Anorexia Topical Trends in Self-declared Reddit Users

Razan Masood[♣], Mengjiao Hu[♣], Hermenegildo Fabregat[♢], Ahmet Aker[♣], and Norbert Fuhr[♣]
[♣] University of Duisburg-Essen, Duisburg, Germany
[♢] Universidad Nacional de Educación a Distancia, Madrid, Spain
firstname.lastname@uni-due.de
gildo.fabregat@lsi.uned.es

ABSTRACT

Social Media platforms have been a vital environment to share experiences and seek knowledge. People with various interests form online communities in which they can accumulate many experiences from many peers. Among these communities are the mental health-related ones that have been growing on Social Media in the last few years. However, users can show alarming behavioral signs at the stage of their mental illness that should be identified before it is too late. Hence, equipping social media platforms with the needed tools to monitor its users, identify risks, and intervene on time has been of great concern recently. In this paper, we target users who self disclose as being diagnosed with an eating disorder, namely Anorexia. We provide a dataset of manually labeled Reddit users' posts, focused on the extraction of some potentially relevant topics for the study of eating disorders. E.g. *diets*, *exercises*, *body image*, etc. These topics can be utilized to find patterns in Anorexic users' behaviors to distinguish them from users who are less likely to have Anorexia. They can also be used to interpret afflicted users' attitudes. We support our labeling with baseline experiments to learn how to differentiate between these topics.

CCS CONCEPTS

• **Human-centered computing** → **Social networking sites**; • **Applied computing** → *Psychology*.

KEYWORDS

mental health, Reddit, social media, Anorexia, machine learning

1 INTRODUCTION

Humanity has come a long way in maintaining a high level of societies' and individuals' well-being, including physical health, education and freedom. Still, a lot needs to be done in the mental health domain, which is getting more attention in the modern age of prospering technologies [17]. More people with mental health issues resort to Social Media (SM) platforms either to directly seek support and information or to communicate their thoughts and feelings indirectly. Recently, the data that such users produce on SM has proved to predict their mental health state and its severity [7]. Besides, it provides precious resources for practitioners and experts as a possible tool for mental health-related research. Moreover, predicting mental health issues in the early stages is essential to provide the needed support in alarming situations like preventing suicide, self-harming, and eating disorders [12, 16, 28].

In this paper, we target SM users who have explicitly stated that they were clinically diagnosed with Anorexia Nervosa (AN). Anorexia is "an eating disorder characterized by abnormally low body weight, an intense fear of gaining weight, and a distorted perception of weight. People with anorexia place a high value on controlling their weight and shape"[1]. We use posts extracted from *Reddit*, "an online network of communities based on people's interests". The different *Reddit* communities are referred to as subreddits. Each subreddit is devoted to a specific topic. Plenty of subreddits are related to AN and other eating disorders such as *EatingDisoders* and *AnorexiaRecovery* subreddits. People resort to such communities for many purposes. Some communities promote sharing recovery experiences and emotional support, and others can cause more harm like pro-Anorexia communities, which promote unhealthy body-image and diets. Hence, the chances are that there are users who may face serious risks, which obliged SM platforms to keep their environments under control and provide possible intervention when needed.

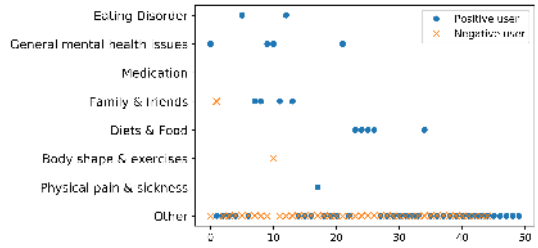
By investigating the specific type of information or topics AN diagnosed users post about, we observed that the most frequently discussed topics are *diet and eating routines*, *weight*, *family and relationship issues*, *anxiety*, and *depression* problems. Figure 1 shows an example of a pair of a positive user (diagnosed with AN) and a negative user (not diagnosed with AN). The timeline of the first 50 posts for the two users and their post topics are plotted. The example shows that the positive user (blue) posts more frequently on topics related to their mental health state (4 times), eating disorders (2 times), diet (4 times) and physical pain (1 time). On the other hand, the negative user posts (orange) about family and exercises, among other non-significant topics. Based on this analysis, we suggest that we can use particular topical patterns to analyze and explain AN users behaviour in a more understandable way, which can be helpful to distinguish risky users. Furthermore, when topical patterns are combined with, e.g., emotions [5, 8] or other aspects like stance, they can help to reveal the severeness level of the illness [26]. Besides, these patterns could be extended and adapted to other mental health issues such as substance abuse and depression. Hence, we believe that post-level classification could be useful for medical researchers and psychiatrists to analyze topical extracts of SM history and evaluate the prevalence of the pattern of certain topics among AN sufferers.

