A Dataset for Research on Depression in Social Media

Conference Paper · July 2020

DOI: 10.1145/3340631.3394879

CITATIONS
2

READS
2,288

3 authors:

Esteban Andrés Ríssola
University of Lugano
13 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Seyed Ali Bahrainian
École Polytechnique Fédérale de Lausanne
21 PUBLICATIONS **255** CITATIONS
SEE PROFILE

Fabio Crestani
University of Lugano
415 PUBLICATIONS **5,470** CITATIONS
SEE PROFILE

A Dataset for Research on Depression in Social Media

Esteban A. Ríssola
esteban.andres.rissola@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

Seyed Ali Bahrainian
bahres@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

Fabio Crestani
fabio.crestani@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

ABSTRACT

Language provides a unique *window* into thoughts, enabling direct assessment of mental-state alterations. Due to their increasing popularity, online social media platforms have become promising means to study different mental disorders. However, the lack of available datasets can hinder the development of innovative diagnostic methods. Tools to assist health practitioners in screening and monitoring individuals under potential risk are essential.

In this paper, we present a new a dataset to foster the research on automatic detection of depression. To this end, we present a methodology for automatically collecting large samples of depression and non-depression posts from online social media. Furthermore, we perform a benchmark on the dataset to establish a point of reference for researchers who are interested in using it.

KEYWORDS

Online mental state assessment, social media, text mining

ACM Reference Format:
Esteban A. Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2020. A Dataset for Research on Depression in Social Media. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20), July 14–17, 2020, Genoa, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340631.3394879

1 INTRODUCTION

Worldwide more than 350 million people, of different ages and communities, suffer from depression[1]. The lack of an appropriate treatment can lead to disability, psychotic episodes, thoughts of self-harm and, at its worst, suicide. For this reason, the importance of mental health has driven the search for cutting-edge and innovative methods for collecting reliable information and evidence about mental disorders. Many useful cues about an individual's mental state [26, 27] as well as personality [22, 29], social and emotional conditions, and even personal values [5] can be discovered by examining the patterns of language use [1, 8]. The main reason, as argued by Pennebaker et al. [25], is because such latent mental-related variables are encoded in the words that people use to communicate.

Despite the great importance of mental health research through the analysis of online social media activities, datasets for the automatic detection of depression are limited and resources are very scarce [13]. This is mainly due to the need for the collection of large amounts of labelled data which is usually a complex and time-consuming endeavour. In order to solve this problem, we present a weak-supervision framework for collecting such data.

The main contributions of this work are as follows:

- We present a methodology for automatically gathering depression and non-depression post samples[2].
- We release the dataset created following the proposed methodology which can advance research on depression detection[3].

To the best of our knowledge, this is the first step towards automatically deriving large samples of data for the study of depression in online social media settings. We empirically validate our methodology and show it can be effectively used for automatically collecting posts samples. Furthermore, we release the dataset created and present a benchmark to foster the research on data-driven approaches for automatic identification of depression.

The remainder of the paper is organised as follows: Section 2 summarises existing and similar datasets; Section 4 outlines the methodology conducted to create the dataset; Section 5 presents the empirical evaluation conducted to validate the proposed methodology. Finally, conclusions are presented in Section 6.

2 EXISTING DATASETS

The Computational Linguistics and Clinical Psychology [10] was the first initiative to foster research on the interaction between language and mental disorders in online social media. They collected a dataset which contained instances of individuals who have experienced depression and post-traumatic stress disorder from Twitter[4]. In a similar fashion, the Early Risk Prediction on the Internet (eRisk) [17] initiative gathered a collection of users who suffered from depression and eating disorders from Reddit[5]. Both CLPysch and eRisk datasets were created following the methodology proposed by Coppersmith et al. [9].

Shing et. al [30] created a dataset for studying the assessment of suicide risk in Reddit. Positive instances of suicidality, that are users at risk of suicide, were collected based on their participation in a discussion forum (better known as *subreddits*[6]) called *SuicideWatch*. Such users were assessed by human annotators, including clinicians and crowsourcing workers.

