MUCS@LT-EDI2023: Detecting Signs of Depression in Social Media Text

Sharal Coelho[a], **Asha Hegde**[b],
Kavya G[c], **Hosahalli Lakshmaiah Shashirekha**[d]
Department of Computer Science, Mangalore University, Mangalore, India
{[a]sharalmucs,[b]hegdekasha,[c]kavyamujk}@gmail.com,
[d]hlsrekha@mangaloreuniversity.ac.in

Abstract

Depression is a term used to characterize mental health disorders and it can worsen over time if left untreated, leading to more severe mental health problems and a lowered quality of life. Regardless of age, gender, or social background, anyone can be a victim of depression. Social media platforms are open to anyone including users suffering from depression, to write opinions on anything, post photos, videos etc., seek online help and so on. As depression can lead to significant changes in the individuals' posts on social media, analysing social media posts can provide insights into their mental health and reveal the signs of depression. However, manually analyzing the growing volume of social media text to detect signs of depression is time-consuming. To address the challenges of identifying signs of depression in social media content, in this paper, we - team MUCS, describe Transfer Learning (TL) and Machine Learning (ML) approaches, submitted to "Detecting Signs of Depression from Social Media Text" shared task, organised by LT-EDI@RANLP-2023. The objective of the shared task is to identify the signs of depression from social media posts in English and classify them into one of three categories: "not depressed", "moderately depressed", and "severely depressed". The TL model with fine-tuning Bidirectional Encoder Representations from Transformers (BERT) and ML models (Logistic Regression (LR) and Multinomial Naive Bayes (MNB)) trained with Term Frequency-Inverse Document Frequency (TF-IDF) of word n-grams in the range (1, 3) are submitted to the shared task. Among the two proposed models, the TL model performed better with a macro averaged F1-score of 0.361 for the Test set.

1 Introduction

Depression is a mental health condition resulting in feelings like sorrow, emptiness, loss of interest or distress and these feelings vary from individual to individual (Salas-Zárate et al., 2022). In severe conditions, depression may cause thoughts of suicide or death.

The user-friendly social media platforms allow users to share their posts or seek help from the online community (Hegde et al., 2022b). Depressed people might feel more ease in sharing their feelings, difficulties, and experiences, via posts on social media. Further, some people find it therapeutic to talk about their problems with others to get guidance, mental support, and sympathy from their online community. According to research, it might be possible to predict the signs of depression an online user is facing by reading the content of such users' posts on social media sites (Chiong et al., 2021). As depression can lead to significant changes in the individuals' posts on social media, analysing such posts can help in identifying the signs of depression in users. Once identified, any help or support can be extended to the users suffering from depression.

The increase in the number of social media users is increasing the user-generated text drastically (Kayalvizhi and Thenmozhi, 2022). Such user-generated text consists of hashtags, emojis, alphanumeric characters, slangs, short forms, etc. in addition to the actual content. Processing and analyzing this complex social media text using conventional text analysis techniques to get valuable insights into the data is challenging. This necessitates the need for automated tools/approaches to process social media text.

The automated approaches can identify depressive symptoms by systematically examining signs of depression in social media texts, providing a possibility for timely intervention and support for depressed individuals (Saqib et al., 2021). Though researchers have explored several techniques to detect signs of depression in social media text, it still remains a challenge because of the complexities of

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion,
pages 295–299, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_045

social media text.

To address the challenges of identifying signs of depression in social media text, in this paper, we - team MUCS, describe the models submitted to "Detecting Signs of Depression from Social Media Text" shared task, organised by DepSign-LT-EDI@RANLP-2023[1] (Sampath et al., 2023). The goal of this task is to detect the signs of depression from social media posts and classify them into one of three categories: "not depressed", "moderately depressed", and "severely depressed". The shared task is modeled as a multi-class text classification problem and two approaches: i) TL model with fine-tuning BERT and ii) ML classifiers (LR and MNB) trained on TF-IDF word n-grams, are proposed for the task.

The rest of the paper is as follows: related work is contained in Section 2 followed by the methodology in Section 3. Experiments with their results are explained in Section 4 and the paper concludes with future work in Section 5.