Our main contributions in this paper are as follows: (1) We define topics of importance to identify *Reddit* users who are more likely to have AN. (2) We provide a dataset of users' posts annotated with defined relevant topics. (3) We present baselines to predict the different posts categories based on the labeled dataset [2].

[1] https://www.mayoclinic.org/diseases-conditions/anorexia-nervosa/symptoms-causes/syc-20353591

[2] The dataset and best performing models code are released for research purposes https://github.com/razanmasood/Anorexia_Topical_Trends_in_Self_declared_

Figure 1: Positive vs. negative users' posts topics shown for the first 50 posts in two users' timeline. The x-axis shows the order of the posts, and the y-axis the assigned topics.



2 RELATED WORK

Online Social Media has driven a wide range of investigations on mental health by exploiting the growing users' data [14, 27]. Many datasets have been collected from SM platforms for language or communication analysis and risk prediction [10, 26, 30]. The datasets collection is based on different rules and methods [15]. One method is to identify users affected by mental illnesses using psychiatric surveys assessed by experts. The selected users' SM accounts are then explored based on the results of the survey [13].

Another method is to consider users who have mentioned that they have been diagnosed with a mental illness on their social media as positive cases [9, 18]. Nonetheless, the datasets mentioned above are labeled on users level, i.e a user have or does not have the targeted mental illness.

A third method is to annotate posts based on the signals it holds and that characterize the mental illness in question manually. The annotations are either determined from the data or based on theory [15]. The vast majority of post level annotations regarding eating disorders characteristics were done as part of content analysis work by experts. Mowrey et al. defined an annotation scheme for labeling tweets according to depressive symptoms and psychosocial stressors [21]. The goal of the final corpus is to understand the depression language and to identify the differences between psychological factors. Moreover, Sowles et al. extended the annotation to coding the attitude and the support behaviors of the comments [23]. On the other hand, the frequent topics brought up by mental health online communities has been explored using topic modeling such as LDA (Latent Dirichlet Allocation) and other methods [7, 22, 29].

However, the problem with automatic topic modeling, when applied to Reddit posts, is that posts as documents are not long enough for topic modeling. Moreover, when all posts of a user are joined in one document, it is more likely to undergo topic shifts, variation in tone, and hence, be out of context [10]. To our knowledge, a manually annotated post-level Reddit dataset for topics related to Anorexia is not available as a basis for both enhanced supervised and unsupervised classification models. Besides, our topic annotations are more descriptive than bare automatic topics. Our manual annotation criteria are defined based on the dataset observations and on the previous work that defined frequently mentioned topics by people who show symptoms of eating disorder online. Unlike the experts based annotations, we do not involve the attitude of the writer towards the mentioned topics in the post when assigning the labels.

3 DATASET

We use eRisk 2019 dataset [3]. eRisk is a part of CLEF (Conference and Labs of the Evaluation Forum) 2019 labs. The lab has designated a task for the early detection of Reddit users with signs of Anorexia [19]. The training data is a set of users under two categories. One is the users who stated in at least one of their posts that they were diagnosed with Anorexia, and the other category did not. For each user in the dataset, all posts and comments made by that user (which are up to 1000 posts and 1000 comments) are chronologically sorted [18]. The post in which a user declares their diagnosis was filtered out. Posts and comments of the users can belong to any subreddit. For our purposes, we selected 55 positive users and labeled 50-100 posts starting from the earliest post/comment. This research is oriented towards investigating the topics of interest of positive users, but to examine the occurrence of similar topics in negative users' posts, we picked ten negative users to label their posts using the same criteria.

