK-IRST
melissa.donati@studenti.unitn.it, strappa@fbk.eu

Abstract

Eating disorders (EDs) constitute a widespread group of mental illnesses affecting the everyday life of many individuals in all age groups. One of the main difficulties in the diagnosis and treatment of these disorders is the interpersonal variability of symptoms and the variety of underlying psychological states that are not considered in traditional approaches. In order to gain a better understanding of these disorders, many studies have collected data from social media and analysed them from a computational perspective, but the resulting dataset were very limited and task-specific. Aiming to address this shortage by providing a dataset that could be easily adapted to different tasks, we built a corpus collecting ED-related and ED-unrelated comments from *Reddit* focusing on a limited number of topics (fitness, nutrition, etc.). To validate the effectiveness of the dataset, we evaluated the performance of two classifiers in distinguishing between ED-related and unrelated comments. The high-level accuracy of both classifiers indicates that ED-related texts are separable from texts on similar topics that do not address EDs. For explorative purposes, we also carried out a linguistic analysis of word class dominance in ED-related texts, whose results are consistent with the findings of psychological research on EDs.

Keywords: Corpus Linguistics, Text Classification, Eating Disorders

1. Introduction and motivation

The term Eating Disorders (EDs) groups a number of mental illnesses characterized by abnormal or disturbed eating habits that have an adverse effect on both mental and physical health. Despite the commonality of these health issues that constitute one of the prevalent types of psychological disorders nowadays, EDs are still underdiagnosed and interventions are oftentimes ineffective because traditional "one-size-fits-all" approaches in treatment do not allow to target the specific psychological variables for each individual (Zhou et al., 2020). The self-protective nature of EDs represents an additional obstacle for researchers that are willing to investigate deeper the factors that promote EDs, because people suffering from these disorders are not likely to communicate their experiences and emotions with physicians and doctors (Zhou et al., 2020). However, the increasing engagement of social media users in health-related conversations and discussions (Lenhart et al., 2010) could constitute a potential solution to such problems. Indeed, given the community-building nature of social media, individuals with an ED tend to engage more and more openly in discourse about their disorders with users sharing similar experiences (Kenny et al., 2020), thus making available large amount of ED-related linguistic data. As a consequence, applying data mining techniques to extract and analyse data from social media has become a popular methodological approach in health care research. So far, however, the investigations that where conducted involving EDs led to the collection of small datasets created *ad hoc* for single studies. Besides not being representative and therefore not allowing to generalise the observed trends, the small size of such datasets constitutes also an issue for the implementation of machine learning approaches. Given the need for a larger collection of ED-related data, in this paper we present an English corpus of ED-related posts extracted from *Reddit*. The aim of this work is to create a dataset that can be used for different purposes, from linguistic and content analysis of ED-related discourse in order to gain crucial insights into the factors that can motivate and trigger EDs behaviours, to the development of classifiers that could detect ED-relevant contents on social media. The paper is organized as follows: Section 2 describes the related works; Section 3 is devoted to the dataset creation process; Section 4 presents some statistics on the dataset; Section 5 shows our evaluation of the dataset based on a machine learning approach and a short linguistic analysis; finally Section 6 presents our conclusions and future directions of the work.

2. Related work

The extraction and analysis of health-related data from social networks is now a well-established methodology in different areas of healthcare research (Mullany et al., 2015) The main advantage of electronic communication is that it allows to discuss medical concerns in a less direct way, making users feel less vulnerable and thus allowing them to express their opinions and emotions more openly (Suler, 2004). This is particularly true for people (especially teenagers and adolescents) suffering from EDs. Indeed, researchers have observed that, due to a desire for anonymity, a significant portion of information seeking and discussion with respect to EDs takes place on the Internet and through social media (Oh et al., 2013). For this reason, recently many studies in the field of psychology and medicine have investigated EDs adopting a corpus-based approach to analyse linguistic data extracted from social media (Lukač and others, 2011;