2 Related Work

Researchers have experimented many techniques to recognise signs of depression in social media content, and the description of some of the most useful studies are given below:

To address the early sign of depression in Reddit social media posts, Tadesse et al. (2020) developed ML models (Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Extreme Gradient Boosting) trained with TF-IDF and statistical features and Deep Learning (DL) models (combination of Long Short Term Memory and Convolutional Neural Network (LSTM+CNN) model) trained with Word2Vec embeddings. Among the proposed models, LSTM+CNN model obtained an accuracy of 93.8%. To identify the signs of depression in social media posts, Hegde et al. (2022a) used two learning models: i) TL model with fine-tuning BERT and ii) Ensemble (RF, Multilayer Perceptron, MNB, and Gradient Boosting (GB)) model with soft voting. As the dataset is imbalanced resampling technique (i.e. randomoversampling) is used to balance the dataset. Among the two models, the TL model performed better with a macro-average F1-score of 0.479.

Aswathy et al. (2019) utilized Word2Vec for generating word embeddings to train LSTM+CNN and SVM models to identify the signs of depression from the tweets. Their models obtained the weighted average F1-scores of 0.97 and 0.85 for the LSTM+CNN and SVM models respectively. Mowery et al. (2016), developed ML (Decision Tree, Linear Perceptron, RF, LR, SVM, and NB) models for determining whether a tweet represents evidence of depression or not and experimented on "Depressive Symptoms and Psychosocial Stressors Associated with Depression (SAD)" dataset which contains 9,300 tweets. To train their models, they used features including unigrams, emoticons, age, gender, linguistic inquiry word counts, etc. and obtained 0.52 average F1-score for SVM classifier. Janatdoust et al. (2022) proposed an ensemble of fine-tuned BERT models (A Lite BERT (ALBERT), DistilBERT, Robustly optimized BERT (RoBERTa), and BERT base model) with majority voting and experimented on social media comments in English (Kayalvizhi et al., 2022) and obtained a macro F1 score of 0.54.

To summarize, different learning approaches including ML, DL, and TL models are explored for detecting signs of depression in social media text. Though several techniques are experimented to detect the signs of depression in social media text in English, not all models have performed well. Further, the dynamic nature of user-generated content on social media makes the task more challenging. This emphasises the need for developing models to enhance the performance of identifying depressive symptoms from social media text.

3 Methodology

To detect the signs of depression in social media texts, the proposed methodology comprises of two learning models: i) TL model with fine-tuning BERT and ii) ML model trained with TF-IDF n-grams. Description of the two models are given below:

3.1 Pre-processing

User-generated social media texts consist of noise that includes non-ASCII characters, digits, hashtags, user mentions, URLs, and emojis. The given English text is converted to lowercase, contractions are expanded, and the URLs, digits, non-ASCII characters, punctuation, and extra spaces, are removed from the text as they do not contribute to the classification task. Pre-processing step remains the same for both the approaches.

[1] https://codalab.lisn.upsaclay.fr/competitions/11075

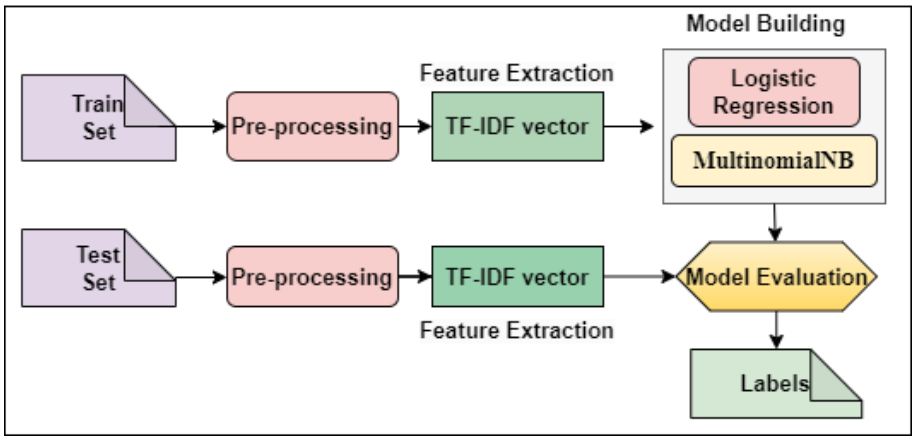

Figure 1: The proposed framework of Machine Learning classifiers

Dataset	Classes		
	Moderate	Not Depression	Severe
Train set	3,678	2,755	768
Development set	2,169	848	228