Jamil et al. [14] gathered a collection of Twitter posts from individuals who took part in #BellLetsTalk 2015, a campaign to help

[1] See: https://www.who.int/mental_health/advocacy/en/#Factsheets

[2] Code available at: https://github.com/earissola/umap20
[3] To get access to the data contact the first author.
[4] See: https://twitter.com
[5] See: https://www.reddit.com
[6] Titled forums on Reddit are denominated *subreddits*.

UMAP '20, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-6861-2/20/07…$15.00
https://doi.org/10.1145/3340631.3394879

decrease social stigma and foster awareness and understanding of mental disorders. These tweets were examined by human annotators to verify whether they were providing some evidence of depression or not. The resulting dataset was highly imbalance as only a five percent of the tweets contained references to depression and 95% did not.

3 MOTIVATIONS

One drawback of the studies conducted using the collections presented in Section 2 is that users are usually represented by the concatenation of all their posting history [18]. This assumption might not result completely accurate. In fact, it is unlikely that every message posted by an individual potentially affected by a mental disorder could provide some evidence that could be used to identify the onset of such disorder [6]. If processed automatically, such messages might introduce unnecessary noise in the development and tuning of classification systems used for risk-assessment and decision-making.

Our goal is to build a post-level annotated dataset large enough to enable the development of robust models. Such models should be able to extract evidences from each message posted by a user who could be potentially tracked in order to anticipate the onset of depression. The strength and trend of such evidences can be utilised to detect the point where the mental disorder is starting to manifest [28]. Furthermore, post-level annotations provide a useful way to discover life circumstances related with mental disorders not captured by traditional depression diagnostic criteria [2]. Finally, we sought to alleviate the great effort and time that involves manually annotating individual posts by proposing a set of heuristics to automatically gathering samples of posts providing some evidence of depression signs. Thus, we are able to obtain higher-level supervision over unlabelled data.

4 METHODOLOGY

In this section we outline the methodology we propose to automatically generating large datasets of depression and non-depression posts. Consider a set of social media users that we have definitive knowledge that they are suffering from depression. This knowledge could come from a survey (such as the PHQ-9 [15][7]) or it could be self-declared. Given a chronology of textual posts and based on previous findings in the literature, we propose different heuristics to characterise depression signs and use this information to automatically selecting posts for building the dataset.

Let D^+ be the candidate set of positive posts samples. We retrieve such posts from a set of users suffering from depression. Since the goal is to filter out less useful messages, we define two heuristics:

- Filtering posts by their sentiment polarity score
- Filtering posts by their topical similarity with a *depression taxonomy*.

Let D^- be the control posts samples, that are the posts not providing any reference to depression signs. Such posts are randomly collected from a set of users which are not affected by the mental disorder. In the following subsections we explain in detail how we apply the heuristics to create the dataset.

[7] *PQH* stands for *Patient Health Questionnaire.*

4.1 Heuristic: Sentiment Polarity Score (H_s)

Research has shown that the sentiment polarity score (also referred to as *semantic orientation* or *valence*) of a post can be linked with the emotions evoked by a piece text [23]. Based on previous findings [3, 7, 24] and our own empirical analysis, we hypothesise that when this value is negative (*i.e.*, below zero) it can be a good indicator of distress or unhappiness, especially when the posts are written by users experiencing depression. The sentiment polarity score of a post ranges from -1 to 1 and is calculated using a lexicon-based approach. Given the computed scores, the posts in D^+ are sorted from lower to higher polarity retaining those messages whose score is less/equal to 0. We use TextBlob[8] to obtain the polarity score of a post.

According to the *Diagnostic and Statistical Manual of Mental Disorders* [2] (DSM)[9] depressive moods are characterised by the predominance of two emotions: *sadness* and *disgust*. Based on psychology literature, we refine this heuristic by mandating the presence of words related with these emotions in the posts. The goal is to clean the set of messages obtained after filtering by sentiment polarity score from *false positive* cases, such as "*That's the worst name I ever heard*" (this message also has a low sentiment polarity score). To this end, we utilise the terms included in EmoLex [20].