3.1 Labels

To choose the posts' labels, we manually examined the different topics that appeared more frequently than other topics in positive users' posts. Then, we verified and expanded the topics using related work that analyzed posts of social media users, which indicate symptoms of Anorexia [6, 7, 29]. The selected topics are related to *mental health disorders* and *anxiety*, *self-harm*, *suicidal thoughts*, *pain*, and hints of the *desire to be skinny*. In addition, we used additional topics that were shown to be related to general mental health evaluation, like *family*, and *sleep*, as found in [11, 25]. The selected set of topics was rearranged under seven labels of posts with related AN topics and an additional label for posts that cannot be labeled under any of the seven defined topics. The number of labels is qualified for automatic classification experiments. The posts labels are as the following [4]: (1) *Eating disorder*: Posts with explicit mentions of experiences that indicate eating disorder (Anorexia, Bulimia, ED), and behaviors like binging and induced throw-ups. (2) *General mental health*: Under this label are posts with mentions of signs of mental disturbances and inconveniences. Examples of these are: a. Posts with indications of depression, anxiety, and sadness expressions. b. Posts with signs of harming oneself and suicidal expressions. c. Posts with mentions of issues related to sleep like lack of sleep or oversleep. d. Posts with mentions of alcohol drinking problems and other addiction issues like drugs and smoking. (3) *Medication*: Posts with mentions of medication names. Some medications could be used for treatment reasons or for inducing throw-ups. (4) *Family & friends*: Posts that contain stories on friends or family members. (5) *Diets & Food*: Posts with mentions of specific foods, recipes, and diets that include fasting, skipping meals, and purging. (6) *Body shape & exercises*: Posts with mention of the body's weight, height, BMI, and other body-image expressions. In addition to posts that mention exercise routines and

Reddit_Users. The labels are released by the IDs of the original dataset because of the signed user agreement.

[3] https://early.irlab.org/2019/index.html

[4] According to the user agreement signed with eRisk organizers, it is not allowed to show contents from the dataset.

Table 1: Labels with number of instances (#) for each and agreement scores considering (Fleiss Kappa κ). The number of instances for each label in Training, Development and Test sets are shown in the corresponding columns.

Label	κ	#	Train	Dev	Test
Eating disorder	0.72	126	85	24	17
General mental health	0.49	170	71	43	56
Medication	0.4	60	33	22	5
Family & friends	0.49	146	84	36	26
Diets & food	0.69	234	171	39	24
Body shape & exercises	0.75	280	188	48	47
Physical pain & sickness	0.49	130	108	8	14
Other	0.69	3959	2549	765	645

other physical activities. (7) *Physical pain & sickness*: Posts with mentions of physical sickness or illness. (8) *Other*: Any post not related to the categories mentioned above.

We define the topics in a way that makes it more straightforward for non-clinician annotators. The labels' definitions do no involve judgment on the severeness, emotions, or attitude the writer has towards the reported topics. This separation is necessary to ensure annotation with fewer inaccuracies and to separate emotions and attitude factors from the plain topic labels.

3.2 Annotation

Five master students from the Computer Science department annotated the data. The annotators were paid per hour. We dedicated a session to train the annotators and made sure that they follow the definition of the labels through a selected sample of posts. The posts were annotated with as many labels as the topics mentioned. Taking into account possible further experiments, we fixed one label for each post as the main label. We define the main label as the one that the annotator found being the most representative/dominant label of the post [4].

In the case of multiple labels, we calculated the agreement reached on individual labels using Fleiss' Kappa for multiple raters using the Python library statsmodels 0.11.0[5]. Because each post can have multiple labels, we calculate the agreement for each label separately, i.e., to observe the agreement between raters to choose a specific label for the post. The agreement results are shown in the second column of Table 1. The least agreement value we get is on the label *Medication*, which can be due to fewer posts available on the topic. Another reason is that the annotators are not experts, and in many cases, it is hard to recognize medication names easily. The third column of Table 1 shows the number of instances for each label taken as a main label only. We use Fleiss' Kappa as well to compute the agreement on the main label. The agreement score obtained is **0.65**, which is considered to be in an acceptable range [20].

