TBD3: Thresholding-Based Dynamic Depression Detection from Social Media for Low-Resource Users

Hrishikesh Kulkarni[1], **Sean MacAvaney**[2], **Nazli Goharian**[3], **Ophir Frieder**[3]
[1] Georgetown University, [2] University of Glasgow, [3] IR Lab, Georgetown University
[1] Washington, DC, USA, [2] Glasgow, UK, [3] Washington, DC, USA
[1] hpk8@georgetown.edu, [2] first.last@glasgow.ac.uk, [3] first@ir.cs.georgetown.edu

Abstract

Social media are heavily used by many users to share their mental health concerns and diagnoses. This trend has turned social media into a large-scale resource for researchers focused on detecting mental health conditions. Social media usage varies considerably across individuals. Thus, classification of patterns, including detecting signs of depression, must account for such variation. We address the disparity in classification effectiveness for users with little activity (e.g., new users). Our evaluation, performed on a large-scale dataset, shows considerable detection discrepancy based on user posting frequency. For instance, the F1 detection score of users with an above-median versus below-median number of posts is greater than double (0.803 vs 0.365) using a conventional CNN-based model; similar results were observed on lexical and transformer-based classifiers. To complement this evaluation, we propose a dynamic thresholding technique that adjusts the classifier's sensitivity as a function of the number of posts a user has. This technique alone reduces the margin between users with many and few posts, on average, by 45% across all methods and increases overall performance, on average, by 33%. These findings emphasize the importance of evaluating and tuning natural language systems for potentially vulnerable populations.

Keywords: depression detection; social media; thresholding; mental health; early risk detection

1. Introduction

Depression affects between 3% (UN, 2017) and 5% (WHO, 2021) of the global population and can lead to a variety of negative outcomes including suicide, self-harm, dementia, and premature mortality. Hence, early depression detection is crucial for preventing exacerbation of the condition. Best performing mental health monitoring methods are highly data-driven, often bound by the pertinence and volume of the dataset (Harrigian et al., 2021). Reaching actionable accuracy is particularly difficult when a user has a short posting history, providing systems little data to make an assessment.

We investigate the identification of depression in users with short posting histories (which we call low-resource users). We provide an evaluation of the performance of low-resource users (relative to their high-resource counterparts) using an existing large-scale dataset: the Reddit Self-reported Depression Diagnosis dataset (Yates et al., 2017, RSDD). By stratifying the dataset by the number of posts a user has, we find that a variety of classifiers perform substantially worse on low-resource users. For instance, the F1 score of a CNN-based model is 0.803 for high-resource users, but only 0.365 for their low-resource counterparts.

Recognizing that confounding factors could affect the aforementioned evaluation, we produce a new resource that aims to better control for the textual content across lower- and higher-resource users. This resource simulates low-resource users, namely, new entrants or simply users with fewer posts, by sampling the first posts from high-resource users. This sampling strategy is based on the intuition that high-resource users necessarily start out as low-resource when they begin using a platform. We find that the performance gradually increases as more posts are available (e.g., BERT achieves F1 scores of 0.237, 0.385, 0.454, and 0.508 when users have 100, 200, 300, and 400 posts). We observe similar trends for other models and using two other sampling strategies (most recent posts, and random posts). When considered in tandem with the dataset stratification experiments, these results heavily suggest a bias against low-resource users.

Further, prior social-media based methods for depression detection require a vast volume of data per user for accurate predictions. Hence, by the time prediction is made, users may already be suffering from later-stage depression.

We analyze the performance of various methods for the task of identifying depressed social media users with limited platform activity. We infer that the low performance is not just a result of limited data but other factors are likewise responsible. To address these shortcomings and to further establish the utility of our new resource, we propose a novel *Thresholding-Based Dynamic Depression Detection* (TBD3) method for classifying depression in social media posts. Through experiments on both the stratified and our new simulated dataset, we find that TBD3 successfully addresses the issues of disparity in depression detection from social media posts.

Our contributions are as follows:

- We establish a performance discrepancy between low- and high-resource users in depression detection

- We provide a new resource for evaluating these discrepancies in a controlled fashion
- We propose a new method (TBD3) designed to help systems improve performance for low-resource users
- Through extensive evaluation over four models, we find that TBD3 consistently improves the depression detection performance on low-resource users

2. Related Work

With the advent of social media and the anonymity it provides, forums that help users with mental health problems appeared. Some of these forums have moderators to counsel those in acute distress as providing them urgent counseling can potentially avoid their self-harm. For identification and prioritizing of such users, triaging methods with high accuracy were proposed (Cohan et al., 2016; Cohan et al., 2018b; Coppersmith et al., 2018; Adrian et al., 2020). Social media posts can provide mental health cues. The importance of automated screening for depression detection was highlighted along with evaluation criteria for existing language resources (Losada and Gamallo, 2020). Attempts were also made to detect depression from clinical interview transcripts by approximating mental lexicons from available lexicons (Villatoro-Tello et al., 2021).

