An Exploratory Analysis of the Relation Between Offensive Language and Mental Health

Ana-Maria Bucur[1], Marcos Zampieri[2], and Liviu P. Dinu[1]
[1]University of Bucharest, Romania
[2]Rochester Institute of Technology, USA
ana-maria.bucur@drd.unibuc.ro, marcos.zampieri@rit.edu
ldinu@fmi.unibuc.ro

Abstract

In this paper, we analyze the interplay between the use of offensive language and mental health. We acquired publicly available datasets created for offensive language identification and depression detection and we train computational models to compare the use of offensive language in social media posts written by groups of individuals with and without self-reported depression diagnosis. We also look at samples written by groups of individuals whose posts show signs of depression according to recent related studies. Our analysis indicates that offensive language is more frequently used in the samples written by individuals with self-reported depression as well as individuals showing signs of depression. The results discussed here open new avenues in research in politeness/offensiveness and mental health.

1 Introduction

The use of offensive language is pervasive in social media and it has been studied from different perspectives. A popular line of research is the study of computational models to identify offensive content online relying on traditional machine learning classifiers (e.g. naive bayes and SVMs) (Xu et al., 2012; Dadvar et al., 2013), neural networks (e.g. LSTMs, GRUs) with word embeddings (Aroyehun and Gelbukh, 2018; Majumder et al., 2018), and more recently, transformer models like ELMO (Peters et al., 2018) and BERT (Devlin et al., 2019) which have shown to obtain competitive scores topping the leaderboards in recent shared tasks on offensive language and hate speech detection (Liu et al., 2019).

Offensive language is related to the notion of impoliteness (Culpeper, 2011) and it can take various forms from general and often harmless profanity to abusive language intended to cause harm, such as cyberbullying and hate speech (Waseem et al., 2017). Computational models have been applied not only to identify the various types of offensive content (Basile et al., 2019) but also to, for example, study the relation between profanity and hate speech (Malmasi and Zampieri, 2018) and the different functions and intentions of vulgarity in social media (Holgate et al., 2018).

Most of the datasets used in the aforementioned studies contain data sampled from the general population and therefore very little light has been shed on the use of offensive language in online communication by specific groups such as individuals with mental health conditions. A notable exception is the recent study by Birnbaum et al. (2020) which shows that users with mood disorders (bipolar disorder, major depressive disorder) and schizophrenia spectrum disorders use more swear words in their Facebook messages than healthy users.

To address this shortcoming, in this paper, we build on recent work on offensive language identification and apply it to mental health datasets. More specifically, we look at the role of offensive language in the communication of users with depression using two publicly available datasets containing posts by individuals with self-reported depression diagnosis.

To the best of our knowledge, this study is the first to apply state-of-the-art offensive language identification models to mental health datasets. We aim to answer two research questions:

RQ1: Are posts from individuals suffering from depression more likely to contain offensive language in existing datasets?

RQ2: Are there differences in the nature of offensive language used by individuals with depression compared to control groups?

WARNING: This paper contains offensive words.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3600–3606
August 1–6, 2021. ©2021 Association for Computational Linguistics

2 Related Work

Offensive language identification is a popular topic in NLP. Researchers have been working to improve the performance of systems trained to identify conversations that are likely to go awry (Zhang et al., 2018) and to detect the various types of offensive posts in social media (Basile et al., 2019; Kumar et al., 2020). More recently, with the goal of improving explainability, offensive language identification at the token-level has received more attention (Mathew et al., 2021; Ranasinghe and Zampieri, 2021). A number of computational models have been applied to this task ranging from traditional machine learning classifiers, most notably SVMs (MacAvaney et al., 2019), to various deep learning models (Liu et al., 2019). While the clear majority of studies on this topic deal with English, some studies have addressed offensive language in other languages like Greek (Pitenis et al., 2020) and Turkish (Çöltekin, 2020) while a few others have applied cross-lingual models to take advantage of existing English datasets when making predictions in languages with fewer resources (Ranasinghe and Zampieri, 2020).

Several studies have applied machine learning and NLP methods to address research questions related to mental health in social media such as identifying users with a particular mental health condition and predicting the risk of self-harm or suicide ideation (De Choudhury et al., 2013; Preoţiuc-Pietro et al., 2015; Malmasi et al., 2016; De Choudhury et al., 2016; Chancellor and De Choudhury, 2020). The CLPsych workshop co-located with international NLP conferences has hosted multiple competitions on these topics providing participants with important benchmark datasets and attracting a large number of teams (Coppersmith et al., 2015; Milne et al., 2016; Zirikly et al., 2019).