4 AUTOMATIC TOPIC CLASSIFICATION

To further understand the complexity of classifying users' posts according to the defined topic labels, we present baseline results obtained using well-known classification approaches based on Logistic Regression Classifier, LSTM (Long Short-Term Memory), and CNN (Convolutional Neural Networks). Firstly, we divided the corpus into three sets (training, development, and test), where each set comprised different users to avoid learning user-specific features like individual writing styles. We choose to experiment with the main label only rather than dealing with multiple labels for a post. The distribution of posts can be seen in Table 1. Then, we pre-process the posts' textual content by lemmatizing, lower-casing, and removing expressions related to the Reddit platform like tagging forums and users. The cleaned posts and comments are the input to the ML models, and the assigned labels are the targets to be learned.

To analyze the task at different levels of complexity, we consider two experimental frameworks, namely, binary and multiclass classifiers. For the binary task, we transform our eight labels into two labels, one is the *Related* label that has all the seven labels relevant to AN, and the *Unrelated* label that has the *Other* label. The second task is a fine-grained multi-classification task that is set to distinguish the eight labels individually.

The three used models are set as the following:

Logistic Regression with TF-IDF (LR-TFIDF). We use the logistic regression implementation by Python's Sklearn package. We feed the classifier with Term Frequency-Inverse Document Frequency (TF-IDF) features of uni- and bi-word grams.

LSTM with inner-attention (LSTM-Att). For this model, we represented each post by its term embeddings extracted using GloVe [3]. Then each term was weighted by the average value of the embeddings of certain recurrent terms. The recurrent terms are selected by extracting the significant terms for each label against the other labels by the Chi-square test on TF-IDF features of unigram words. We selected the most significant 200 terms for each label. We then calculated the average embedding GloVe vector for each set of terms for each label. The inner attention mechanism is based on weighting each term of a post/comment by the average vector of each label. The model was implemented as in [1, 24]. For the LSTM, we used a single forward layer, eight neurons, and Hyperbolic Tangent activation function.

CNN with Meta-Map (CNN-MM) with which we explored the addition of more focused knowledge using concepts extracted by Meta-Map, an NLP tool focused on information retrieval from the biomedical domain and enriched with several thesauri [2]. As Meta-Map provides for each identified concept the semantic category to which it belongs, we explored an approach using this knowledge. In total, we studied 50 semantic groups manually selected based on their relationship with the labels, including *Activity*, *Behavior*, *Disease or Syndrome*. In short, each post has been represented as a sequence of terms and its respective category. We used GloVe to represent the words and a trainable embedding vector of 50 dimensions to represent the semantic categories. The CNN is applied with a fixed window of 5 elements and a total of 128 neurons.

4.1 Results and Error Analysis

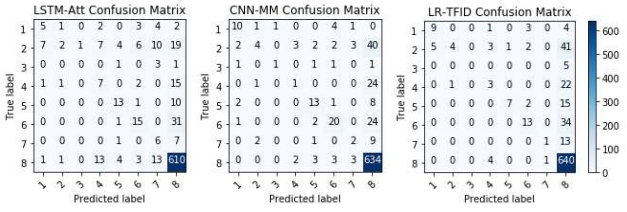
The overall classification performance including precision (P), recall (R) and F1 shown in Table 3 confirms that the multiclass classification is trickier than the binary one. CNN, combined with Meta-Map

[5]https://www.statsmodels.org/stable/generated/statsmodels.stats.inter_rater.fleiss_kappa.html

Table 2: Performance of the models reported on each label individually on the test set. (1) Logistic Regression with TF-IDF features (LR-TFIDF), (2) LSTM with attention (LSTM-Att), and (3) CNN with MetaMap (CNN-MM)