Additionally, we compute for each post a *depression score*. To calculate this score we use the NRC Affect Intensity Lexicon (AIL) [19] which associates words with real-valued scores of intensity for different emotions. Given a word and an emotion, the intensity ranges from 0 to 1, where an intensity of 1 means that the word evokes the highest intensity of that emotion. The sadness score is computed as the average of the intensities of the words which evoke "sadness"[10] found in the text of a post. The rationale of calculating this score is to be strict with the words considered to evoke sadness. Moreover, the AIL includes words that might not predominantly convey a certain emotion and still tend to co-occur with words that do. For instance, the words *failure* and *death* describe concepts that are usually accompanied by sadness and, thus, they denote some amount of sadness.

Hence, following heuristic H_s a post in D^+ is considered a valid candidate for the final set if it contains at least one word related with either sadness or disgust and a depression score higher than a certain threshold. We decide to use a threshold of 0.1 as we want to omit words marginally related with depression without being too stringent. A sensitivity analysis to determine the best threshold is left for the future.

4.2 Heuristic: Topical Similarity (H_t)

We define H_t as a heuristic in which less useful posts in D^+ are filtered out based on their topical similarity with a *depression taxonomy*. We first build this topic using the lexicon released by Choudhury et al. [7]. This lexicon consists of words closely associated with texts written by individuals discussing depression or its symptoms in online settings. We enlarge this set by collecting all possible

[8] See: https://textblob.readthedocs.io/en/dev/index.html

[9] The DSM determines a common vocabulary and standard criteria to group and characterise the different mental disorders. Its three main components are: the diagnostic classification, the diagnostic criteria sets and the descriptive text.

[10] The lexicon only provides scores of intensity for four basic emotions: anger, fear, joy, and sadness.

online vocabularies with concepts and terms commonly related to depression. The goal is to select from all these words those that are considered closely associated with this mental disorder and define a compact but accurate list. With the aid of three human experts (one experienced clinical psychologist and two computational linguistic experts, none co-authoring this paper) we selected a subset of the words obtaining a list of 78 depression-related terms[11] which comprised the depression taxonomy.

Subsequently, we use Latent Dirichlet Allocation [4] (LDA) to obtain the topics that emerged from the posts in D^+ and compute the cosine similarity between each post and the depression taxonomy. The goal is to obtain a ranked list of the posts according to their similarity with the taxonomy of depression. Thus, the higher ranked posts are those with the highest association to depression. In order to compute the ranked list, each post is mapped into the LDA space and compared with the depression taxonomy as a reference point. The optimal number of topics in our trained LDA is 200 and this number is obtained following the method of Griffiths et al. [12].

4.3 Resulting Dataset

In order to automatically select posts of our dataset we use the eRisk 2018 collection[12] [16]. This publicly available collection consists of two groups of Reddit users, namely depressed and non-depressed. The maximum number of posts per user is 2,000 and the labels are assigned to users but not to individual posts. It is noteworthy that the methodology we propose can be potentially applied to any collection where information about the depression status of the users is available. For example, in the case of eRisk 2018 collection some users have self-declared being diagnosed with depression.

Following H_s the candidate posts in D^+ are ranked according to their polarity and sadness scores. We select the top 3500 posts to build the final set, which is defined as D_s^+. In a similar way, applying H_t we create a second set of posts, defined as D_t^+, where we select the top 3500 posts with the highest topical similarity to the depression taxonomy. Instead, D_s^- and D_t^- are two sets of 3500 posts each that are randomly sampled from set of users which are not diagnosed with depression and used them as control sets.

To have a better understanding on how the data is distributed, Figure 1 depicts the Gaussian kernel density estimation (KDE) of the automatically derived sets. We observe that there is some overlap between the curves. This intersection is desirable since it enables to improve the decisions in the boundary cases. We empirically found that selecting samples with no such overlap has a negative impact on the performance of a depression post classifier, caused by an overfit. For example, if the vocabularies obtained from the posts in D^+ and D^- are disjoint and too specific they hamper the classifier to learn a useful representation.