There have been multiple studies on the impact of offensive and hateful speech on the individual's psychological mental health and well-being (Bannink et al., 2014; Saha et al., 2019). The use of offensive language by individuals with mental health conditions, however, has not been substantially studies with the exception of Birnbaum et al. (2020) that analyzed the use of offensive language in Facebook messages from individuals with mood disorders. Our work fills this important gap by providing further empirical evidence of the use of offensive language by individuals with diagnosed depression or showing signs of depression.

3 Data

In our experiments, we use three publicly available English datasets with data collected from social media: one with offensive language annotation, and two datasets with posts from users with self-reported depression diagnosis.

Offensive Language We use the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019a) to train offensive language identification models. OLID contains a total of 14,100 manually annotated posts from Twitter and it was released as the official dataset of SemEval-2019 Task 6 (OffensEval) (Zampieri et al., 2019b). We chose OLID due to its general hierarchical annotation taxonomy with the following levels:

Level A: Offensive language identification: offensive (OFF) vs. non-offensive (NOT)

Level B: Categorization of offensive language: targeted insult or threats (TIN) vs. untargeted profanity (UNT).

Level C: Offensive language target identification: individual (IND) vs. group (GRP) vs. other (OTH).

This hierarchical taxonomy provides us with a flexibility as it represents multiple types of offensive content in a single annotation scheme (e.g. posts targeted at an individual are often *cyberbullying* and posts targeted at a group are often *hate speech*) making it a great fit for this kind of analysis. In our experiments, we consider level A (offensive vs. non-offensive) and level B (target vs. untargeted).

Mental Health We run all our experiments on the Reddit Self-reported Depression Diagnosis (RSDD) dataset (Yates et al., 2017) and on the Early Risk Prediction on the Internet (eRisk) 2018 dataset (Losada and Crestani, 2016), two publicly available datasets containing posts from Reddit. The RSDD dataset consists of users annotated as having depression by their mention of diagnosis and control users, which are users who do not suffer from depression (there is not any mention of diagnosis in their posts). To prevent users labeled with depression to be easily identified by specific keywords, the authors removed posts containing depression terms (e.g. *depression, depressive*) or belonging to mental health related subreddits. The authors made the training, validation, and test splits available and in our experiments we use the training split, which contains over 5 million posts from users with depression and over 30 million posts from users in the control group.

The eRisk 2018 dataset contains users labeled with depression by their mention of diagnosis and control users. In this paper, we use both train and test splits, consisting of a total of approximately 90,000 submissions from users annotated as having depression and 985,000 posts and comments from the users in the control group. As opposed to the RSDD dataset, the authors removed only the posts containing the exact mention of diagnosis.

4 Methods

Offensive Language Detection and Categorization We address **RQ1** and **RQ2** by studying the language of users from the two groups, self-reported depression diagnosis and control, in social media. We start by computing an offensive score, which measures the extent to which a post is offensive, and whether it is a targeted insult or an untargeted post (most often profanity). These two tasks correspond to OLID levels A and B respectively Zampieri et al. (2019a).

For the task of offensive language detection, we fine-tune a BERT model on the OLID dataset on level A. We train the model for 2 epochs, with a small learning rate of 0.00001 and Adam optimizer (Kingma and Ba, 2015). We use an 80:20 split of the training data to choose the best performing model in terms of F1 score. The model obtains 0.85 Precision, 0.74 Recall and 0.77 F1 score on the test data from the OLID dataset. These numbers are consistent with the baselines reported in (Zampieri et al., 2019a). The offensive score is computed as a probability taken from the softmax output of the BERT model.

For the task of offensive language categorization (targeted insult or untargeted profanity) we also choose a transformer-based approach, using another BERT model trained on OLID level B. We fine-tune BERT for 7 epochs with the same aforementioned train-validation split, with a learning rate of 0.00002 with Adam optimizer and a linear warm-up schedule with a 0.05 warm-up ratio, as proposed by Rosenthal et al. (2020). To account for the class imbalance, we use cross-entropy loss with balanced class weights. The effectiveness of the model is also evaluated on the OLID test data, using the same metrics and achieving 0.78 Precision, 0.84 Recall and 0.80 F1 score.