	Eating disorder			General mental health			Medication			Family & friends			Diets & Food			Body shape & exercises			Physical pain & sickness			Other		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
M/LR-TFIDF	**0.64**	0.53	0.58	**0.80**	**0.07**	**0.13**	0	0	0	**0.27**	0.12	0.16	**0.88**	0.29	0.44	**0.65**	0.28	0.39	**0.50**	0.07	0.12	0.83	**0.99**	0.90
M/LSTM-Att	0.36	0.29	0.32	0.4	0.04	0.07	0	0	0	0.24	**0.27**	**0.25**	**0.54**	0.54	**0.54**	0.50	0.32	0.39	0.17	**0.43**	**0.24**	**0.88**	0.95	0.91
M/CNN-MM	0.62	**0.59**	**0.61**	0.50	**0.07**	0.12	**0.50**	**0.20**	**0.29**	0.17	0.04	0.06	0.59	**0.54**	**0.57**	**0.65**	**0.43**	**0.51**	0.22	0.14	0.17	0.86	0.98	**0.92**

Figure 2: Confusion matrices obtained on test set. The numbers refer to the labels in the order they are mentioned in section 3.1



(a) (b) (c)

semantic groups, performed significantly better than the other two models for the multiclass task[6].

We further list detailed results on each of the labels in multiclass settings in Table 2. The highest performance according to F1 measure are on *Eating disorder*, *Diets & food*, *Body shape & Exercises* and *Other* label. The unbalanced distribution of labels highly influences the performance. Hence, the high performance on the *Other* labels. CNN-MM performed well on the labels that contained terms related to the medical semantic groups in Meta-Map in their assigned posts, e.g., Body parts terms and Eating disorders terms. However, CNN-MM performed better than the other models on the *Medication* Label despite the few items in training data due to the medication terms used for features encoding. However, the performance on the *General mental health* label is not as expected. This might be due to the fact that this label involves multiple domains that confused the classifiers. In other words, the confusion matrice in Figure 2(a) shows that the LSTM-Att model confused *General mental health* label (2) mostly with *Physical pain & sickness* label (7) besides the *Other* label (8). This confusion can be due to that many posts that mention general mental health issues like lack of sleep also mention pain aspects, which made the classifier choose the *Physical pain* label (7). LSTM-Att also shows better performance on *Family & Friends* label (4) as it uses terms related to this topic to weigh the post terms, unlike the CNN model (Figure 2(b)) with which Meta-Map does not have such terms.

The classification results show that terms play an important role in identifying the topics. This is shown by the improvement in performance when achieved when supporting the models with targeted related terms in comparison with using TF-IDF features alone (Figure 2(c)). Nevertheless, the problem is that a post, especially the longer ones, can discuss many topics. Therefore, we suggest ML models with multiple outputs of labels scores. Besides, the labeling process can be enhanced by highlighting the related sentences according to each label rather than labeling longer posts to make the labeled text more focused.

[6]McNemar's test, $p < 0.0125$ after Bonferroni correction.

Table 3: Results using Binary (B) and Multi-class (M) classification with each of the models Logistic Regression with TF-IDF features (LR-TFIDF), LSTM with attention (LSTM-Att), and CNN with MetaMap (CNN-MM)

	Dev			Test		
Model	P	R	F1	P	R	F1
B/LR-TFID	0.83	0.80	0.81	0.77	0.76	0.77
B/LSTM-Att	0.84	0.78	0.80	0.85	**0.80**	0.82
B/CNN-MM	**0.85**	0.80	**0.82**	**0.88**	0.79	**0.83**
M/LR-TFIDF	0.42	0.29	0.33	**0.57**	0.29	0.34
M/LSTM-Att	0.39	0.33	0.33	0.39	0.35	0.34
M/CNN-MM	**0.52**	**0.37**	**0.41**	0.51	**0.37**	**0.41**