The goal of Figure 2 is to study whether some relationship between the posts selected with H_s and H_t exists. A Pearson's correlation coefficient of 0.1484 (significant at p-value < 0.001) denotes a very small positive correlation between the topic similarity score and the polarity/sadness scores (top plot). We also analyse whether some relationship holds between the polarity and sadness scores both used in the definition of H_s. A Pearson's correlation coefficient of -0.1388, shows a small negative correlation (significant at p-value < 0.001) between them (bottom plot). In conclusion, there is no considerable correlation between variables analysed.

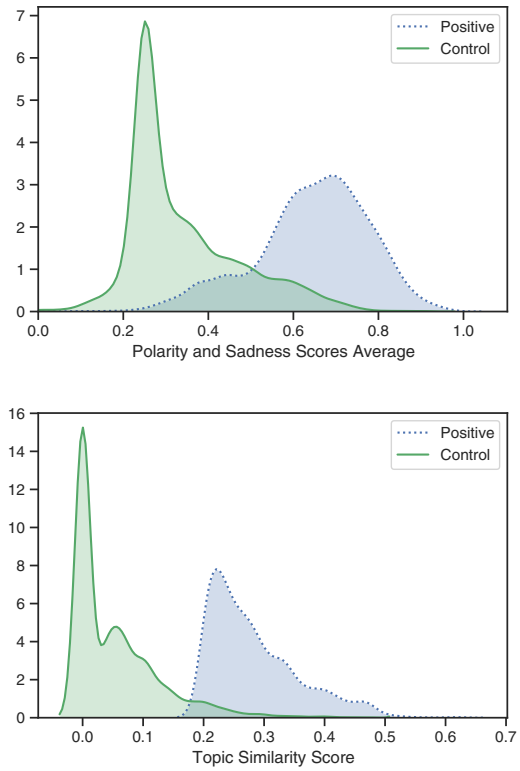

Figure 1: KDE computed for the automatically derived sets. On the top, candidate posts in D^+ are selected based on H_s (Section 4.1). While on the bottom, the posts are selected according to H_t (Section 4.2). In both cases, control posts samples are randomly picked from set of users which are not affected by depression.

5 EMPIRICAL VALIDATION

In this section we describe the process conducted to validate both the methodology and the resulting dataset. The goal is to determine whether the two automatically generated sets, created following H_s and H_t, can be used to train a classification model. Such model should be able to effectively distinguish posts providing some evidence of depression signs from those which do not.

Validation Set. We created a validation set of manually annotated posts. To this end, we randomly sampled a total of 400 posts, 200 from each class, in the eRisk 2018 collection[13] and asked three human experts, none co-authoring this paper, to label them. The annotation process followed a similar procedure to that defined by Moreno et al. [21]. Annotators were asked to determine which posts can be considered as a reference to depression following the DSM criteria. The references found should point to symptoms or

[11] Examples of words included in the list are: *insomnia*, *anxiety*, *suicidal*, *delusions*.
[12] See: https://early.irlab.org/2018/index.html

[13] These posts were removed from D^+ and D^-.

