KUCST@LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text

Manex Agirrezabal and **Janek Amann**
Centre for Language Technology
Department of Nordic Studies and Linguistics
University of Copenhagen
manex.aguirrezabal@hum.ku.dk, ja@developdiverse.com

Abstract

In this paper we present our approach for detecting signs of depression from social media text. Our model relies on word unigrams, part-of-speech tags, readabilitiy measures and the use of first, second or third person and the number of words. Our best model obtained a macro F1-score of 0.439 and ranked 25th, out of 31 teams. We further take advantage of the interpretability of the Logistic Regression model and we make an attempt to interpret the model coefficients with the hope that these will be useful for further research on the topic.

1 Introduction

Depression[1,2] is a mental illness that affects to the 5% of adults. The *World Health Organization* states that depression is the leading cause of disability worldwide. Human beings have varying mood, but depression is a condition that affects further than solely the mood. Depending on the degree of intensity of its symptoms, it may become a serious health condition. In spite of the magnitude and risk, there is effective ways of treating mild, moderate and severe depression.

In this paper we present our attempt for automatically classifying whether a social media post shows signs of moderate or severe depression. This is part of the *Shared Task on Detecting Signs of Depression from Social Media Text* at the LT-EDI-2022 workshop (Sampath et al., 2022). All our code is available in the following repository.[3]

The paper is structured as follows. First we introduce some related work on the topic. Then, we introduce the data that we employed. We continue with the used features and the actual models that we trained. After that we present the results and briefly discuss the model coefficients, and finally we conclude the paper with some possible future directions.

2 Related work

There have been several attempts to model the language of people with depression. In some works the focus is on detection of social media posts from users with different degrees of depression and in some other cases, the goal was to analyze the language style of people with depression.

There is a large number of works that have attempted to detect depression from Social Media text. Some works employ Twitter (Coppersmith et al., 2015; De Choudhury et al., 2021; Cavazos-Rehg et al., 2016; Mowery et al., 2016; Pirina and Çöltekin, 2018; Tadesse et al., 2019) and they work with different degrees of granularity with depression or depression-related symptoms.

Other researchers have employed other social media that contain longer essays, such as Reddit (Ireland et al., 2020; Iavarone and Monreale, 2021) for the detection of depression, by employing posts of users that were self-reported to have depression. Reddit posts have been further employed for, for instance, Bipolar disorder detection (Sekulic et al., 2018) or anxiety detection (Shen and Rudzicz, 2017).

With regards to features, many works make use of the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007). As in other Natural Language Processing related tasks, models based on contextual word embeddings have shown a good performance for depression detection, e.g. (Martínez-Castaño et al., 2020). As they report, the performance of the model is high but the interpretability of the model could be improved.

Besides, the use of personal pronouns have been analyzed by many researchers. For instance in Rude et al. (2004), they analyzed the language

[1] https://www.who.int/news-room/fact-sheets/detail/depression
[2] http://purl.bioontology.org/ontology/SNOMEDCT/35489007
[3] https://github.com/manexagirrezabal/depression_detection_EDI2022

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 245 - 250
May 27, 2022

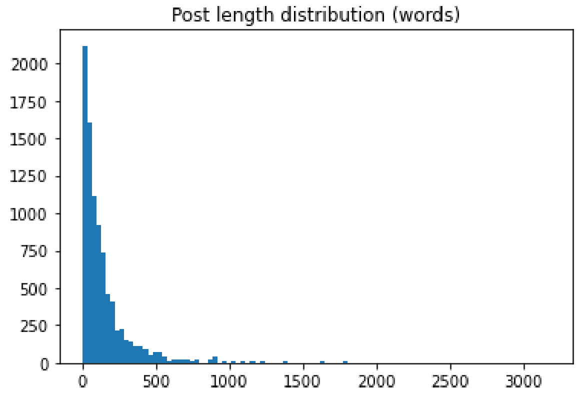

Figure 1: Histogram that shows the distribution of the length of social media posts in the current data set.

use of currently-depressed, formerly-depressed and never-depressed college students. Among other factors, they analyze the use of the first person pronoun "I" and they found that formerly-depressed and currently-depressed participants used the word "I" more often than the never-depressed participants. Furthermore, Tackman et al. (2019) claim that depressive symptomatology is manifested in a greater use of the first-person singular pronoun and find a small but reliable positive correlation between depression and I-talk.