Malson et al., 2011; Leonidas and Dos Santos, 2014; Hunt and Harvey, 2015; Mullany et al., 2015). In particular, the analysis of ED-related forums using linguistic inquiry tools such as term frequency analysis, part of speech (POS) analysis and sentiment analysis allowed to uncover some specific linguistic properties that characterize ED-related discourse and that could be useful for clinicians to understand the underlying needs and emotions of individuals suffering from these disorders (Oh et al., 2013). All previous works, however, share a relevant limitation: the linguistic analyses were carried out on small datasets created ad hoc for single studies and they were often focused on a limited number of keywords. This is the case, for example, for Bohrer's work (Bohrer et al., 2020), that analysed online ED-related forums targeting the process of recovery; but also for McCaig and colleagues' works (McCaig et al., 2018; McCaig et al., 2019; McCaig et al., 2020), whose thematic analysis of ED-related forums was centred on calorie counting apps and fitness tracking technology; as well as for Moessner's study (Moessner et al., 2018), that focused on identifying topics related to social support in EDs treatment. Besides reducing the generalizability of the observed trends to the population, the small dataset size does not allow to implement machine learning algorithms for ED-relevant contents identification and recognition tasks. So far, there has been only a single attempt to develop a machine learning-based classifier to identify tweets related to EDs (Zhou et al., 2020). In this case the authors did manage to collect a quite large dataset, but the nature of the texts they collected, that are short, often convoluted and difficult to analyse because of the presence of hashtags and abbreviations, does not allow to adapt the dataset to different tasks, thus limiting its applications.

3. Methods

Our goal was to create a corpus on EDs that could be used for various types of analyses, being at the same time easily adaptable for different machine learning tasks. In order to accomplish this, we focused on the American social news website and forum *Reddit*, a discussion website where registered members submit contents such as links, text posts, images, and videos, which are then voted up or down by other members. Registrations is free and posts are organized by subject into user-created boards called "communities" or "subreddits", which cover a wide variety of topics and can be accessed via keyword search. The reason why we selected this platform is twofold: on the one hand, not imposing any limitation in the length of the posts, *Reddit* makes available longer and more complex texts (compared for example with the largely investigated *Twitter*, where the maximum length of a post is 280 characters); on the other hand, the discussion-oriented nature of the website also naturally leads to more articulated and linguistically rich comments. For these reasons, the type of linguistic data that can be extracted from *Reddit* appears to be suitable for the task at hand, that is building a multi-purpose corpus on EDs. Indeed, it has been shown that performance of NLP methods increase with the length of the documents being analyzed (Curiskis et al., 2020). In addition, as suggested by Shen & Rudzicz (2017), the lenght of *Reddit*'s posts, together with the website organization, constitute a "considerable potential for sophisticated methods of feature extraction as well as qualitative analysis" (Shen and Rudzicz, 2017, pag.63).

3.1. Keyword-based search on *Reddit*

In order to get access to the relevant comments we performed a keyword-based search using the Python Reddit API Wrapper[1] (PRAW), a Python package that allows for simple access to Reddit's API[2]. As keywords we used the list of names of the most prevalent EDs that was obtained from a dedicated website[3]. For each ED name we obtained the related subreddits titles and collected all the posts classified under each subreddit (see Table 1 for the complete list). Each text was then annotated according to the ED it describes using an abbreviation of the corresponding ED name (ex. BU for Bulimia). The abbreviations are reported below the ED names in the first column of Table 1. Given that we aimed to build a corpus that could be used for EDs classification purposes, we needed to collect an equally extensive sample of EDs-unrelated comments that would constitute the negative class. Following the same procedure that was described above, we performed a keyword-based search on *Reddit*, and we extracted all the posts classified under each subreddit related to the keywords. One of the most common ways of building negative class dataset is via random selection, however, in this case the comments were extracted starting from a list of manually selected keywords that refer to frequently occurring topics in EDs discourse (i.e. *food, fitness*). In doing so, we could hypothesize that the main feature(s) distinguishing the positive class (ED-related posts) from the negative class (ED-unrelated posts) are exactly the features that characterize EDs discourse. Table 2 reports the list of selected topics, the corresponding subreddit titles and the number of posts extracted for each subreddit.

3.2. Comments selection and cleaning

In the cleaning step we cleared each comment from emoticons, hyperlinks and hashtags, but we did not remove punctuation marks because they have often been shown to provide a crucial contribution for understanding the psychological state of the speaker (Say and Akman, 1996; Oh et al., 2013). In order to standardise the texts and to maximize the quantity and quality of linguistic information, we replaced contractions (i.e.