Depression is often reflected through articulation, language, and selection of words (Coppersmith et al., 2021). Depressed users can be identified from their use of language alone (Coppersmith et al., 2017; Kelly et al., 2020). Data from social media platform Reddit were also used for investigating linguistic aspects of mental health leading to depression (Zainab and Chandramouli, 2020). Neural frameworks outperform traditional methods in the task of identifying depressed users from their use of language (Yates et al., 2017). Other efforts echo these findings with reference to other applications (Kalchbrenner et al., 2014; Xiao and Cho, 2016). Attempts were also made to collect temporal cues from self-reported depression diagnosis. Researchers also worked to identify temporal cues based on the RSDD dataset to cope up with dynamism of this problem (MacAvaney et al., 2018).

Depression detection is a crucial part of generalized mental health (Chancellor and Choudhury, 2020). Interactions with a forum does not always help distressed users. Classifying whether it will help or not and matching a user to the right forum can have great impact (Soldaini et al., 2018). Multiple mental health conditions can be interrelated, and analyzing them together can provide better insights about the user's mental health (Cohan et al., 2018a).

Various efforts evaluate performance of detection methods using the RSDD dataset (Cong et al., 2018; Rao et al., 2020). Classifiers supporting incremental learning with ability to visually explain its rationale were proposed for the task of depression detection (Burdisso et al., 2019). While most research efforts considered a large number of posts, there were a few attempts of early depression detection (Bucur et al., 2021). For example (Cohan et al., 2018b) showed that users' last posts tend to be of lower severity than their first post in a window of 36 months. This evaluation was performed on dedicated forum posts of a Reachout dataset (ReachOut, 2016) where users register to seek help. Lower severity was potentially due to interactions with counselors.

Researchers also experimented considering first, last and, random selections of posts in RSDD dataset (Yates et al., 2017). It was observed that the best performance for CNN model is achieved when it considers 100 terms per post and up to 1750 posts for each user.

Others used thresholding to improve classification accuracy in cases of imbalanced data (Sun et al., 2009) or selecting a number of classes in multi-label text classification (Yang and Gopal, 2011; Azarbonyad et al., 2021). Distinguishing from existing methods, TBD3 dynamically sets, rather than relying on predetermined, thresholds to detect conditions suffered by low-resource users.

In summary, prior efforts did not focus on detecting depression in the case of low-resource and early entrant users on social media platforms. Their findings did however establish a need for developing early depression detection systems that cope with variable number of posts without compromising accuracy. That precise need is our focus.

3. Language Resources

3.1. Dataset

The Reddit Self-reported Depression Diagnosis (RSDD) dataset (Yates et al., 2017) is suitable for evaluating natural language systems on potentially vulnerable populations. Though RSDD is a comprehensive dataset, for focused testing of individuals with varying levels of interactions, there is a need to create instances and filters of RSDD.

We derived a set of new language resources from the existing RSDD dataset. It contains a dataset with filters: first, random, last posts of users where the number of posts can be 100, 200, 300, 400 and all. This facilitates evaluation for early depression detection for low-resource users.

The RSDD dataset consists of 107,274 control and 9210 diagnosed users, split equally in three sets: training, validation and testing. Along with establishing baselines for evaluation, we release the code[1] to extract the derived set of language resources from the RSDD dataset. Access to the RSDD dataset is available through an already established and widely used

[1]The resource code repository can be found at: https://github.com/Georgetown-IR-Lab/lrec2022-tbd3

Data Usage Agreement (DUA). Further, the released code facilitates others to create language resources for early depression detection efforts with reference to duration of the presence of the users on the platform.

3.2. Ethics and Privacy

Data related to mental health on social media are often sensitive. We minimized the risk associated with use of these data for experimentation. The RSDD dataset uses publicly available Reddit posts. All precautions related to ethics and privacy per (Yates et al., 2017) are taken into account while working with these data, including anonymizing user identifiers, storing the data on secure servers, and making no attempt to re-identify users.

4. Method: TBD3

TBD3 formulates a personalized dynamic threshold for every user based on interactions with the platform. In TBD3, personalized dynamic thresholds are mathematically derived using a regression line along with saturation points. A validation set is used to establish this mathematical model. Let th be the set of possible thresholds with $th[i]$ being the i^{th} possible threshold. Let $F1[i]$ denote the $F1$ score corresponding to the i^{th} threshold value. Then, we define the ideal threshold as:

$$th_{ideal} = th[arg\max_{i}(F1[i])] \quad (1)$$

Based on the values of the thresholds obtained in our experiments over validation data, we formulate the threshold value solely in terms of the number of posts. We use TBD3 with Logistic Regression (LR), Support Vector Machine (SVM), Convolutional Neural Network (CNN) and Bidirectional Encoder Representations from Transformers (BERT) to formulate respective mathematical models. Based on these mathematical models derived from regression, the dynamic thresholds are obtained for different methods. We also threshold low and high number of posts. The upper bound for threshold refers to the threshold determined for all posts from validation data. We set the lower bound threshold to the threshold obtained for 165 posts as per the mathematical models of respective methods. The bounds resulting in optimal performance were determined empirically as we found that the performance is sensitive to the cutoff threshold values.

