Predicting Moments of Mood Changes Overtime from Imbalanced Social Media Data

Falwah AlHamed[1,3], **Julia Ive**[2], and **Lucia Specia**[1]

[1]Department of Computing, Imperial College London, London, UK
[1]{f.alhamed20,l.specia}@imperial.ac.uk
[2]Queen Mary University of London, London, UK
[2]j.ive@qmul.ac.uk
[3]King Abdulaziz City for Science and Technology(KACST), Riyadh, Saudi Arabia

Abstract

Social media data have been used in research for many years to understand users' mental health. In this paper, using user-generated content we aim to achieve two goals: the first is detecting moments of mood change over time using timelines of users from Reddit. The second is predicting the degree of suicide risk as a user-level classification task. We used different approaches to address longitudinal modelling as well as the problem of a severely imbalanced dataset. For the first task, using BERT with undersampling techniques performed the best among models tested, including LSTM and random forests models. For the second task, extracting features related to suicide from posts' text contributed to the overall performance improvement. Specifically, a feature representing of a number of suicide-related words in a post improved accuracy by 17%.

1 Introduction

Social media platforms are widely used nowadays. The nature of these platforms allows people to be open and express themselves and share daily details about their activities and thoughts. As a result, social media data have been used in research for many years to understand users' mental health. A number of techniques have been proposed in the recent literature on monitoring mental health state over time. For example, a study by (Sawhney et al., 2020) was conducted to investigate suicidal risks from Twitter. The authors used a time-aware transformer model with a pre-collected data set for suicide ideation and applied their model on 34,306 tweets from 32,558 users. The main goal was to classify if the person is at risk based on their sequence of tweets. Another study was conducted for detecting mood change by (Pruksachatkun et al., 2019). They proposed a predictive model to determine if a post is associated with a moment of cognitive change.

In this paper, we explain our approach to the CLPsych (Tsakalidis et al., 2022a) shared task, which consists of two subtasks, as follows:

Subtask A: Subtask A tries to capture those moments when a user's mood deviates from their baseline mood based on a user's postings throughout a specific time period — this is a post-level sequential classification task. The full task description can be found in (Tsakalidis et al., 2022b).

Subtask B: A user-level classification task on predicting the degree of suicide risk. An individual/user is considered to belong to one of four categories: no, low, medium or severe risk based on their posts on Reddit "r/SuicideWatch". The full task description can be found in (Zirikly et al., 2019).

2 Dataset

Data used for this shared task was pulled from Reddit. This well-known social media platform contains communities known as "subreddits", each of which covers a different topic.

For Subtask A, subreddits relating to mental health were used in this task. A total of 186 users were included in this study, with 256 timelines and a total of 6205 posts. The average time span for each user is 2 months. Data annotation was carried out by four annotators with multiple training rounds and mediation. Timelines were manually checked to ensure that they contain content indicating mood. Each post was labelled with one of three labels: IS for Switches i.e (mood shifts from positive to negative, or vice versa), IE for Escalations – gradual mood progression from negative (positive) to very negative (very positive), and 0 for no change. Subtask A data can be found on (Losada and Crestani, 2016; Losada et al., 2020; Shing et al., 2018). The data for this task was severely imbalanced. The values distribution were 79% for 0, 15% for IE, and only 6% labelled as IS.

For subtask B, four clinical experts annotated the user based on data from the *SuicideWatch* subreddit to one relative suicide risk severity. SubTask B data

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 239 - 244
July 15, 2022 ©2022 Association for Computational Linguistics

can be found on (Shing et al., 2018).Each user was labelled with one of four labels: "None", "Low", "Moderate", "Severe" representing their suicidal risk level. The classes “Low” and “None” were merged together to address the class sparsity issue. The resulting class set is composed of the “Severe”, “Moderate” or “Low” classes for 127 users with the frequencies of 48%, 43% and 9% respectively.

All authors have signed a Data User Agreement (DUA) and Non-Disclosure Agreement (NDA) to have access to the dataset.

3 Methods

In this section, we will describe the methods we developed to address these two shared subtasks.

3.1 Subtask A

We looked at various strategies to address the problem of data imbalance and also to consider longitudinal modelling.

3.1.1 Pre-processing

Different preprocessing techniques were applied on the posts in sequential manner using regular expressions operations. This includes cleaning for special characters and words such as users’ mentions (special character ‘@’). Some characters were defined to be word boundaries characters which include comma, period, colon, question mark and semicolon. All these characters are replaced with a white space. Also, all URL hyperlinks were removed from posts with Regex.