[1] https://github.com/praw-dev/praw

[2] https://www.reddit.com/dev/api

[3] https://www.freedeatingdisorders.org/patient-family-support/types-of-eating-disorders/

ED names	subreddit titles	posts
Eating Disorder(s) ED	*EatingDisorders', 'eating_disorders', 'EatingDisorderHope', 'edsupport', 'EDAnonymous', 'EdAnonymousAdults', 'EDRecovery_public', 'EDRecovery'*	5089
Anorexia AN	*'AnorexiaNervosa', 'AnorexiaRecovery', 'ProAnaBuddies', 'anorexiaflareuphelp'*	3957
Bulimia BU	*'bulimia',' BulimiaAndAnaSupport'*	685
Binge Eating BE	*'BingeEatingDisorder', 'bingeeating'*	811
Purging PU	*'PurgingDisorder'*	6
Not Otherwise Specified NOS	*'NotOtherwiseSpecified', 'Ednos*	23

Table 1: List of ED-related word included in the search, corresponding subreddit titles and total number of posts retrieved for each subreddit

Common EDs topics	subreddit titles	posts
Nutrition	*'nutrition', 'EatCheapAndHealthy', 'ketogains', 'SportNutrition', 'EatingHealthy', 'EatHealthy', 'intuitiveeating'*	2332
Food	*'HealthFoodChat', 'FitnessFood', 'Macrofoodients'*	17
Fitness	*'xxfitness', 'veganfitness', 'runmeals', 'workout', 'bodyweightfitness'*	1833
Diet	*'Dietandhealth', 'diet', ' dieting', 'PlantBasedDiet', 'Pescetarian'*	1351

Table 2: List of frequently occurring topics in EDs discourse included in the search, corresponding subreddit titles and total number of posts retrieved for each subreddit

don't), abbreviations and slang forms (i.e. *asap*), and medical acronyms (i.e. *AN*) with the corresponding extended forms (respectively: *do not, as soon as possible* and *Anorexia Nervosa*). Finally, given the already discussed potential of longer texts for both quantitative and qualitative analysis (see Section 3), we decided to exclude from the cleaned dataset comments that were shorter than the maximum length of a tweet (280 char.).

4. Corpus statistics

In this section we highlight some additional statistics regarding CorEDs. These statistics refer to the total number of posts that were collected, the total number of words and the average post length for each of the two datasets. As shown in Table 3, the ED-related dataset contains 7662 posts (more than 1.4 million of words) whose average length is 194 words, while in the ED-unrelated dataset there are 6538 posts (around 1.2 million of words) whose average length is 184 words. The whole corpus contains 14200 posts for a total of almost 2.7 million of words and is available for research purposes on request from the corresponding author [CS].

5. Experiments and Results

In this section we describe, the different experiments we carried out to test the validity of the datasets. In particular, we trained two machine learning classifiers and compared their performances. We also performed a short linguistic analysis of the datasets by identifying the dominant word classes.

Dataset	ED-rel	ED-unrel	Total
# of posts	7662	6538	14200
# of words	1 486 325	1 210 495	2 696 820
Av.len.	194	184	189

Table 3: Statistics of the two datasets and of the whole corpus: total number of posts, total number of words and average length of posts (in number of words).

5.1. Classification

The two classifiers selected for the task were the Multinomial Naive Bayes (MNB) and the Support Vector Machine (SVM). Both are well known machine learning algorithms that have been shown to be accurate and highly effective in binary classification tasks. The dataset was split 80% for training and 20% for testing.

5.2. Results

In this subsection we report on and discuss the performance of the two classifiers on our corpus. In order to evaluate and compare their results we used the usual metrics in text classification: Precision (*P*), Recall (*R*), F-score (F_1) and Accuracy (*Acc)*. The results achieved with the two classifiers are reported in Table 4. The high classification performance of both classifiers indicates that good separation between ED-related and unrelated posts can be obtained by using automatic classifiers. As can be seen, overall the SVM performed better than the MNB and this might be due to the fact

that the nature of the SVM is probabilistic and it takes into account the interaction between features, while the MNB is geometric and based on the assumption that the features are independent. For explorative purposes, we decided to extract from the MNB the most informative features that the classifier selected to distinguish between ED-related and ED-unrelated texts. We reported in Table 5 the 20 most informative features for the positive (ED-related) and negative (ED-unrelated) datasets. Interestingly, many features are shared between the two datasets, indicating a high level of similarity between the texts they contain. This is probably due to the fact that the ED-unrelated dataset was specifically built using texts covering topics that overlap with those in the ED-related dataset, such as dieting, fitness and healthy eating, with the only difference that in the ED-related texts these topics are discussed from the perspective of individual suffering from EDs. Indeed, features like 'food', 'eat/eating', 'weight' are shared by the two datasets, while distinguishing features (highlighted in bold in Table 5) are strongly ED-related in the positive dataset ('recovery', 'binge', 'help') and connected to fitness and fitness dieting in the negative one ('protein', 'workout', 'good').

