Detecting and Explaining Depression in Social Media Text with Machine Learning

Rida Zainab
Stevens Institute of Technology
New Jersey, USA
rzainab@stevens.edu

Rajarathnam Chandramouli
Stevens Institute of Technology
New Jersey, USA
mouli@ieee.org

ABSTRACT

Depression affects one in 15 adults every year. It presents a serious challenge in personal and public health. In this paper, we identify depression from personal social media posts. Explainable artificial intelligence (AI) and natural language processing are combined to analyze and rank depression-related linguistic biomarkers. Experimental results show that the proposed approach is promising for mental health screening. Data from experiments on English and Urdu texts reveal the importance of semantic and cultural variations in language and people for depression detection.

KEYWORDS

depression, machine learning, explainable artificial intelligence, natural language processing

1 INTRODUCTION

Depression is a common but serious mood disorder. According to the World Health Organization, more than 300 million people globally suffer from depression. An individual with depression is 20 times more likely to die from suicide than someone without depression [14]. Although depression is among the most treatable of mental disorders, between 76% and 85% of people in low and middle-income countries receive no treatment for their disorder, primarily due to lack of resources and social stigma associated with mental disorders [20]. Estimated to account for 6.2% of the global burden of diseases in 2030 [1], depression is especially prevalent in young individuals in the 15-29 year age group wherein suicide is the second leading cause of death.

The prevalence of depression globally increased by 18% between 2005 and 2015 [2]. For patients suffering from depression, therapy is often the first level of treatment. Early recognition and treatment of depression has been shown to improve the negative impacts of the disorder [8]. However, due to steep increase in the number of depressed patients in recent years, in-person diagnosis and therapy is not accessible to all. Research has shown that online therapy, in addition to being more accessible and less time consuming, can be just as effective as standard therapy [9]. Hence, there is a need for online tools of depression detection and treatment.

Young individuals increasingly use social media platforms, such as Reddit, Twitter, Instagram and Facebook, to share their thoughts and opinions. These platforms are valuable source of databanks for curating natural language datasets to identify people's attitudes and behaviours. Therefore, analyzing content on such platforms could provide insights into how individuals discuss their depression.

The challenge of detection and treatment of depression in online settings involves identification of indicative features which do not necessarily manifest in in-person diagnosis, as well as, optimized use of online methods and forums of expression. A number of studies have shown that linguistic biomarkers such as increased frequency of personal pronouns and hedonic tone are symptoms of depression [11], [19], [16]. It has been observed that depressed patients used more words related to sadness [4]. The ratio of pronouns to nouns was also found to be an indicator of self-destructive behavior [18].

While previous studies have looked into mental illness detection and classification with Twitter corpora [6], [5], some studies have used the social media platform Reddit as it contains richer natural language information on depression since tweets, although available in large volumes, are limited in length [23], [7], [21]. Social media platforms such as Twitter or Facebook are also often associated with permanent online identities that potentially deter individuals from sharing their mental health problems. Reddit is a unique social website for news aggregation, content rating, and discussion in an anonymous setting as users can choose to create temporary accounts called "throwaway accounts" that are not associated with their main account in order to make posts, comments or ask questions regarding sensitive information [17]. Therefore, in this paper we use depression-related natural language English text data from Reddit.

After text classification, we employ an interpretable artificial intelligence (AI) model [10] to observe the differences between key features contributing to English vs. Urdu-based depression detection. Interpretable or explainable AI (XAI) methods are tools and techniques that attempt to describe reasoning behind decisions made by black box machine learning and AI engines in order to develop systems that are transparent about their reasoning and biases.

This paper is organized as follows. Section 2 describes the depression related dataset used in this study and the machine learning and XAI methods. In Section 3, we present the results and discussions from experiments. Concluding remarks are provided in Section 4.

2 METHODOLOGY

2.1 Depression Data

We use a custom dataset created from two subreddits: /r/depression and /r/CasualConversation. Social media posts were collected from the two subreddits using the Python Reddit API Wrapper (PRAW) to create the dataset. All posts from /r/depression were considered "depression posts" and posts from /r/CasualConversations were

considered "non-depression posts". We collected 20,000 posts each from /r/depression subreddit and /r/CasualConversation subreddit. The dataset was filtered to include only existing posts in the subreddit, deleted or removed posts were not included. The dataset was also filtered to remove posts containing images.