5. Experimental Setup

To minimize selection bias when the data are stratified based on the number of user posts, we filtered the dataset in various ways. These filtered datasets were used to validate the initial findings. For every machine learning algorithm considered, we experimented with 100, 200, 300, 400 and all posts. The posts were selected in three ways: first, last and random. For efficiency, thresholds were obtained from 10% validation data and tested on the complete test data for all methods. We experiment on the early stage of user activities with 20% and 40% initial posts of each user and use dynamic thresholds from the regression based mathematical model. The models for the experiments were implemented in python 3.9.1.

5.1. LR & SVM

As input, lexical bag-of-words (BoW) features were used. Words with a minimum document frequency of 12 were considered. The BoW tokenizer was fitted on the training set. These traditional machine learning methods were also used to establish baselines for early depression detection both with and without TBD3.

5.2. CNN

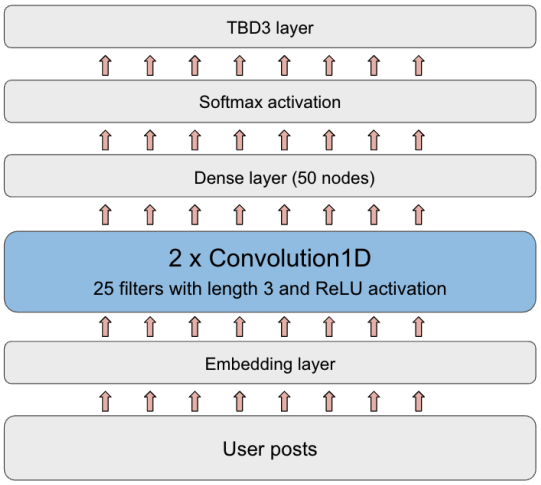

Figure 1: CNN architecture with TBD3 layer

As evident in figure 1, CNN had an embedding size of 50 per token, two Convolution1D layers with 25 filters, filter length of 3 and a ReLU activation function. It then had a dense layer of size 50 and an output layer with softmax activation function. Best performing combination of hyperparameters obtained after grid search was a learning rate $= 0.001$ and epochs $= 5$.

5.3. BERT

We likewise experimented with BERT. Considering the vast size of RSDD data per user and memory constraints, 310 x 128 tokens were considered per user. To compare CNN and BERT results, experiments were performed with exactly similar training and testing setups. BERT configurations (Devlin et al., 2019) tried were 'small_bert/bert_en_uncased_L-4_H-512_A-8' and 'small_bert/bert_en_uncased_L-2_H-128_A-2' with the latter giving the best results. Here, L denotes the number of transformer blocks, i.e., layers. H denotes the hidden size, and A denotes the number of self-attention heads. As the prior model is more sophisticated, it can accommodate only 175 x 128 tokens. Hence, there is a trade off. The later smaller model performs better than the prior model as it can accommodate significantly more data per user, i.e., 310 x 128 tokens as depicted in

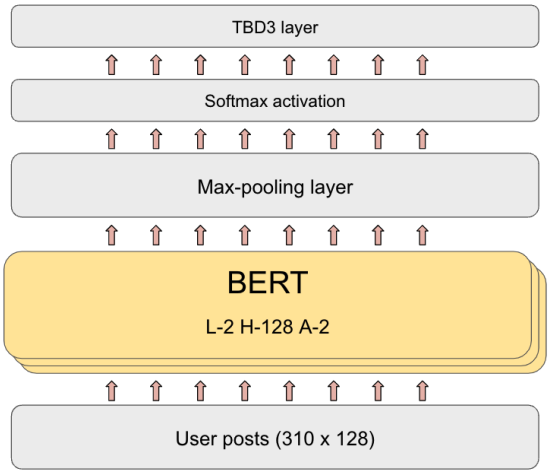

Figure 2: BERT architecture with TBD3 layer

figure 2. The best performing combination of hyperparameters obtained after grid search was a learning rate $= 2e-5$ and epochs $= 3$.