Signs of Depression Detection Furthermore, we are interested in distinguishing the posts that show signs of depression from all the posts of individuals from the depression group. This way, we filter out the noise added by the texts which do not contain any cues of depression. We are using the Semantic Polarity Score heuristic (H_s heuristic) proposed by Ríssola et al. (2020) to detect posts showing signs of depression written by individuals with a self-reported depression diagnosis.

H_s uses a mix of sentiment polarity, depression score, and emotion detection. The authors use TextBlob[1] to obtain the polarity score of each post, ranging between -1 and 1. The terms from EmoLex (Mohammad and Turney, 2013) are used in order to detect the emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) contained in the texts. The depression score of each post is computed using the NRC Affect Intensity Lexicon (Mohammad, 2017), ranging from 0 to 1. In order to distinguish the posts showing signs of depression from other posts of users with self-reported depression diagnosis, we follow the criteria from Ríssola et al. (2020). Posts are labeled as showing signs of depression if the texts have a negative polarity, if sadness or disgust emotions are present, and if they have a depression score higher than 0.1.

5 Results and Discussion

Using the H_s heuristic, we demonstrate that there is a statistically significant difference (Welch t-test, p-value <0.001) in terms of offensive language use between individuals with self-reported depression diagnosis that manifest signs of depression in their posts and users who do not show any signs of depression. Posts containing signs of depression have a higher offensive score than posts from users diagnosed with depression without any signs, in both eRisk 2018 and RSDD datasets, as shown in Figure 1.

For labeling the offensive posts, we use the same 0.50 threshold as used during training. We show in Table 1 that more posts from users diagnosed with depression are labeled as offensive than from control. Using the H_s heuristic, we filter the posts containing signs of depression and find that there is a higher percentage of posts with signs of depression labeled as offensive. These findings are consistent for both eRisk 2018 and RSDD datasets. The higher degree to which depressed individuals use offensive language in comparison to individuals in the control group can be explained via the

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

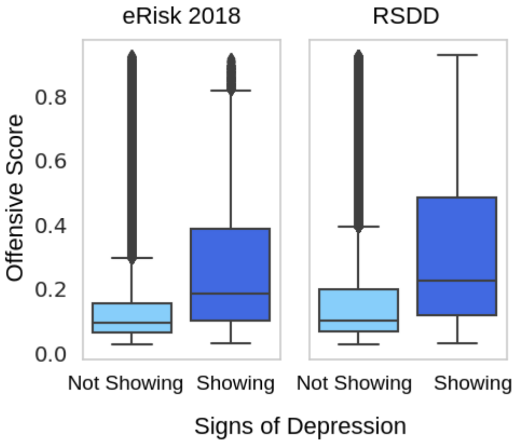

Figure 1: Distribution of the offensive language score for posts written by users with self-reported depression diagnosis and showing or not showing signs of depression measured with the H_s heuristic.

	Self-reported		Signs of depression	
Dataset	**Depression**	**Control**	**Showing**	**Not showing**
eRisk 2018	8.24%	5.91%	18.50%	7.40%
RSDD	11.31%	8.91%	24.33%	10.10%

Table 1: Percentage of posts labeled as offensive from total posts of self-reported individuals and of individuals showing/not-showing signs of depression measured with the H_s heuristic.

emotion regulation framework (Gross, 1999). The use of offensive language could be an emotion regulation strategy through which depressed individuals relieve some of their distress. Similarly, pain and distress studies indicate that the use of offensive language when experiencing pain significantly diminishes the level of pain experienced (Stephens and Robertson, 2020), suggesting that the use of offensive language can relieve distress.

Although there are more posts with signs of depression labeled as offensive, the majority of them are untargeted (containing swears, profanity) and only 10.71% and 10.72%, respectively, are targeted insults (Table 2).

	Self-reported		Signs of depression	
Dataset	**Depression**	**Control**	**Showing**	**Not showing**
eRisk 2018	24.12%	21.72%	11.48%	26.68%
RSDD	16.63%	23.94%	8.29%	18.48%

Table 2: Percentage of posts labeled as targeted insult from the offensive posts of self-reported individuals and of individuals showing/not showing signs of depression measured with the H_s heuristic.

The fact that depressed individuals tend to use more self-deprecating content and less deprecation of others, as evidenced in our analysis, is a result that is in line with the broad spectrum of cognitive studies, which indicates that negative evaluation of the self is a main interpretation bias in depressed individuals (Everaert et al., 2017). Depressed individuals tend to view themselves as less valuable than others. By self-deprecating language, we use the definition from Speer (2019). This broader definition includes, but is not limited to, insults towards self, if they have a negative intention. Finally, studies show that there is also a self-focused attention tendency in depressed individuals (Brockmeyer et al., 2015), where just like in other conditions (e.g. anxiety), individuals tend to be unable to detach from their own perspective focusing primarily on their side of the story, their pain, etc.