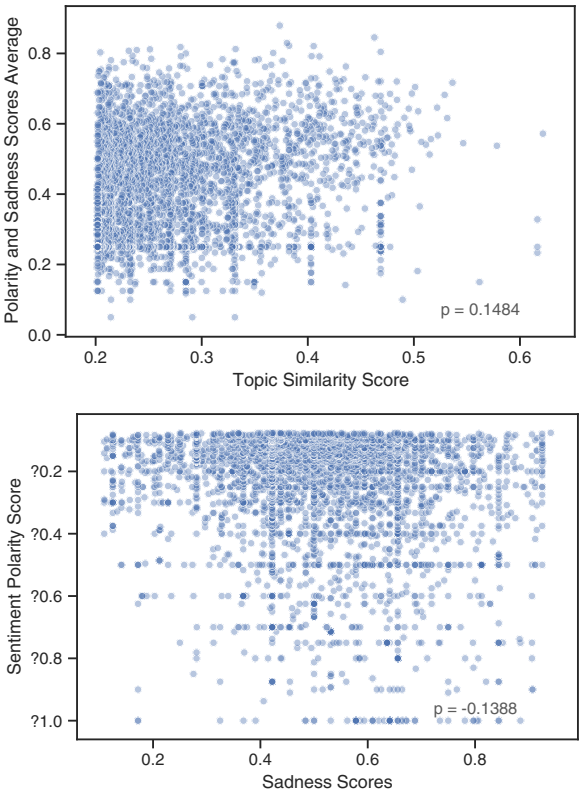

Figure 2: Correlation study. The top plot analyse the relationship between the posts selected following H_s and H_t, respectively. The bottom plot studies the association between sentiment polarity and sadness scores both used in the definition of H_s. Pearson's correlation coefficients are reported (both significant at p-value < 0.001).

feelings experienced by the individual and not by a third person. Furthermore, general comments about daily ordeals and common experience of having a bad day do not meet the criteria as a depression symptom.

Each message was assigned with one of the following codes:

(1) No depression reference is expressed;
(2) One or more depression references are expressed;
(3) Unable to make a judgement.

The three annotators achieved a pairwise Cohen's Kappa score ranging between 0.577 and 0.7492. Achieving a high inter-rater agreement can be a difficult task. In some cases taking a decision is complex without any additional information. The final validation set comprised of 55 positive posts (*i.e.*, references to depression) and 93 control posts.

Experiments and Results. Table 1 shows the results of various models trained separately on each automatically derived set and evaluated on the validation set previously described. The goal of the classification task is to identify posts that show some evidence of depression signs from those which do not. The effectiveness is evaluated using classic metrics such as Precision, Recall, F_1 and Area Under the Curve ROC (AUC).

We created the benchmark using Logistic Regression (LR) classifiers and testing a variety of commonly used features for text classification tasks. The features are the following: (1) "Unigrams" which represents the raw frequency count of the words extracted from each post; (2) "Unigrams++" which extends unigram features including the post length (word count), the sentiment polarity score and sadness scores obtained from the posts; (3) "LIWC" uses LIWC[14] psychometric word categories counts collected from the posts; (4) "BERT_Embeddings" BERT[15] are the contextual language representations derived from the posts . The LR classifiers are trained with L2 regularisation.

Table 1: Benchmark models trained on the automatically derived sets and evaluated using the validation set (manually annotated posts). All models are trained using Logistic Regression classifiers.

H_s	**Precision** (%)	**Recall** (%)	F_1 (%)	**AUC** (%)
Unigrams	**78.94**	81.81	80.35	**84.45**
Unigrams++	75.92	74.54	75.22	80.28
LIWC	64.78	83.63	73.01	78.37
BERT_Embeddings	72.97	**98.18**	**83.72**	**88.33**
H_t	**Precision** (%)	**Recall** (%)	F_1 (%)	**AUC** (%)
Unigrams	80.39	74.54	77.35	81.89
Unigrams++	**83.33**	72.72	77.66	82.06
LIWC	65.00	94.54	77.03	82.21
BERT_Embeddings	73.61	**96.36**	**83.46**	**87.96**

We observe that, overall a higher recall is achieved following heuristic H_s. While a higher precision is obtained when applying H_t. Given the nature of the task a higher recall is usually preferable. For instance, under certain circumstances, producing a considerable amount of false alarms (*false positives*) can be tolerated at the benefit of discovering most of the *real* cases. Finally, F_1 allows to conclude that any of the heuristics can effectively be used to automatically derived a set of posts samples and used such set to train a classification model. We hope this benchmark to serve as a starting point for further research.

6 CONCLUSIONS

Recently, there has been an increasing research interest in the identification of mental state alterations through the exploitation of online digital traces. The lack of available resources, such as datasets, might hinder the development of cutting-edge technologies to assist health practitioner in their daily labour. In this paper we introduced a methodology for automatically gathering post samples of depression and non-depression. Furthermore, we will release the automatically crated dataset, as well as the validation set of manually annotated posts, in order to contribute to the research on automatic detection of depression.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for the constructive suggestions. This work was supported in part by the Swiss Government Excellence Scholarships and Hasler Foundation.