3 Data

We make use of the data provided by the organizers of the shared task, built from Reddit posts (Sampath and Durairaj, 2022).[4] Some of these posts have no depression signs, others show moderate depression signs and finally, there are the ones that show severe depression signs. The dataset contains 8891 posts, from which the ones with no, moderate and severe depression signs are 1971, 6019 and 901, respectively. All posts are written in English. Figure 1 shows a histogram with the length of the posts.

4 Features and models

In this section we present the features that we employed. Many of the features have been widely used for text classification and authorship analysis.

Words. Bag of words as implemented by the CountVectorizer package from the `scikit-learn` library (Pedregosa et al., 2011). The expectation was that word usage might differ from depressed to non depressed users, and therefore, we expected that this feature would result beneficial.

Pos-tags. We also included part-of-speech tags among the employed features. But, we did not incorporate them as single counts, but we normalized them in a way that we got a probability distribution of pos-tags. We simply counted the frequency of each pos-tag in each post and then normalized them using the *softmax* function.

Readability and style. On top of that, we employ several readability and style related features as returned by a Python package called `readability`.[5] This package includes readability metrics, such as the Automated Readability Index (ARI), Coleman-Liau, Dale-Chall, and so on, [6] and some further stylistic features.

Person and number. In addition, following previous research on the topic, we also decided to include information about the usage of first person, second person or third person and also singular vs. plural word distribution. The difference of the usage ratio of the first person is visualized in Figure 2 for posts with different levels of depression signs. In order to calculate those, we used the `stanza` library (Qi et al., 2020).

Example: ***I am*** *lost because* ***I*** *do not like* ***them***.

In this example there are three words that express information in first person, there is one word that is in third person and there is no word expressing the second person. Therefore, the vector encoding this information would be $(0.75, 0.0, 0.25)$. With regards to number, it finds that there are three singular form words and one expressing a plural form, thus the vector that encodes number will be $(0.75, 0.25)$. The final vector representing person and number is a concatenation of the previous two vectors ($[0.75, 0.0, 0.25, 0.75, 0.25]$).

Models

As our goal was not to test how well different models would perform for the task, we decided to keep it simple and train Logistic Regression models. The main reason for doing this is the interpretability of the model, as the Logistic Regression is a relatively simple model.

[4] https://competitions.codalab.org/competitions/36410

[5] https://github.com/andreasvc/readability

[6] Please refer to the Github repository for a full list of outcomes.

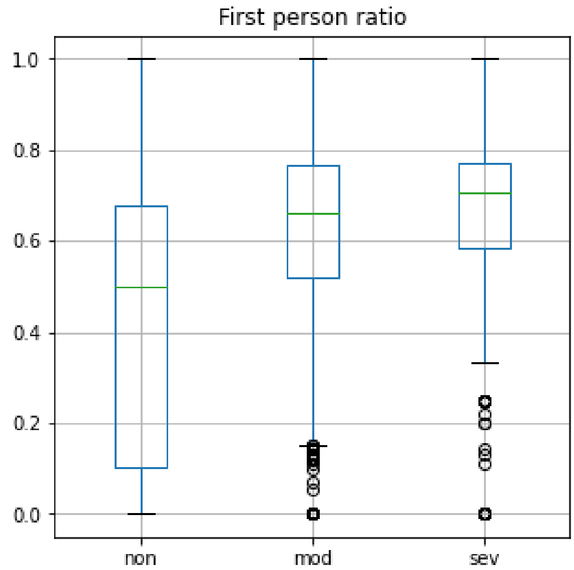

Figure 2: Usage ratio of the first person in posts with no depression signs, moderate signs and severe signs.

We trained two different Logistic Regression models[7,8] with the following feature configuration:

- Model 1: Words, POS-tags, Readability and style
- Model 2: Words, POS-tags, Readability and style, Person and number

5 Results and Discussion

Our best model, the second one, obtained a macro F1-score of 0.4429 on the test data. The first model performed marginally worse with a macro F1-score of 0.439. When performing our own experiments based on the training data, using a balanced train/test split, we had observed a rather higher performance, from which we could say that our model does not generalize well enough.

From the results, and by comparing to the rest of participants, we can say that our model has several aspects to be improved. In the team wise classification our model ranked 25^{th}, out of 31 teams.