Classifier	P	R	F_1	Acc
MNB	0.91	0.91	0.90	0.904
SVM	0.93	0.93	0.93	**0.935**

Table 4: Results obtained with the Multinomial Naive Bayes (MNB) and the Support Vector Machine (SVM), reported in terms of Precision (P), Recall (R), F-score (F_1) and Accuracy (Acc).

Positive		Negative	
coefficient	feature	coefficient	feature
-5.9344	eating	-5.9675	**diet**
-5.4386	just	-5.9924	like
-5.4889	like	-6.0296	eat
-5.5006	feel	-6.0308	weight
-5.6289	eat	-6.0453	just
-5.6370	weight	-6.1417	eating
-5.7025	know	-6.2835	food
-5.7844	want	-6.2852	body
-5.8613	really	-6.3035	day
-5.8731	food	-6.3171	**protein**
-6.0693	time	-6.3339	want
-6.1010	**recovery**	-6.3871	feel
-6.1958	body	-6.3876	really
-6.2067	**going**	-6.4113	**fat**
-6.2152	**did**	-6.4349	**workout**
-6.2329	day	-6.4497	**have**
-6.2649	**binge**	-6.4557	know
-6.2677	**help**	-6.5064	week
-6.2823	**think**	-6.5317	time
-6.3845	does	-6.5494	**good**

Table 5: Most informative features and corresponding coefficients for the positive (ED-related) and the negative (ED-unrelated) datasets.

5.3. Identifying dominant word classes in ED-related text

In order to gain a better understanding of the characteristics of ED-related text, we performed an analysis to identify the dominant word classes in the ED-related dataset. The adopted methodology was inspired by the work of (Mihalcea and Strapparava, 2009), to calculate the saliency (*dominance)* of a word class in a target collection of texts (for a precise description of the methodology see Mihalcea & Strapparava, 2009). The dominance score obtained with such methodology (Mihalcea and Strapparava, 2009) should be interpreted as follows: if the dominance score takes on a value that is close to 1 this means that the target word class is similarly distributed in both datasets; if the value is significantly higher than 1, then the target word class is dominant in the ED-related dataset; and vice-versa, if the value is significantly lower than 1, this indicates that the word class is dominant in the ED-unrelated dataset. The word classes were extracted from the 2007 version of the Linguistic Inquiry and Word Count (LIWC) lexicon, a resource developed for psycholinguistic analysis that has been largely validated (Pennebaker et al., 2001). LIWC 2007 includes 4482 words and word stems grouped into 64 word classes that are considered relevant for analysing psychological processes. Table 6 and 7 report the top ranked classes for both datasets along with their dominance score and a few sample words belonging to the class and also appearing in the texts. In the direction of a clearer discussion of the results, we divided the word classes into two groups using LIWC categories as reference. More specifically, Table 6 shows the "standard function words categories" (Chung and Pennebaker, 2007, pg.344), i.e. *function words, verbs, pronouns, relatives, prepositions* etc., that are useful to analyse the morphosyntactic structure of the texts, in other words to analyze *how* the content is expressed. On the other hand, Table 7 displays content and emotion words (Chung and Pennebaker, 2007), that are needed to analyse the semantics of the texts, that is to say *what* the text is about. As we can see, focusing on the morphosyntactic structure of discourse (Table 6), ED-related texts appear to be characterized by the large presence of negations, first-person narrative, extensive use of pronouns and preponderance of past tense. This is coherent with the literature, as it has been shown that the use of the first person singular pronoun (Oh et al., 2013), as well as of negations (Leis et al., 2019), is often linked to depression, isolation and mental distress, all conditions strongly related to EDs. In addition, the large use of past tense and/or reference to the past is also in line with the psychological and emotional condition of both people suffering from EDs, who often use the web to talk about events in the past that might have triggered the onset of the

Class	Score	Sample words
ED-related texts		
negate	1.48	*no, never*
i	1.35	*i, mine, my*
ppron	1.34	*his, our, oneself*
pronoun	1.27	*this, which*
past	1.26	*was, were*
ED-unrelated texts		
you	0.79	*you, yours*
assent	0.84	*absolutely,awesome*
quant	0.85	*any, less*

Table 6: Dominant morphosyntactic word classes in ED-related and ED-unrelated texts along with sample words.