The dataset after pre-processing resulted in a total of 16,060 Reddit posts, consisting of 8030 depression posts and 8030 non-depression posts. Data from subreddit with higher number of posts was downsampled to create balanced classes. The data were randomly shuffled and split in 10 fold cross validation training and testing sets. The final dataframe consisted of one column of text of the post and another column of label for the corresponding post. Each post was labeled with 1 or 0 for "depressed" or "non-depressed", respectively, depending on the subreddit. In the subreddits, similar to [12], we only collected original posts and did not collect comments on the posts. For XAI analysis, we used a randomly selected subset of 8000 posts (4000 depression posts and 4000 non-depression posts) out of 16,060 posts and performed a classification task on it by using 70% data (5360 posts) for training and 30% (2640 posts) for testing. We also performed XAI analysis on complete dataset to study the effect of size of dataset on explainability.

2.2 Depression Classification

We created bag-of-words (BoW) and term-frequency times inverse document-frequency (TF-IDF) features [15] and used them separately for classification for both English and Urdu data. Both of these features are based on estimation of word or term in vector space. With BoW, documents are described by word occurrences represented by a matrix of frequency of occurrence of words in samples. TF-IDF, on the other hand, is a statistical measure used to evaluate how important a word is to a document in a corpus. It is a product of two values - term frequency, which is the count of a word in a sample, and inverse document frequency, which is logarithm of the quotient obtained by dividing the total number of samples by the number of samples containing the word.

In addition to these features, we also explored three lexical richness metrics namely type-token ratio (TTR), Brunet's Index (BI) and Honore's Statistics (HS) but we obtained poor classification results from these features.

We used Logistic Regression and Random Forest classifiers for classification. BoW and TF-IDF features were fed to machine learning classifiers and evaluation metrics were compared using 10 fold cross-validation framework. For model evaluation, we computed precision, recall and F1 score.

2.3 Explainable AI

For XAI analysis, Local Interpretable Model-Agnostic Explanations (LIME) was used to explain individual predictions of black box machine learning models [13]. For text classification, model agnostic interpretation methods offer desirable advantages such as the ability to work with any machine learning model, not limited to a certain form of explanation and the flexibility to use a different feature representation as per the model being explained [10].

LIME works by generating a new dataset consisting of permuted samples and the corresponding predictions of the black box model. In the context of text data, LIME works by creating new text from original text by randomly removing words from the original text. The dataset is then represented with binary features for each word. A feature is 1 if the corresponding word is included and 0 if it has been removed. Using this new dataset, LIME then trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest. The purpose is to approximate the prediction around the vicinity of a particular instance, while the original model may be very complex globally. This is done by treating the model as a black box and perturbing the sample for which explanation is needed and learning a sparse linear model around it.

3 RESULTS AND DISCUSSION

Table 1 and Figures 1 and 2 show the classification results for the entire dataset and XAI results for the reduced dataset. Results for Logistic Regression classifier were found to be slightly better than Random Forest classifier. As seen in Table 1, we obtained higher scores for TF-IDF vectors as compared to BoW in all three evaluation metrics.

Table 1: Depression detection results for English text

Logistic Regression

	BoW	TF-IDF
Precision	0.86	0.90
Recall	0.88	0.88
F1 score	0.87	0.89

Random Forest

	BoW	TF-IDF
Precision	0.85	0.86
Recall	0.82	0.83
F1 score	0.84	0.84

Figure 1: LIME - English Text - Logistic Regression

Results of LIME XAI were interesting. We saw frequent use of personal pronouns as reported in earlier works [19] and shown in Table 2. This table shows the frequency of occurrence of a personal pronoun in the top 3 keywords as computed by LIME for 2638 posts in the English language test dataset using logistic regression classifier and TF-IDF features. We also observed higher number of unique keywords (more than twice) given by LIME for non-depression posts as compared to depression posts.

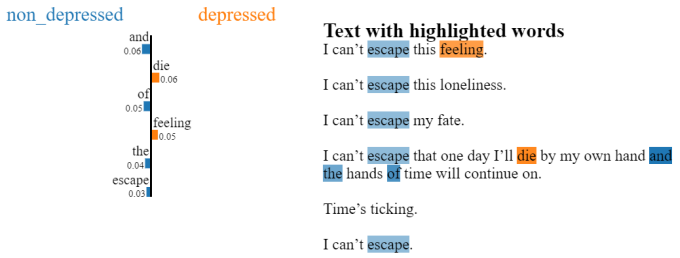

Figure 2: LIME - English Text - Random Forest

Table 2: Personal Pronouns in top 3 LIME Keywords

English Text		
Predicted class	True class	% Personal Pronoun
depression	depression	43%
depression	non-depression	52%
non-depression	depression	38%
non-depression	non-depression	27%