Table 1: Class-wise distribution of the dataset

Text	Label
My life is objectively very easy but my depression makes it all feel like a struggle	Moderate
I'm so tired and I hate everything.	Severe
i like being alone but i hate being alone anyone else	Not Depression

Table 2: Sample texts of 'Signs of Depression' from the dataset

Classifier	Development set	Test set
LR	0.457	0.346
MNB	0.313	0.236
TL model with BERT	**0.557**	**0.361**

Table 3: Performances of the proposed models in terms of macro-averaged F1-score

3.2 Model Construction

Methodology of the proposed TL model and ML models, to detect signs of depression in social media texts are given below:

TL Model - consists of training a model for source task and then applying the knowledge acquired from that task to the target task (Hegde and Shashirekha, 2022). It enables the model to start learning from a partially trained state, saving time and resources. BERT[2] is a Language Model (LM), pre-trained on 800 million English words from the Huggingface Book Corpus and 2,500 million English words from the Wikipedia corpus (Devlin et al., 2018). Using the concept of masked language model, it learns to predict the next words in the sentence. Further, it captures the context of words within a given sentence considering the nearby words on both the left and right sides and generates contextualized representations for words.

For the proposed TL model, the bert-base-uncased[3] - a BERT variant is used to represent text. From the huggingface library[4], BERT LM is loaded and fine-tuned with the Train set. A transformer based classifier - ClassificationModel is used to make the predictions.

Machine Learning models - consists of Feature Extraction and Classifier Construction. The framework of the ML model is shown in Figure 1. The significance of a word in a document relative to its frequency across all the documents in a corpus is captured by TF-IDF (Hegde et al., 2021). The proposed work utilizes TF-IDF of word n-grams in the range (1, 3), obtained using TfidfVectorizer[5]. 51,505 word n-grams are obtained from the Train set to train the classifiers.

Two classifiers: i) LR and ii) MNB are employed to predict the class labels for the input text. The regularization techniques in LR classifier helps to control the complexity of the model and discourage it from fitting noise in the data, making them effective tools for preventing overfitting in high-dimensional environments. The MNB classifier is

[2]https://huggingface.co/docs/transformers/model_doc/bert
[3]https://huggingface.co/bert-base-uncased
[4]https://huggingface.co/docs/hub/models-libraries
[5]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Text	Actual label	Predicted label	Remarks
It's the closest thing to dying. This life sucks.	Severe	Moderate	The words "thing", "sucks" and "life" are frequently employed with samples of "Moderate" class in the Train set. Hence, the model has classified this sample as "Moderate".
I miss when I was happy and life wasn't pointless.	Moderate	Not Depression	None of the words represent the signs of depression. Hence, this sample is classified as "Not Depression".
I wonder how much longer I can continue like this	Moderate	Not Depression	The words "wonder", "like", and "much" indicate the absence of depression. Hence, the model has classified as "Not Depression".

Table 4: Few misclassified samples from the Test set obtained by TL model

the probabilistic model (Harjule et al., 2020) which computes the prior probabilities of given classes and the dependent probabilities of words given the class. The class with the highest probability is selected as the predicted class for the given input text.

4 Experiments and Results

The proposed models aim to detect the signs of depression from the social media posts and classify them into one of levels of the signs of depression: "not depressed", "moderately depressed", and "severely depressed".

The statistics of the dataset provided by the shared task organizers (S et al., 2022) is shown in Table 1. The given dataset consists of social media posts in English and few samples from the dataset are shown in Table 2.