Class	Score	Sample words
ED-related texts		
family	2.10	*mum, dad, family*
swear	1.98	*shit, suck*
friend	1.75	*friend, roomate*
anx	1.75	*scared, guilty*
anger	1.63	*angry, fucked*
negemo	1.46	*alone, hate*
excl	1.26	*without, rather*
ED-unrelated texts		
leisure	0.52	*relax, party*
work	0.72	*duty, hardwork*
body	0.75	*abdomen, hips*
money	0.81	*cheap, discount*
posemo	0.86	strong, amazing

Table 7: Dominant semantic word classes in ED-related and ED-unrelated texts along with sample words.

disorder, and people recovering or already recovered who tell their story and share the steps of their healing process (Wolf et al., 2007). Interestingly, at the content level (Table 7) we can see that words indicating personal relationships (*family* and *friends*) appear to be dominant in ED-related texts. This observation does not conflict with the self-protective nature of EDs and it can be explained in different ways. In some cases the comments that we collected were written by people expressing their concern for a loved one struggling with an ED, but in most cases the comments are produced by the people suffering from EDs themselves, who talk about how the disorder affected their social relationships or, unfortunately just as often, describe what role their relatives played in the onset of the ED. Other dominant word classes that emerged are those connected to negative emotions (*anx*, *anger* and *negemo*) and exclusion (*excl*), that describe emotional and psychological conditions typically shared by people suffering from EDs. It is also worth noticing the high dominance score obtained by the class grouping swear words, that are often associated either broadly with the ED (ex. *"fucking anorexia"*) or more specifically with its symptoms (ex. *"purging is shit"*), indicating the "friend and foe" relationship that pulls individuals towards and against their ED (Serpell et al., 1999). By contrast, word classes relating to entertainment (*leisure*), wealth and income (*money*) and positive emotions (*posemo*) are less likely to be found in ED-related texts and resulted dominant in ED-unrelated texts. The high presence of words related to body parts (*body*) and work (*work*) in the dataset that we collected as negative for the classification, is due to the fact that the texts were extracted from subreddits on fitness, diet and healthy eating, where one of the main topics of discussion is the *work*out routine of users and how exercising benefits the physical appearance.

6. Conclusions and future directions

In this paper, we have described an English corpus on EDs, containing comments extracted from *Reddit* covering a limited number of topics (*diet, healthy eating, fitness, nutrition* etc.). The texts are labeled as ED-related or ED-unrelated depending on whether they were extracted from a subreddit on EDs or not. Within the ED-related dataset, the texts are further annotated with an abbreviation referring to the type of ED the text is about. The corpus contains a total of 14200 comments (2 696 820 words), whose average length is 189 words. Since our aim was to build a corpus that could be easily adapted to different tasks and used to perform various types of analysis, while normalizing the text we did not remove punctuation nor stop words and we did not add Part-of-Speech (POS) tags. Moreover, in order to validate the effectiveness of the dataset, we also proposed a machine learning approach for automatically detecting ED-related comments in texts. Our preliminary results are promising, as they show that ED-related texts are separable from texts covering very similar topics but not addressing EDs. However, given the complexity of automatic EDs detection and classification, further experiments need to be carried out to test the soundness of our dataset on different tasks. Finally, the linguistic analysis performed to explore the word classes that characterize ED-related discourse revealed some interesting patterns of word usage –such as the prevalence of first-person narratives, the predominance of negations and a vocabulary that emphasizes the sense of exclusion and negative emotions– that are in line with the related findings in psychology and psychotherapy. As future work we plan to perform more experiments on the datasets, applying other techniques and testing different classifiers with the purpose of understanding how the corpus could be improved and refined. We would also like to extract the most informative features from the SVM classifier in order to compare them with those extracted from the MNB. At the level of linguistic analysis, we would like to implement more fine-grained investigations on the ED-related texts, in particular trying to better handle the fact that words in LIWC can exist within more than one category and can have more than one meaning, which

could have skewed the current results to some degree. In conclusion, we consider this corpus as a first attempt to build a flexible tool that could be used to investigate more extensively possible automatic approaches to EDs detection and discourse analysis and we look forward to further work that could address, from a computational perspective, the main issues in diagnosis and treatment of these pervasive disorders.