5.4. Mathematical Model

Our dynamic threshold has a linear relationship with the number of available user posts. In TBD3, we model the depression flag as True if the inequality holds. Here, n is the number of initial posts of the user available to the model for depression detection. Equations 2, 3, 4 and 5 represent linear models of threshold for SVM, LR, CNN and BERT, respectively where f is a flag indicating early depression. TBD3 can be deployed for any machine learning algorithm using two modules. Module 1 gives the score for the n posts of a user using the trained machine learning model. Module 2 gives the threshold for n posts based on the dynamic threshold linear model. If the score is higher than the dynamically determined threshold, we infer that the given low-resource user is depressed.

$$f = \text{SVM}(n\ posts) > 5.63e-07 * n + 0.0859 \quad (2)$$

$$f = \text{LR}(n\ posts) > 5.59e-05 * n + 0.5310 \quad (3)$$

$$f = \text{CNN}(n\ posts) > 1.59e-04 * n - 0.0039 \quad (4)$$

$$f = \text{BERT}(n\ posts) > 9.52e-04 * n - 0.0165 \quad (5)$$

6. Results

We now present our experimental findings. Initially we illustrate our results using a stratified dataset and then demonstrate the detection gains achieved using dynamic thresholding.

6.1. Discrepancy in Stratified Dataset

We conducted initial experimentation on a stratified set of users as per median number of posts. To assess the impact of the number of posts on depression detection accuracy, we divided the RSDD dataset into two segments. The segment with posts greater than or equal to the median number of posts (646) is referred to as 'ABOVE' while the one with less than the median is referred to as 'BELOW'. Evaluation of both segments was conducted separately, considering all posts of each user, using all four methods. Here, median value is used to logically separate low-resource users to carry-out effective experimentation.

ABOVE segment results are significantly greater than BELOW segment results. Table 1 depicts the comparative performance for the same. Here, thresholds are determined empirically from 10% validation data of the respective segments. Lower values of thresholds for SVM unlike other methods can be attributed to Platt scaling which is used for probability calculations. Table 1 clearly suggests that depression detection results are compromised when the number of posts by users are lower (typically below the median), and hence, there is a pressing need to develop effective and efficient depression detection approach specifically for users with a lower number of posts, including new entrants on the platform. Note that here, the default threshold used for baseline comparison is 0.5.

SVM				
Segment	Threshold	F1	Prec	Recall
BELOW	default	0.227	0.339	0.171
BELOW	0.083676	0.264	0.224	0.322
ABOVE	default	0.416	0.274	0.868
ABOVE	0.084480	0.586	0.668	0.523
LR				
Segment	Threshold	F1	Prec	Recall
BELOW	default	0.327	0.236	0.531
BELOW	0.646	0.374	0.352	0.400
ABOVE	default	0.608	0.671	0.555
ABOVE	0.648	0.607	0.734	0.518
CNN				
Segment	Threshold	F1	Prec	Recall
BELOW	default	0.365	0.790	0.237
BELOW	0.214	0.441	0.612	0.344
ABOVE	default	0.802	0.852	0.758
ABOVE	0.473	0.803	0.844	0.767
BERT				
Segment	Threshold	F1	Prec	Recall
BELOW	default	0.459	0.423	0.502
BELOW	0.599	0.484	0.516	0.456
ABOVE	default	0.731	0.702	0.763
ABOVE	0.625	0.744	0.788	0.704

Table 1: Impact of thresholding on performance for users stratified as per median number of posts

		SVM			LR			CNN			BERT		
First	Thr.	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall
100	def.	0.003	0.667	0.001	0.351	0.287	0.450	0.033	0.961	0.017	0.237	0.716	0.142
100	tun.	**0.310**	0.326	0.295	**0.376**	0.404	0.352	**0.430**	0.411	0.451	**0.404**	0.370	0.445
200	def.	0.045	0.651	0.023	0.391	0.334	0.470	0.133	0.922	0.072	0.385	0.662	0.271
200	tun.	**0.351**	0.350	0.353	**0.413**	0.350	0.402	**0.481**	0.432	0.541	**0.460**	0.478	0.444
300	def.	0.130	0.655	0.072	0.413	0.359	0.485	0.252	0.901	0.147	0.454	0.635	0.353
300	tun.	**0.367**	0.311	0.448	**0.428**	0.411	0.447	**0.526**	0.569	0.489	**0.492**	0.533	0.457
400	def.	0.235	0.600	0.146	0.423	0.371	0.491	0.369	0.899	0.232	0.508	0.636	0.422
400	tun.	**0.388**	0.332	0.467	**0.435**	0.428	0.442	**0.552**	0.546	0.559	**0.520**	0.510	0.530
All	def.	0.386	0.279	0.624	0.471	0.415	0.544	**0.692**	0.814	0.602	0.643	0.619	0.669
All	tun.	**0.474**	0.577	0.403	**0.481**	0.486	0.476	0.681	0.717	0.648	**0.651**	0.673	0.630

Table 2: Impact of thresholding on performance for SVM, LR, CNN and BERT on RSDD (def: default, tun: tuned)

6.2. Impact of Thresholding

The proposed method, TBD3, alleviates the aforementioned discrepancy. We first evaluate thresholding on the proposed language resource in combination with a lexical bag-of-words, and conventional CNN-based and transformer-based machine learning methods. Results are obtained on multiple levels of filtered data from the new language resource derived from the RSDD dataset.