Predictions obtained from the proposed TL model and ML models are evaluated by the shared task organizers based on macro-averaged F1-score. The performances of the proposed models on Development and Test sets are shown in Table 3. Among the two proposed approaches, the TL model obtained a macro-averaged F1-score of 0.361 for the Test set.

The performances of the proposed models are influenced by issues like: the imbalanced dataset, an incorrect spelling of words, and limited vocabulary in the dataset. Further, the given Train set consists of very less number of samples for 'severe' class compared to the other classes. As a result, the proposed model failed to understand the features and patterns associated with the 'severe' class during the training process. Few misclassified samples in the Test set along with the actual and predicted labels and remarks are shown in Table 4.

5 Conclusion and Future work

In this paper, we describe two models: TL model with fine-tuning BERT and ML models (LR and MNB), for detecting signs of depression in social media text and classify them into one of three categories: "not depressed", "moderately depressed", and "severely depressed". These models are submitted to the "Detecting Signs of Depression from Social Media Text" shared task at LT-EDI@RANLP2023. Among proposed models, the TL model outperformed the ML models with a macro-averaged F1-score of 0.361. Efficient techniques will be explored to handle the imbalanced dataset and improve the performance of the proposed models.

References

KS Aswathy, PC Rafeeque, and Reena Murali. 2019. Deep Learning Approach for the Detection of Depression in Twitter. In *Proceedings of the International Conference on Systems, Energy Environment (ICSEE)*.

Raymond Chiong, Gregorius Satia Budhi, Sandeep Dhakal, and Fabian Chiong. 2021. A Textual-based Featuring Approach for Depression Detection using Machine Learning Classifiers and Social Media Texts. volume 135, page 104499. Elsevier.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Priyanka Harjule, Astha Gurjar, Harshita Seth, and Priya Thakur. 2020. Text Classification on Twitter Data. In *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, pages 160–164.

Asha Hegde, Mudoor Devadas Anusha, and Hosahalli Lakshmaiah Shashirekha. 2021. Ensemble Based Machine Learning Models for Hate

Speech and Offensive Content Identification. In *Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org*.

Asha Hegde, Sharal Coelho, Ahmad Elyas Dashti, and Hosahalli Shashirekha. 2022a. MUCS@ Text-LT-EDI@ ACL 2022: Detecting Sign of Depression from Social Media Text using Supervised Learning Approach. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 312–316.

Asha Hegde, Sharal Coelho, and Hosahalli Shashirekha. 2022b. MUCS@ DravidianLangTech@ ACL2022: Ensemble of Logistic Regression Penalties to Identify Emotions in Tamil Text. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 145–150.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2022. Learning Models for Emotion Analysis and Threatening Language Detection in Urdu Tweets.

Morteza Janatdoust, Fatemeh Ehsani-Besheli, and Hossein Zeinali. 2022. KADO@ LT-EDI-ACL2022: BERT-based Ensembles for Detecting Signs of Depression from Social Media Text. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 265–269.

S Kayalvizhi, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, et al. 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*, pages 331–338.

S Kayalvizhi and D Thenmozhi. 2022. Data Set Creation and Empirical Analysis for Detecting Signs of Depression from Social Media Postings. *arXiv preprint arXiv:2202.03047*.

Danielle L Mowery, Y 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.

Kayalvizhi S, 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*, pages 331–338, Dublin, Ireland. Association for Computational Linguistics.

Rafael Salas-Zárate, Giner Alor-Hernández, María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde, Maritza Bustos-López, and José Luis Sánchez-Cervantes. 2022. Detecting Depression Signs on Social Media: A Systematic Literature Review. In *Healthcare*, volume 10, page 291. MDPI.

Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Jerin Mahibha C, Kogilavani Shanmugavadivel, and Pratik Anil Rahood. 2023. Overview of the second shared task on detecting signs of depression from social media text. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Kiran Saqib, Amber Fozia Khan, and Zahid Ahmad Butt. 2021. Machine Learning Methods for Predicting Postpartum Depression: Scoping Review. *JMIR mental health*, 8(11):e29838.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2020. Detection of Suicide Ideation in Social Media Forums Using Deep Learning. volume 13, page 7. Multidisciplinary Digital Publishing Institute.