7. Bibliographical References

Bohrer, B. K., Foye, U., and Jewell, T. (2020). Recovery as a process: Exploring definitions of recovery in the context of eating-disorder-related social media forums. *International Journal of Eating Disorders*, 53(8):1219–1223.

Chung, C. and Pennebaker, J. W. (2007). The psychological functions of function words. *Social communication*, 1:343–359.

Curiskis, S. A., Drake, B., Osborn, T. R., and Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 57(2):102034.

Hunt, D. and Harvey, K. (2015). Health communication and corpus linguistics: Using corpus tools to analyse eating disorder discourse online. In *Corpora and Discourse Studies*, pages 134–154. Springer.

Kenny, T. E., Boyle, S. L., and Lewis, S. P. (2020). # recovery: Understanding recovery from the lens of recovery-focused blogs posted by individuals with lived experience. *International Journal of Eating Disorders*, 53(8):1234–1243.

Leis, A., Ronzano, F., Mayer, M. A., Furlong, L. I., Sanz, F., et al. (2019). Detecting signs of depression in tweets in spanish: behavioral and linguistic analysis. *Journal of medical Internet research*, 21(6):e14199.

Lenhart, A., Purcell, K., Smith, A., and Zickuhr, K. (2010). Social media & mobile internet use among teens and young adults. millennials. *Pew internet & American life project*.

Leonidas, C. and Dos Santos, M. A. (2014). Social support networks and eating disorders: An integrative review of the literature. *Neuropsychiatric Disease and Treatment*, 10:915.

Lukač, M. et al. (2011). Down to the bone: A corpus-based critical discourse analysis of pro-eating disorder blogs. *Jezikoslovlje*, 12(2):187–209.

Malson, H., Bailey, L., Clarke, S., Treasure, J., Anderson, G., and Kohn, M. (2011). Un/imaginable future selves: A discourse analysis of in-patients' talk about recovery from an 'eating disorder'. *European Eating Disorders Review*, 19(1):25–36.

McCaig, D., Bhatia, S., Elliott, M. T., Walasek, L., and Meyer, C. (2018). Text-mining as a methodology to assess eating disorder-relevant factors: Comparing mentions of fitness tracking technology across online communities. *International Journal of Eating Disorders*, 51(7):647–655.

McCaig, D., Elliott, M. T., Siew, C. S., Walasek, L., and Meyer, C. (2019). Profiling commenters on mental health–related online forums: A methodological example focusing on eating disorder–related commenters. *JMIR mental health*, 6(4):e12555.

McCaig, D., Elliott, M. T., Prnjak, K., Walasek, L., and Meyer, C. (2020). Engagement with myfitnesspal in eating disorders: Qualitative insights from online forums. *International Journal of Eating Disorders*, 53(3):404–411.

Mihalcea, R. and Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 309–312.

Moessner, M., Feldhege, J., Wolf, M., and Bauer, S. (2018). Analyzing big data in social media: Text and network analyses of an eating disorder forum. *International Journal of Eating Disorders*, 51(7):656–667.

Mullany, L., Smith, C., Harvey, K., and Adolphs, S. (2015). 'am i anorexic?'weight, eating and discourses of the body in online adolescent health communication. *Communication & medicine*, 12(2-3):211–223.

Oh, J. S., He, D., Jeng, W., Mattern, E., and Bowler, L. (2013). Linguistic characteristics of eating disorder questions on yahoo! answers–content, style, and emotion. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–10.

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

Say, B. and Akman, V. (1996). Current approaches to punctuation in computational linguistics. *Computers and the Humanities*, 30(6):457–469.

Serpell, L., Treasure, J., Teasdale, J., and Sullivan, V. (1999). Anorexia nervosa: friend or foe? *International Journal of Eating Disorders*, 25(2):177–186.

Shen, J. H. and Rudzicz, F. (2017). Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326.

Wolf, M., Sedway, J., Bulik, C. M., and Kordy, H. (2007). Linguistic analyses of natural written language: Unobtrusive assessment of cognitive style in eating disorders. *International journal of eating disorders*, 40(8):711–717.

Zhou, S., Zhao, Y., Bian, J., Haynos, A. F., and Zhang, R. (2020). Exploring eating disorder topics on twitter: Machine learning approach. *JMIR Medical Informatics*, 8(10):e18273.