As the logistic regression model features are interpretable, we decided to analyze them more thoroughly, with the hope that this analysis is helpful for further research. For this analysis, we used the second model that makes use of the all the features and they were obtained after training the model with all available training data. Figures 3, 4, 5 and 6 show the same sorted ranking of the features. In each figure we mark the position of the top 5 features, for each output class and for each feature template.[9]

Figure 4 shows that punctuation marks and nouns are can be good predictors. In figure 5 we can observe that the *Flesch Reading Ease* metric seems to be a good predictor together with the type token ratio. From figure 6 we can observe that the first person ratio and the plural ratio seems to have a rather high effect in at least two classes of posts, meaning that they could be good predictors. Finally, figure 3 shows the importance of the top 5 words. These last features seem to have more importance than other features. This is because the vectorizer for words[10] was used in the default configuration and no normalization was done afterwards (all other features had values between 0 and 1). This means that at the current stage we cannot compare the importance of specific features across feature templates based on the coefficients of the model.

All the observations regarding feature importance should be taken with a grain of salt. A better approach would be to use a bootstrapping approach, training several models from subsets of the training corpus and analyzing the weight importance among several of those models.

6 Conclusion and Future Work

In this paper we presented our attempt to classify whether a social media post from Reddit shows signs of depression. We employed simple features and a linear model and we made an attempt to interpret the learned coefficients. As mentioned above, the model has several aspects that could be improved given its performance. Below we outline some possibilities for further research.

Following recent advances in Natural Language Processing, we think that including a pretrained word embedding model, such as BERT (Devlin et al., 2019) would have positively contributed to the performance. These features could be additional features to the ones that we currently use or we could even fine-tune a pretrained model for this specific task.

Another aspect we believe that could improve

[7] All parameters are set to the default values.

[8] `https://scikit-learn.org/0.24/modules/generated/sklearn.linear_model.LogisticRegression.html`

[9] Our feature templates are words, POS-tags, readability & style and person & number.

[10] We used `CountVectorizer` from Scikit-Learn.

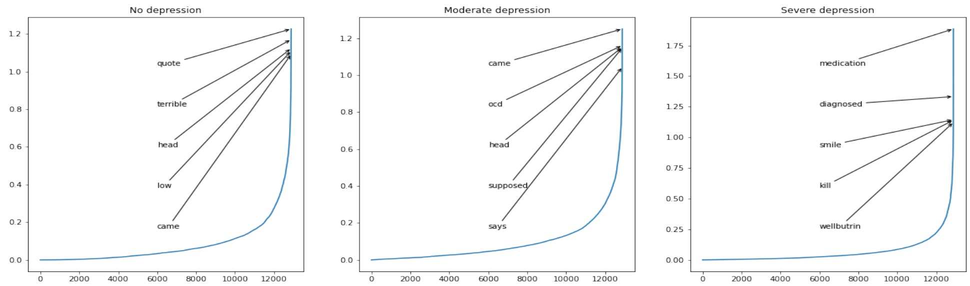

Figure 3: Sorted absolute values of Logistic Regression coefficients. We mark the rank of the top 5 features regarding words.

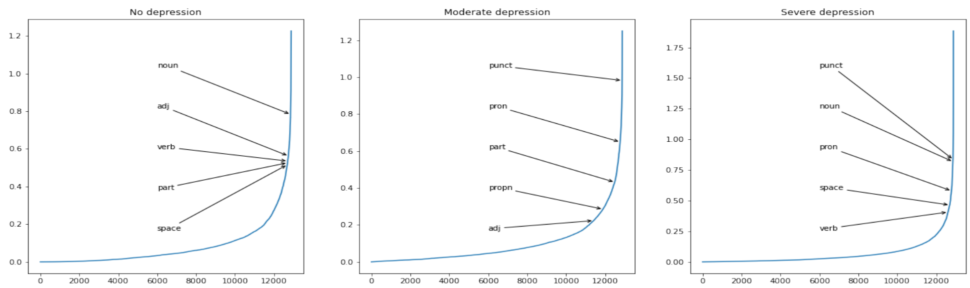

Figure 4: Sorted absolute values of Logistic Regression coefficients. We mark the rank of the top 5 features regarding POS tags.