Upon increasing the size of the dataset from 8000 reduced sample size to 16,060 total sample size (2638 test samples and 5296 test samples for LIME analysis, respectively), we did not observe significant difference in the evaluation metrics. However, we observed an increase in the occurrence of more depression related words for depression posts in the list of keywords generated by LIME. For instance, we observed frequent occurrence of more curse words in LIME keywords after increasing the size of the dataset. We also observed that LIME interpretation was sensitive to pre-processing of the dataset, especially the choice of removal or no removal of stopwords from the dataset. For instance, LIME analysis without removing the stopwords from the dataset resulted in occurrence of personal pronouns and negative form of auxiliary verbs (such as don't) for depression posts in the list of keywords. However, since most methods of removal of stopwords (such as stopwords list in NLTK library) remove personal pronouns and auxiliary verbs, we did not see personal pronouns and negative form of auxiliary verbs for depression posts after removal of stopwords from our dataset in the pre-processing stage. The results described in this paper are without removing stopwords from the dataset in order to study the behaviour of occurrence of personal pronouns in depression dataset.

3.1 Depression Detection in Low-Resource Languages

As noted earlier, research into distinctive linguistic features of text of depression patients has been a major topic in the field of NLP. However, it is worth noting that these results are obtained from analysis of data from sources which are often in English language from English speaking population. Research and applications in NLP suffers from the problem of an overall unbalanced set of data resources when it comes to NLP tasks in languages other than high-resource languages such as English.

To explore this idea, we used the same English language reddit corpus and translated it into Urdu language using Google Translate. Urdu is the first language of around 70 million people, predominantly in Pakistan and India, and second language of more than 100 million people. Despite it's widespread use, Urdu is a low-resource language which is a major challenge for Urdu Natural Language Processing [3]. The quality of the Urdu language reddit translated text was manually checked by reading 50 random samples by one of the authors of this paper who is a native Urdu speaker. This was found to be contextually similar to the subject and content of the original English post despite grammatical errors.

The Table 3 and Figures 3 and 4 show the classification and XAI results for translated Urdu dataset. Results were found to be similar for English and Urdu text which was expected. However, the frequency of occurrence of personal pronouns in the top 3 LIME keywords using logistic regression classifier and TF-IDF features were found to be different for Urdu. As shown in Tables 3 and 4, we can observe a similar trend for Urdu and English personal pronouns occurrence but the percentage of personal pronouns occurrence for Urdu text are higher than English text. This is due to the fact that the Urdu word for "I" and "in" is the same. Moreover, there were errors in the translated Urdu text as social media posts are likely to have grammatical and spelling errors. We observed frequent use of the word "idk" (contraction of "I do not know") and "im" instead of "I am" or "I'm" which did not get translated into Urdu. Therefore, prior to employing text based diagnostic tools in NLP, the semantic and cultural variations of the language and population under consideration should be taken into account.

Table 3: Depression detection results for Urdu text

Logistic Regression		
	BoW	TF-IDF
Precision	0.85	0.89
Recall	0.87	0.87
F1 score	0.86	0.88

Random Forest		
	BoW	TF-IDF
Precision	0.83	0.84
Recall	0.82	0.83
F1 score	0.83	0.83

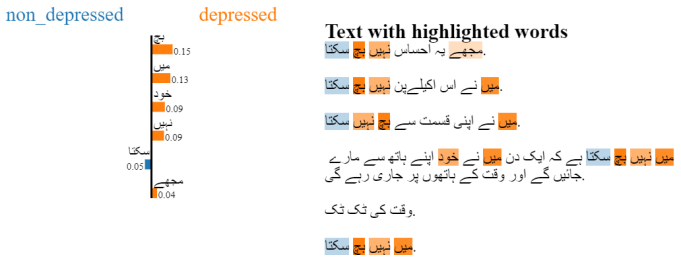

Figure 3: LIME - Urdu Text - Logistic Regression

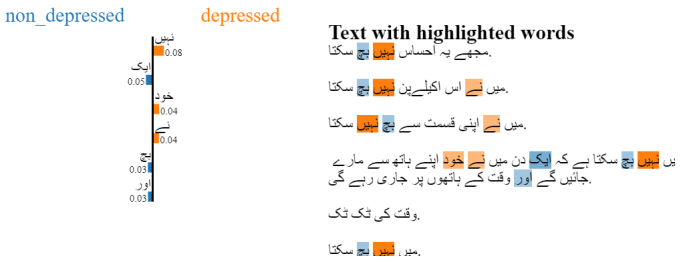

Figure 4: LIME - Urdu Text - Random Forest

Table 4: Personal Pronouns in top 3 LIME Keywords

Urdu Text		
Predicted class	True class	% Personal Pronoun
depression	depression	70%
depression	non-depression	83%
non-depression	depression	49%
non-depression	non-depression	36%

Using social media text for depression detection is a multi-layered problem. Most of social media activity online is in English language or Roman English script even amongst populations where the first language is not English. For instance, all Urdu language content on social media groups might not follow the same Nastaliq Urdu script used for this paper. A popular form of writing Urdu text on social media websites is in Roman alphabet called Roman Urdu. Roman Urdu, however, still lacks a standard script and spellings of words vary widely between users, although recently some work has been done in this area [22]. We also note that, a lot of Urdu language content in the original Urdu script online does not appear as text but rather as images of text. Hence, there is a need to address the various problems and challenges of low-resource languages in the field of NLP research.