Table 2 details performance for SVM and LR for inputs in increasing order of number of posts from 100 to 400 and also considering all posts. Here, tuned threshold is determined empirically from 10% validation data for similar setting. Changes in performance with variations in thresholds are depicted in the figure 3 for SVM and in the figure 4 for LR. Further, table 2 also shows performance evaluation in case of initial posts for CNN and BERT respectively. Note that the impact of thresholding is evident throughout. That is, in virtually all cases, with the rare exception of all posts for CNN and even there only a minor difference, accuracy improves using tuned thresholds (represented as "tun" in the table). Further, figure 5 and figure 6 depict performance variations with changing thresholds for CNN and BERT, respectively.

We obtain an F1 of 0.474 and 0.481 for SVM and LR, respectively when tested on all posts of all test users by applying thresholding. Similarly an F1 of 0.681 is obtained after using thresholding with CNN on all posts of all test users. Memory constraints restricted us to 310 x 128 tokens per user for BERT. We obtain an F1 of 0.651 for BERT by applying thresholding. To compare BERT and CNN, we obtained results for CNN also on similar data specifications (310 x 128 tokens per user) giving F1 = 0.623. Statistically significant improvement in performance i.e. on average 33% is observed by using thresholding across all machine learning algorithms when limited data for early depression detection was considered.

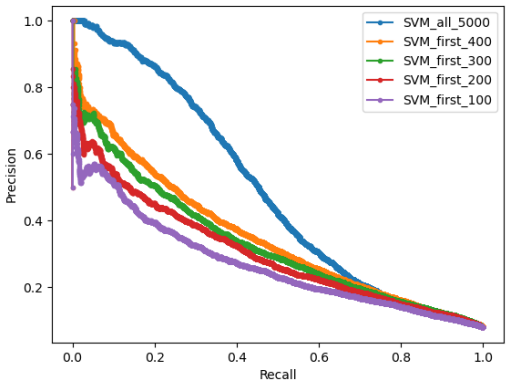

Figure 3: Precision vs Recall for SVM considering first n posts

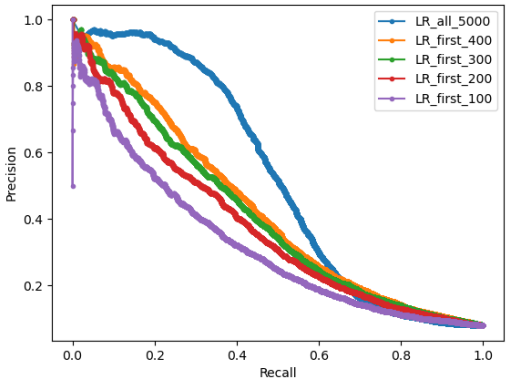

Figure 4: Precision vs Recall for LR considering first n posts

6.3. Efficacy of TBD3

With thresholding improving results, we evaluated TBD3 on 20% and 40% of the initial posts of each user. The TBD3 regression component of the devel-

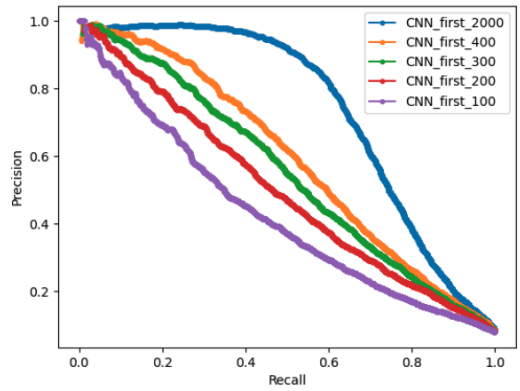

Figure 5: Precision vs Recall for CNN considering first n posts

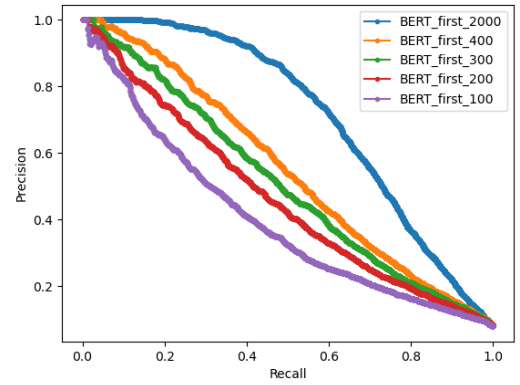

Figure 6: Precision vs Recall for BERT considering first n posts

oped mathematical model depends on the number of initial posts of the user that obviously varies from user to user. It is depicted in figure 7 (a) and (b) for SVM and LR, respectively. Similarly it is depicted in figure 7 (c) and (d) for CNN and BERT, respectively. As evident in table 3, we obtain an F1 of 0.613 and 0.561 for CNN and BERT for 20% of posts. Note that we observe improvements of more than 0.1. Similarly, we obtain an F1 of 0.632 for CNN and 0.621 for BERT for 40% of posts. We observe significant improvements across all methods, illustrating the efficacy of TBD3.