In order to further understand the differences in the use of offensive language, we analyze the words from posts written by individuals with depression. We compute the keyness score (Kilgarriff, 2009; Gabrielatos, 2018) of content words (removing stop words) from posts labeled as offensive written by users with self-reported diagnosis. The keyness is computed in order to show which words occur more often in the texts from depressed individuals showing signs of depression (target corpus) in comparison to the texts from users diagnosed with depression that do not show signs of depression (reference corpus). We calculate the frequencies of words from the two corpora and then the log-likelihood Ratio (G^2) (Dunning, 1993) for each word. In Figure 2 we present the top 20 words, ordered by G^2 from each corpus, in the two datasets.

We show that, while users without signs of depression refer more to sexual and profane terms, posts by users showing signs of depression include more negative words such as *bad, hate, sick, death*. This result corroborates the findings described in the literature on cognitive errors or biases in depression (Beck and Haigh, 2014). It is well known that depressed individuals tend to view life events more negatively than their non-depressed peers (Everaert et al., 2017). Furthermore, depressed individuals are more likely to recall negative life events than positive events and also more likely to pay closer attention to negative information (Beck and Haigh, 2014). Signs of this biased view of life are expected to be noticeable in language and there are studies that indicate that depressed individuals tend to have a more negative discourse than their non-depressed depressed peers (Rude et al., 2004). Keywords with a negative polarity, such as *bad, die* or *pain*, seem to be pervasive in the speech of depressed individuals as confirmed in our study. Finally, the reduced sexual drive is a well-known indication of depression (Manohar et al., 2017), therefore, it is to be expected that depressed individuals tend to use fewer words with sexual connotation as confirmed in our study.

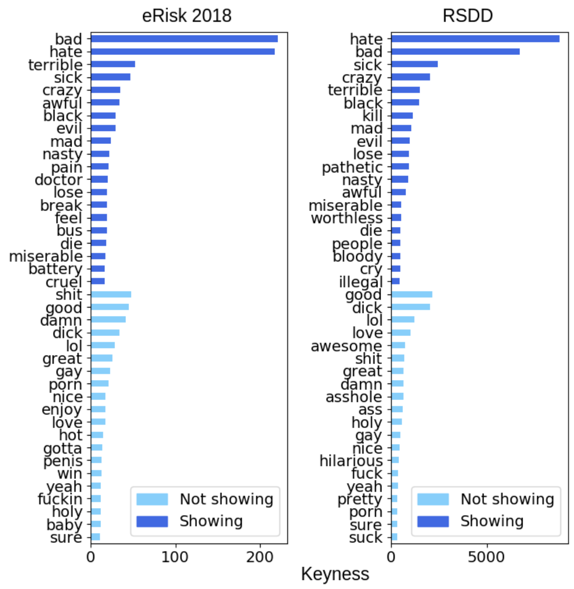

Figure 2: Keyness for words from posts showing/not showing signs of depression.

6 Conclusion and Future Work

This paper is the first to apply offensive language identification techniques to posts by individuals with a mental health condition with the purpose of interpreting the use of profanity and offensive language by this group. We showed how the offensive language use differs substantially between individuals with depression (in samples with self-reported diagnosis or showing signs of depression) answering our **RQ1**. Our findings indicate that users with self-reported depression diagnosis are more likely to use offensive language in their posts compared to the control group. From the posts of individuals with depression, the ones showing signs of depression contain more offensive language than the ones not showing any signs.

In terms of the nature of offensive content, our results indicate that posts from individuals with signs of depression are less likely to contain targeted offensive language. Furthermore, while analyzing the texts of users with depression, we observed a larger frequency of words with negative polarity (e.g. *bad, hate, sick, suffer*) in the posts of users showing signs of depression, where the discourse of users not showing any signs contains more sexual-related content, addressing our **RQ2**. These findings are consistent with the existing literature from psychology (Stephens and Robertson, 2020; Everaert et al., 2017; Beck and Haigh, 2014).