4 CONCLUSION

There have been relatively few studies using Reddit for investigating linguistic aspects of mental health. The cultural context of expression of depression varies between geographies, communities, gender and age groups. Even for depression related text metrics which can be generalized across interlingual populations, the linguistic features of the target language should be taken into consideration as observed by our analysis of personal pronoun occurrence in two different languages. One limitation in this work is the inability of the features to take word ordering into account. It would be useful to perform similar analysis using attention-based mechanisms and observe the results.

Our machine learning and Explainable AI analysis of depression and non-depression reddit text data in English and Google translated English to Urdu version reveals: (a) an increased usage of personal pronouns when depressed, (b) variation in NLP results and social media text in inter-lingual text and population, (c) high ratio of of the number of unique keywords for non-depression to depression posts, (d) XAI sensitivity to pre-processing of the data.

REFERENCES

[1] [n.d.]. https://www.who.int/mental_health/management/depression/wfmh_paper_depression_wmhd_2012.pdf

[2] [n.d.]. https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf;sequence=1

[3] Farah Adeeba and Sarmad Hussain. 2011. Experiences in Building Urdu WordNet. In *Proceedings of the 9th Workshop on Asian Language Resources*. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 31–35.

[4] Sonnenschein AR, Hofmann SG, Ziegelmayer T, and Lutz W. 2018. Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cogn Behav Ther.* (2018).

[5] Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-Task Learning for Mental Health using Social Media Text. *CoRR* abs/1712.03538 (2017).

[6] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 13* (2013). https://doi.org/10.1145/2470654.2466447

[7] 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. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1485–1497.

[8] Cacheda F, Fernandez D, Novoa FJ, and Carneiro V. 2019. Early Detection of Depression: Social Network Analysis and Random Forest Techniques. *J Med Internet Res* (2019).

[9] Havigerová J. M., Haviger J., Kučera D., and Hoffmannová P. 2019. Text-Based Detection of the Risk of Depression. *Frontiers in psychology* (2019).

[10] Christoph Molnar. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

[11] Ramírez-Esparza N., Chung C. K., Kacewicz E., and Pennebaker J. W. 2008. The psychology of word use in depression forums in English and in Spanish: testing two text analytic approaches. (2008).

[12] Inna Pirina and Çağrı Çöltekin. 2018. Identifying Depression on Reddit: The Effect of Training Data. *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task* (2018). https://doi.org/10.18653/v1/w18-5903

[13] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (2016). https://doi.org/10.18653/v1/n16-3020

[14] Hannah Ritchie and Max Roser. 2020. Mental Health. (2020). https://ourworldindata.org/mental-health.

[15] Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation* 60, 5 (2004), 503–520. https://doi.org/10.1108/00220410410560582

[16] Rude S., Gortner E., and Pennebaker J. W. 2004. Language use of depressed and depression-vulnerable college students. (2004).

[17] Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting Anxiety through Reddit. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology - From Linguistic Signal to Clinical Reality* (2017). https://doi.org/10.18653/v1/w17-3107

[18] Litvinova T., Zagorovskaya O., Litvinova O., and Seredin P. 2016. Profiling a set of personality traits of a text's author: a corpus-based approach. (2016).

[19] Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. 2020. Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences. *IEEE Transactions on Knowledge and Data Engineering* 32, 3 (2020), 588–601. https://doi.org/10.1109/tkde.2018.2885515

[20] Philip S Wang, Sergio Aguilar-Gaxiola, Jordi Alonso, Matthias C Angermeyer, Guilherme Borges, Evelyn J Bromet, Ronny Bruffaerts, Giovanni De Girolamo, Ron De Graaf, Oye Gureje, and et al. 2007. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *The Lancet* 370, 9590 (2007), 841–850. https://doi.org/10.1016/s0140-6736(07)61414-7

[21] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2968–2978. https://doi.org/10.18653/v1/D17-1322

[22] Saif Ur Rahman Zareen Sharf. 2017. Lexical Normalization of Roman Urdu Text. *IJCSNS International Journal of Computer Science and Network Security* (2017).

[23] Ayah Zirikly, Philip Resnik, Özlem 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*. Association for Computational Linguistics, Minneapolis, Minnesota, 24–33. https://doi.org/10.18653/v1/W19-3003