7. Analysis

From table 1, the F1 score improves significantly by shifting the threshold from default to a value determined from the validation set for all four methods. Additionally, we see that the optimal threshold value for each method differs. This can be attributed to different training setups and hyperparameters of the machine learning algorithms.

SVM				
Posts	Threshold	F1	Prec	Recall
20% Posts	default	0.269	0.861	0.159
20% Posts	dynamic	0.401	0.461	0.355
40% Posts	default	0.403	0.501	0.337
40% Posts	dynamic	0.434	0.432	0.435
LR				
Posts	Threshold	F1	Prec	Recall
20% Posts	default	0.366	0.306	0.457
20% Posts	dynamic	0.405	0.430	0.383
40% Posts	default	0.410	0.354	0.489
40% Posts	dynamic	0.444	0.453	0.435
CNN				
Posts	Threshold	F1	Prec	Recall
20% Posts	default	0.255	0.986	0.146
20% Posts	dynamic	0.613	0.748	0.520
40% Posts	default	0.468	0.982	0.307
40% Posts	dynamic	0.632	0.676	0.593
BERT				
Posts	Threshold	F1	Prec	Recall
20% Posts	default	0.453	0.899	0.303
20% Posts	dynamic	0.561	0.592	0.532
40% Posts	default	0.584	0.825	0.452
40% Posts	dynamic	0.621	0.655	0.590

Table 3: Impact of dynamic thresholding on performance for limited initial user posts (20% and 40%)

The statistical analysis is carried out to identify different trends across the results obtained. It is observed that results increase with an increase in the number of posts considered per user. For SVM, going from 100 to 400 posts, the F1 score increases from 0.003 to 0.235 with default threshold and from 0.310 to 0.388 with tuned threshold. The same trend is observed for LR, CNN and BERT. For CNN the F1 score increases from 0.033 to 0.692 as we go on increasing the posts from 100 to all posts. The saturation point for LR is attained at a smaller number of posts. The analysis clearly suggests the positive impact of increasing the number of posts on F1 score. Thus, more data lead to more insight in a user's mental state.

It is also observed that the performance improvement given an increase in the number of posts is quite steep, in the beginning. A lower number of posts restricts the performance and settles for lower F1 values across all the methods. Interestingly, in a low-resource scenario, a tuned threshold delivers substantial accuracy rewards. For a lower number of posts, like typically 100 and 200, improvement of 0.167 and 0.075 was observed, respectively for BERT. In the case of BERT, thresholding delivers an improvement of 0.01 for all posts. The results clearly establish the high efficacy of tuning thresholds in low-resource scenarios.

We also conducted similar experiments with random

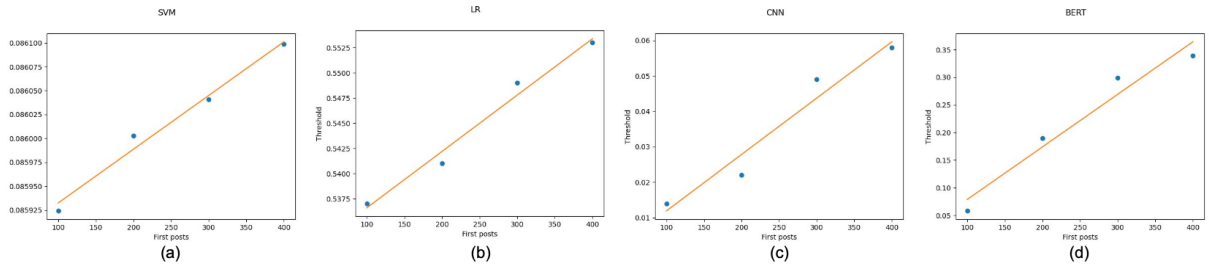

Figure 7: Dynamic threshold determination for (a) SVM, (b) LR, (c) CNN, (d) BERT on first n posts

and recent (i.e., last) post selections. We observed that the performance for random posts is greater than for recent posts, and recent posts provide better performance than initial (i.e., first) posts. This observation helps in limited posts selection. We observe that the transformer-based model performs better than the conventional CNN-based model. Both these models outperform lexical bag-of-words methods. In a similar learning environment and same size of data, the BERT model yields approximately a 4.5% improvement over the CNN model.