While it is clear that depressed users are more likely to write posts with negative polarity, the interplay between offensive language and polarity in the mental health datasets used in this paper has not yet been explored. A polarity score has been used in the heuristic by Ríssola et al. (2020) suggesting that using NLP models to investigate the interplay between polarity and depression is a promising future work direction. Other future work directions include the analysis of the targets of offensive posts using the OLID Level C annotation and a more detailed analysis on the function of profanity and vulgarity in these datasets (Holgate et al., 2018). Finally, we would like to carry out a similar analysis for other languages taking advantage of existing datasets and available cross-lingual embedding models.

Acknowledgments

We would like to thank Ioana Podină for providing us with helpful feedback and suggestions for the results analysis. We further thank the anonymous ACL reviewers for the insightful feedback provided. Finally, we thank the dataset creators for making the data used in this study available.

Liviu P. Dinu was partially supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number 108, within PNCDI III.

Ethics Statement

This paper uses publicly available datasets on offensive language and mental health to train computational models with the purpose of carrying out both quantitative and qualitative data analysis. We are primarily interested in quantifying and analyzing the use of offensive language in the texts included in the two mental health datasets and we do not attempt to predict mental health status or condition from these datasets. Potential biases in our model predictions and in our analysis may arise from the annotation and sampling techniques of these two datasets and are not intentional. Finally, we did not use any form of demographic information in our models or in our analysis.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of TRAC*.

Rienke Bannink, Suzanne Broeren, Petra M van de Looij-Jansen, Frouwkje G de Waart, and Hein Raat. 2014. Cyber and Traditional Bullying Victimization as a Risk Factor for Mental Health Problems and Suicidal Ideation in Adolescents. *PloS one*, 9(4).

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

Aaron T Beck and Emily AP Haigh. 2014. Advances in cognitive theory and therapy: The generic cognitive model. *Annual review of clinical psychology*, 10:1–24.

Michael L Birnbaum, Raquel Norel, Anna Van Meter, Asra F Ali, Elizabeth Arenare, Elif Eyigoz, Carla Agurto, Nicole Germano, John M Kane, and Guillermo A Cecchi. 2020. Identifying signals associated with psychiatric illness utilizing language and images posted to facebook. *npj Schizophrenia*, 6(1):1–10.

Timo Brockmeyer, Johannes Zimmermann, Dominika Kulessa, Martin Hautzinger, Hinrich Bents, Hans-Christoph Friederich, Wolfgang Herzog, and Matthias Backenstrass. 2015. Me, myself, and i: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. *Frontiers in psychology*, 6:1564.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of LREC*.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of CLPsych*.

Jonathan Culpeper. 2011. *Impoliteness: Using Language to Cause Offence*, volume 28. Cambridge University Press.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Dyberbullying Detection with User Context. In *Proceedings of ECIR*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of ICWSM*.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of CHI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Ted E Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Jonas Everaert, Ioana R Podina, and Ernst HW Koster. 2017. A comprehensive meta-analysis of interpretation biases in depression. *Clinical Psychology Review*, 58:33–48.

Costas Gabrielatos. 2018. Keyness analysis. *Corpus approaches to discourse: A critical review*, pages 225–258.

James J Gross. 1999. Emotion and emotion regulation. *Handbook of personality: Theory and research*, 2:525–552.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proceedings of the CL*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of TRAC*.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of SemEval*.

D. Losada and F. Crestani. 2016. A test collection for research on depression and language use. In *Proceedings of CLEF*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8).

Prasenjit Majumder, Thomas Mandl, et al. 2018. Filtering Aggression from the Multilingual Social Media Feed. In *Proceedings TRAC*.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting post severity in mental health forums. In *Proceedings of CLPsych*.

JS Manohar, TS Sathyanarayana Rao, S Chandran, et al. 2017. Sexual dysfunctions in depression. *Clin Depress*, 3(125):1–5.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of AAAI*.

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of CLPsych*.

Saif M Mohammad. 2017. Word affect intensities. In *Proceedings of LREC*.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of NAACL*.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of LREC*.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of CLPsych*.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of EMNLP*.

Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual Detection of Offensive Spans. In *Proceedings of NAACL*.

Esteban A Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2020. A dataset for research on depression in social media. In *Proceedings of UMAP*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *Findings of the ACL*.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of WebSci*.

Susan A Speer. 2019. Reconsidering self-deprecation as a communication practice. *British Journal of Social Psychology*, 58(4):806–828.

Richard Stephens and Olly Robertson. 2020. Swearing as a response to pain: Assessing hypoalgesic effects of novel “swear” words. *Frontiers in Psychology*, 11:723.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of ALW*.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of NAACL*.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of EMNLP*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of SemEval*.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL*.

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 CLPsych*.