5 CONCLUSIONS AND FUTURE WORK

In this paper, we report the annotation process of Reddit posts and comments by self-declared users with Anorexia Nervosa. We define an annotation scheme of fine-grained labels according to topics related to the diagnosis of Anorexia Nervosa. We show that our annotation is rather robust as the Fleiss' Kappa agreement values are in an acceptable range. We further test the possibility of predicting post topics automatically. The classification results show that predicting one main label for long posts is tricky to perform accurately. Hence, making use of the multiple-label annotations to predict multiple labels for each post can be a possible solution in addition to specifying sentence-level annotations. The annotation scheme provided in this paper is related to the topics of which self-declared Reddit users of AN mention more frequently than the other users. However, these topics are not enough to distinguish risky users as this might lead to false-positive predictions because many users use these communities because they want to help someone related to them who are diagnosed with AN. Onward, in our future work, we will explore the possibility of employing the topics to predict risky users. The prediction models can be enriched with the sequential development of emotions and stances that accompany the topics [5, 8]. Furthermore, the different features combinations allow the estimation of the severeness level of the targeted illness. Also, what can be quite interesting is how to make these models adapt and be diverse to learn different forms of mental illnesses.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2167, Research Training Group "User-Centred Social Media". The work has been also partially supported by the Spanish Ministry of Science and Innovation within the projects PROSA-MED (TIN2016-77820-C3-2-R) and EXTRAE-II (IMIENS 2019).

REFERENCES

[1] Ahmet Aker, Alfred Sliwa, Fahim Dalvi, and Kalina Bontcheva. 2019. Rumour verification through recurring information and an inner-attention mechanism. *Online Social Networks and Media* 13 (2019), 100045.

[2] Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 17.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] Victoria Bobicev and Marina Sokolova. 2017. Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective.. In *RANLP*. 97–102.

[5] Craig J Bryan, Jonathan E Butner, Sungchoon Sinclair, Anna Belle O Bryan, Christina M Hesse, and Andree E Rose. 2018. Predictors of emerging suicide death among military personnel on social media networks. *Suicide and Life-Threatening Behavior* 48, 4 (2018), 413–430.

[6] Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina J Costello, Nina Kaiser, Elizabeth S Cahn, Ellen E Fitzsimmons-Craft, and Denise E Wilfley. 2019. "I just want to be skinny.": A content analysis of tweets expressing eating disorder symptoms. *PloS one* 14, 1 (2019), e0207506.

[7] Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1171–1184.

[8] Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 1653–1660.

[9] Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258* (2018).

[10] Arman Cohan, Sydney Young, Andrew Yates, and Nazli Goharian. 2017. Triaging content severity in online mental health forums. *Journal of the Association for Information Science and Technology* 68, 11 (2017), 2675–2689.

[11] Pricewaterhouse Coopers. 2015. The costs of eating disorders: Social, health and economic impacts. *B-eat, Norwich* (2015).

[12] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. 51–60.

[13] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

[14] 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)*. IEEE, 568–573.

[15] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43–49.

[16] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805.

[17] James Lake and Mason Spain Turner. 2017. Urgent need for improved mental health care and a more collaborative model of care. *The Permanente Journal* 21 (2017).

[18] David E. Losada and Fabio Crestani. 2016. A Test Collection for Research on Depression and Language use. In *Conference Labs of the Evaluation Forum*. Springer, 28–39. https://doi.org/10.1007/978-3-319-44564-9_3

[19] David E. Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of eRisk 2019: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019*. Springer International Publishing, Lugano, Switzerland.

[20] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

[21] Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, and Mike Conway. 2017. Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study. *Journal of medical Internet research* 19, 2 (2017), e48.

[22] 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*. 99–107.

[23] Shaina J Sowles, Monique McLeary, Allison Optican, Elizabeth Cahn, Melissa J Krauss, Ellen E Fitzsimmons-Craft, Denise E Wilfley, and Patricia A Cavazos-Rehg. 2018. A content analysis of an online pro-eating disorder community on Reddit. *Body image* 24 (2018), 137–144.

[24] Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *arXiv preprint arXiv:1802.05758* (2018).

[25] Andrew Toulis and Lukasz Golab. 2017. Social Media Mining to Understand Public Mental Health. In *VLDB Workshop on Data Management and Analytics for Medicine and Healthcare*. Springer, 55–70.

[26] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2018. Social interactions in online eating disorder communities: A network perspective. *PloS one* 13, 7 (2018).

[27] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848* (2017).

[28] Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (2015), 1036–1040.

[29] Sicheng Zhou, Yunpeng Zhao, Rubina Rizvi, Jiang Bian, Ann F Haynos, and Rui Zhang. 2019. Analysis of Twitter to Identify Topics Related to Eating Disorder Symptoms. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–4.

[30] Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. 24–33.