As evident in table 3, dynamic thresholds improve the performance for all methods. The steep increase in performance is observed for 20% posts across all methods. Further, for all methods, the F1 score is greater than 0.4. Significant improvements are also observed for 40% posts across all methods. For conventional CNN-based and transformer based models, the F1 score is greater than 0.56 when 20% posts of each user are considered. While in the same setup when 40% posts of each user are taken into account, the F1 score is greater than 0.62.

These results endorse the utility of TBD3 in case of low-resource users and demonstrate that the new proposed language resource can be used effectively for experimentation in such cases.

8. Conclusion

The availability of a limited number of posts impacts adversely on the performance of depression detection methods. Significant performance discrepancy is observed between low-resource and high-resource users in the task of depression detection. We provided a new language resource for the evaluation of the aforementioned discrepancies. We also proposed a novel method TBD3 which proves to be pivotal in early identification of depressed social media users with limited social media activity. TBD3 not only improves performance for low-resource users but also establishes the merit of the proposed language resource for controlled evaluation. The described method is easily adapted to other mental health issues and behavior analysis.

Appendix: Additional results

SVM	Threshold	F1	Prec	Recall
Last 100	default	0.005	0.778	0.002
Last 100	0.085898	0.335	0.267	0.452
Rnd 100	default	0.002	0.750	0.001
Rnd 100	0.085904	0.351	0.306	0.412
Last 200	default	0.044	0.719	0.025
Last 200	0.085993	0.361	0.355	0.366
Rnd 200	default	0.037	0.671	0.019
Rnd 200	0.085975	0.368	0.313	0.448
Last 300	default	0.135	0.686	0.074
Last 300	0.086049	0.378	0.356	0.403
Rnd 300	default	0.128	0.671	0.070
Rnd 300	0.086052	0.383	0.347	0.429
Last 400	default	0.247	0.681	0.151
Last 400	0.086072	0.391	0.333	0.473
Rnd 400	default	0.232	0.630	0.142
Rnd 400	0.086123	0.392	0.374	0.412

Table 4: Impact of thresholding on performance for SVM on RSDD

LR	Threshold	F1	Prec	Recall
Last 100	default	0.369	0.290	0.510
Last 100	0.537	0.398	0.385	0.412
Rnd 100	default	0.389	0.314	0.512
Rnd 100	0.538	0.424	0.306	0.443
Last 200	default	0.410	0.334	0.533
Last 200	0.540	0.438	0.425	0.440
Rnd 200	default	0.437	0.367	0.539
Rnd 200	0.543	0.469	0.480	0.459
Last 300	default	0.425	0.352	0.536
Last 300	0.555	0.444	0.408	0.487
Rnd 300	default	0.448	0.386	0.534
Rnd 300	0.551	0.469	0.448	0.496
Last 400	default	0.428	0.360	0.528
Last 400	0.554	0.445	0.425	0.467
Rnd 400	default	0.458	0.398	0.538
Rnd 400	0.554	0.477	0.470	0.484

Table 5: Impact of thresholding on performance for LR on RSDD

CNN	Threshold	F1	Prec	Recall
Last 100	default	0.035	0.945	0.018
Last 100	0.016	0.459	0.533	0.403
Rnd 100	default	0.032	1.000	0.016
Rnd 100	0.014	0.481	0.477	0.485
Last 200	default	0.143	0.920	0.077
Last 200	0.026	0.512	0.552	0.478
Rnd 200	default	0.125	0.939	0.067
Rnd 200	0.026	0.518	0.528	0.508
Last 300	default	0.272	0.929	0.077
Last 300	0.037	0.538	0.571	0.509
Rnd 300	default	0.254	0.909	0.148
Rnd 300	0.037	0.540	0.520	0.561
Last 400	default	0.361	0.921	0.224
Last 400	0.075	0.546	0.689	0.452
Rnd 400	default	0.380	0.905	0.240
Rnd 400	0.063	0.567	0.582	0.552

Table 6: Impact of thresholding on performance for CNN on RSDD

BERT	Threshold	F1	Prec	Recall
Last 100	default	0.281	0.767	0.172
Last 100	0.097	0.433	0.479	0.396
Rnd 100	default	0.269	0.769	0.163
Rnd 100	0.054	0.446	0.410	0.488
Last 200	default	0.402	0.720	0.279
Last 200	0.184	0.484	0.518	0.454
Rnd 200	default	0.408	0.686	0.290
Rnd 200	0.189	0.487	0.502	0.473
Last 300	default	0.489	0.717	0.371
Last 300	0.253	0.519	0.545	0.495
Rnd 300	default	0.493	0.666	0.391
Rnd 300	0.313	0.520	0.556	0.488
Last 400	default	0.525	0.715	0.414
Last 400	0.280	0.539	0.565	0.515
Rnd 400	default	0.533	0.654	0.449
Rnd 400	0.346	0.541	0.554	0.529

Table 7: Impact of thresholding on performance for BERT on RSDD

9. Bibliographical References

Adrian, M., Coifman, J., Pullmann, M. D., Blossom, J. B., Chandler, C., Coppersmith, G., Thompson, P., and Lyon, A. R. (2020). Implementation determinants and outcomes of a technology-enabled service targeting suicide risk in high schools: Mixed methods study. *JMIR Ment Health*, 7(7):e16338, Jul.