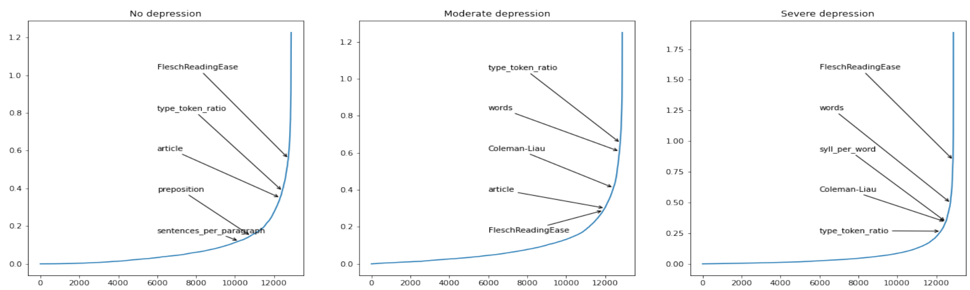

Figure 5: Sorted absolute values of Logistic Regression coefficients. We mark the rank of the top 5 features regarding readability & style.

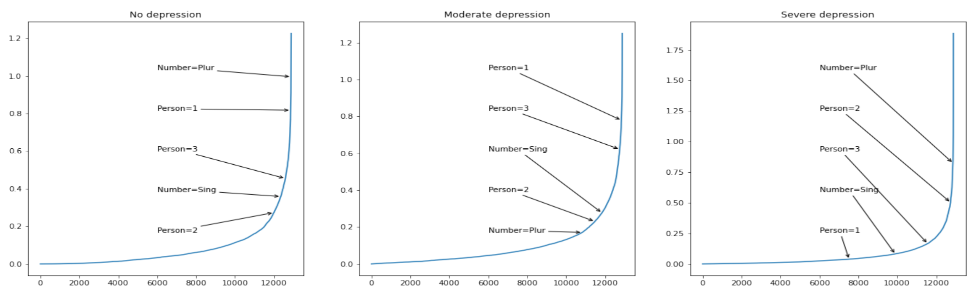

Figure 6: Sorted absolute values of Logistic Regression coefficients. We mark the rank of the top 5 features regarding person & number.

the model is to include further syntactic information. The use of dependency parsing is being currently tested, but besides, there is also an extension of the `readability` package[11], where syntactic information is obtained.

In addition to that, we expect that including the average sentiment of a post could be a relevant feature. Furthermore, recent advances in structured sentiment analysis[12,13] (Barnes et al., 2022) could potentially reveal mood changes.

References

Jeremy Barnes, Laura Ana Maria Oberländer, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval-2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle. Association for Computational Linguistics.

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*, 54:351–357.

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*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2021. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Benedetta Iavarone and Anna Monreale. 2021. From depression to suicidal discourse on reddit. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 437–445.

[11] https://gist.github.com/andreasvc/1fcdcbc2a21d31722facd98e5f02d19a/

[12] https://competitions.codalab.org/competitions/33556

[13] https://github.com/jerbarnes/semeval22_structured_sentiment

Molly Ireland, Jonathan Schler, Gilad Gecht, and Kate Niederhoffer. 2020. Profiling depression in neutral reddit posts: Prediction-insight tradeoffs and mental health technology applications.

Rodrigo Martínez-Castaño, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. 2020. Early risk detection of self-harm and depression severity using bert-based transformers : ilab at clef erisk 2020. *CEUR Workshop Proceedings*, 2696. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020. urn:nbn:de:0074-2696-0.

Danielle L. Mowery, Albert Park, Craig Bryan, and Mike Conway. 2016. Towards automatically classifying depressive symptoms from Twitter data for population health. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 182–191, Osaka, Japan. The COLING 2016 Organizing Committee.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Ron J. Weiss, J. Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

James W. Pennebaker, Roger John Booth, and Martha E. Francis. 2007. Linguistic inquiry and word count (liwc2007).

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on Reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

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

Kayalvizhi Sampath and Thenmozhi Durairaj. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. *CoRR*, abs/2202.03047.

Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C.

2022. Findings of the Shared Task on Detecting Signs of Depression from Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Ivan Sekulic, Matej Gjurković, and Jan Šnajder. 2018. Not just depressed: Bipolar disorder prediction on Reddit. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC. Association for Computational Linguistics.

Allison Mary Tackman, David A Sbarra, Angela L. Carey, M. Brent Donnellan, Andrea B. Horn, Nicholas S. Holtzman, T. Edwards, James W. Pennebaker, and Matthias R. Mehl. 2019. Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of Personality and Social Psychology*, 116:817–834.

Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.