[14]LIWC: *Linguistic Inquiry and Word Count* [31].
[15]BERT: Bidirectional Encoder Representations from Transformers [11].

REFERENCES

[1] Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clinical Psychological Science* 6, 4 (2018), 529–542.

[2] American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing, Washington.

[3] Seyed Ali Bahrainian and Fabio Crestani. 2017. Towards the Next Generation of Personal Assistants: Systems that Know When You Forget. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands*. 169–176.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022.

[5] Ryan L. Boyd, Steven R. Wilson, James W. Pennebaker, Michal Kosinski, David J. Stillwell, and Rada Mihalcea. 2015. Values in Words: Using Language to Evaluate and Understand Personal Values. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, UK*. 31–40.

[6] Patricia A. Cavazos-Rehg, Melissa J. Krauss, Shaina Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J. Bierut. 2016. A content analysis of depression-related tweets. *Computers in Human Behavior* 54 (2016), 351–357.

[7] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, USA*.

[8] Cindy Chung and James Pennebaker. 2007. The Psychological Functions of Function Words. *Frontiers of social psychology. Social communication* (2007).

[9] 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*. Baltimore, USA.

[10] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, USA*. 31–39.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018).

[12] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl 1 (2004), 5228–5235.

[13] 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, Supplement C (2017), 43–49.

[14] Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring Tweets for Depression to Detect At-risk Users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology - From Linguistic Signal to Clinical Reality, CLPsych@ACL 2017, Vancouver, Canada*. 32–40.

[15] Kurt Kroenke, Robert L. Spitzer, Janet B.W. Williams, and Bernd Löwe. 2010. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *General Hospital Psychiatry* 32, 4 (2010), 345 – 359.

[16] David E. Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of eRisk: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14*. 343–361.

[17] David E. Losada, Fabio Crestani, and Javier Parapar. 2019. Early Detection of Risks on the Internet: An Exploratory Campaign. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany*. 259–266.

[18] Razan Masood. 2019. Adapting Models for the Case of Early Risk Prediction on the Internet. In *Advances in Information Retrieval*, Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra (Eds.). 353–358.

[19] Saif Mohammad. 2018. Word Affect Intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan*.

[20] Saif Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[21] Megan A. Moreno, Lauren A. Jelenchick, Katie G. Egan, Elizabeth Cox, Henry Young, Kerry E. Gannon, and Tara Becker. 2011. Feeling bad on Facebook: depression disclosures by college students on a social networking site. *Depression and Anxiety* 28, 6 (2011), 447–455.

[22] Yair Neuman. 2016. *Computational Personality Analysis: introduction, practical applications and novel directions*. Springer.

[23] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 1–135.

[24] Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive Moods of Users Portrayed in Twitter. In *Proceedings of the ACM SIGKDD Workshop On Healthcare Informatics (HI-KDD)*.

[25] James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology* 54, 1 (2003).

[26] Daniel Preoţiuc-Pietro, Johannes C. Eichstaedt, Gregory J. Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle H. Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2015, June 5, Denver, USA*. 21–30.

[27] Esteban Andrés Ríssola, Mohammad Aliannejadi, and Fabio Crestani. 2020. Beyond Modelling: Understanding Mental Disorders in Online Social Media. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17*, Vol. 12035. 296–310.

[28] Esteban Andrés Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2019. Anticipating Depression Based on Online Social Media Behaviour. In *Flexible Query Answering Systems - 13th International Conference, FQAS 2019, Amantea, Italy, July 2-5, 2019*. 278–290.

[29] Esteban Andrés Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2019. Personality Recognition in Conversations using Capsule Neural Networks. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019*. 180–187.

[30] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. 25–36.

[31] Yla R. Tausczik and James W. Pennebaker. 2009. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* (2009).