Azarbonyad, H., Dehghani, M., Marx, M., and Kamps, J. (2021). Learning to rank for multi-label text classification: Combining different sources of information. *Natural Language Engineering*, 27(1):89–111.

Bucur, A.-M., Cosma, A., and Dinu, L. P. (2021). Early risk detection of pathological gambling, self-harm and depression using bert.

Burdisso, S. G., Errecalde, M., and y Gómez, M. M. (2019). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.

Chancellor, S. and Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3, 12.

Cohan, A., Young, S., and Goharian, N. (2016). Triaging mental health forum posts. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 143–147, San Diego, CA, USA, June. Association for Computational Linguistics.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., and Goharian, N. (2018a). 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*, pages 1485–1497.

Cohan, A., Young, S., Yates, A., and Goharian, N. (2018b). Triaging content severity in online mental health forums. *Journal of the Association for Information Science and Technology (JASIST)*.

Cong, Q., Feng, Z., Li, F., Xiang, Y., Rao, G., and Tao, C. (2018). X-a-bilstm: a deep learning approach for depression detection in imbalanced data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1624–1627.

Coppersmith, G., Hilland, C., Frieder, O., and Leary, R. (2017). Scalable mental health analysis in the clinical whitespace via natural language processing. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 393–396.

Coppersmith, G., Leary, R., Crutchley, P., and Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10:1178222618792860. PMID: 30158822.

Coppersmith, G., Fine, A., Crutchley, P., and Carroll, J. (2021). Individual differences in the movement-mood relationship in digital life data. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 25–31, Online, June. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Harrigian, K., Aguirre, C., and Dredze, M. (2021). On the state of social media data for mental health research. In *Proceedings of the 7th Workshop on Computational Linguistics and Clinical Psychology: Improving Access*.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland, June. Association for Computational Linguistics.

Kelly, K., Fine, A., and Coppersmith, G. (2020). Social media data as a lens onto care-seeking behavior among women veterans of the US armed forces. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 184–192, Online, November. Association for Computational Linguistics.

Losada, D. E. and Gamallo, P. (2020). Evaluating and improving lexical resources for detecting signs of depression in text. *Language Resources and Evaluation*, 54(1):1–24.

MacAvaney, S., Desmet, B., Cohan, A., Soldaini, L., Yates, A., Zirikly, A., and Goharian, N. (2018). RSDD-time: Temporal annotation of self-reported mental health diagnoses. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 168–173, New Orleans, LA, June. Association for Computational Linguistics.

Rao, G., Zhang, Y., Zhang, L., Cong, Q., and Feng, Z. (2020). Mgl-cnn: A hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, 8:32395–32403.

ReachOut. (2016). Reachout platform. www.ReachOut.com. Accessed: 2021-30-12.

Soldaini, L., Walsh, T., Cohan, A., Han, J., and Goharian, N. (2018). Helping or hurting? predicting changes in users' risk of self-harm through online community interactions. In *CLPsych@NAACL-HTL*.

Sun, A., Lim, E.-P., and Liu, Y. (2009). On strategies for imbalanced text classification using svm: A comparative study. *Decision Support Systems*, 48(1):191–201. Information product markets.

UN. (2017). United nations depression. United Nations News. Accessed: 2021-30-12.

Villatoro-Tello, E., Ramírez-de-la Rosa, G., Gática-Pérez, D., Magimai.-Doss, M., and Jiménez-Salazar, H., (2021). *Approximating the Mental Lexicon from Clinical Interviews as a Support Tool for Depression Detection*, page 557–566. Association for Computing Machinery, New York, NY, USA.

WHO. (2021). World health organization depression. World Health Organization News. Accessed: 2021-30-12.

Xiao, Y. and Cho, K. (2016). Efficient character-level document classification by combining convolution and recurrent layers. *ArXiv*, abs/1602.00367.

Yang, Y. and Gopal, S. (2011). Multilabel classification with meta-level features in a learning-to-rank framework. *Machine Learning*, 88:47–68.

Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zainab, R. and Chandramouli, R. (2020). Detecting and explaining depression in social media text with machine learning. In *GOOD Workshop KDD*, volume 20, page 2020.