3.1.2 Undersampling

We used undersampling to address the severe class imbalance. For this, we inspect sentiments in texts posts using TextBlob. [1] We found that most posts labelled with “0” have a positive sentiment with polarity greater than 0.2 (polarity ranges between -1 to 1), while “IS” and “IE” posts are connected with negative sentiment. This allowed us to remove 649 (out of 5143) samples labelled with “0" (polarity >= 0.2) and improve the dataset balance. We note that oversampling could be an alternative technique to avoid reducing sample size, which we leave for future work.

3.1.3 BERT

Models built by fine-tuning BERT (Devlin et al., 2019) or related pre-trained language models achieve state-of-the-art performance in a number of NLP tasks. This approach has been shown to give good results in multiple classification tasks, outperforming other algorithms (Acheampong et al., 2021; Al-Garadi et al., 2021). We used BERT with sequence length of 512 for post-level classification. In other words, *the predictions are performed per post without taking the preceding sequence of posts into account*. We experimented with the following different hyperparameters: `batch size:4,8,16,32; epochs: 8,16,32,64`. We reported the best parameters in Section 4.2.

3.1.4 LSTM

LSTMs are widely used for predicting sequential and temporal events, for example in (Chiu et al., 2021; Mirheidari and Christensen, 2019; Sawhney et al., 2020). We used LSTM for monitoring and predicting mood changes over time *taking into account the previous sequence of posts*. Since the baseline model for this task uses LSTM with BERT embeddings, we tried different embedding types, namely GloVe [2] and SpaCy Tok2Vec.[3] We tuned different hyperparameters to improve accuracy of the model. `batch size = [16,32,64,128] epochs=[16,32,40,64] learning rate=[0.01,0.02,0.05,0.1,0.2,0.5].` We reported the best parameters in Section 4.2.

3.2 Subtask B

The aim of this subtask is to classify users to the correspondent suicide risk level. It is clear that the “Low” class is the least represented, which we take into account in our models.

3.2.1 Extra Features

To improve models performance and to account for the class imbalance, we extracted extra features that could positively affect the models’ results. Since data size is small for this task (only 127 user), we used all data without undersampling.

Sentiment: Using TextBlob,[4] we extracted the sentiment of each post in user’s data, then we sum the sentiment and based on the total we assign to each user a value of “Positive” if the total is greater than zero or “Negative” if the total is less than zero.

Polarity: We extracted the polarity of each post as a value between -1 and 1 (where -1 is severe

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

[2] https://nlp.stanford.edu/projects/glove/

[3] https://spacy.io/api/tok2vec

[4] https://textblob.readthedocs.io/en/dev/

Table 1: List of suicidal words for Task B

Suicidal Words	
kill	die
knife	survive
dead	end my life
I'm gone	live anymore
I'm done	taking my life
killing	overdose
jump	suicide
wrist	hang
burn	self-harm
self harm	pesticide
death	take my life
call for help	

negative and 1 is extreme positive) using TextBlob, then we calculated the sum for all posts to get the polarity feature as a numeric value. Most users with severe risk level received a negative value, and most users with low risk levels received a positive value. The polarity was chosen as an indicator of the sentiment intensity.

Number of Suicidal Words: We inspected the posts of the three classes and found that the Severe class contains many words related to suicide attempts and ideas. Thus, we created a list of suicidal words by combining words from (Yang et al., 2022) and other words inferred from manual posts inspection. The word list is shown in Table 1. Then for each user, we calculated the number of words from the suicidal list that occurred in their posts. We added the total frequency of suicidal words as a feature. Related research has shown that combining lexical features besides machine learning models can improve the prediction results (AlHamed and AlGwaiz, 2020; Carvalho and Plastino, 2021).

3.2.2 Random Forests

Random forest is an ensemble machine learning model that relies on constructing multiple decision trees, then comparing the output of trees to predict the class. The class selected by random forests is the class that was selected by most of the trees via majority voting. Random forest was chosen as a non-neural algorithm as it has been shown to achieve higher accuracy in text classification tasks compared to other traditional machine learning algorithms such as KNNs (Biau and Scornet, 2016; Pranckevivius and Marcinkevicius, 2017). We used three random forest models in this task. The first with only word embeddings as features (RF1). The second with word embeddings and the additional extracted featured (RF2). The third with only the extracted features without word embeddings (RF3). We performed random grid search with the following hyperparameters: `no. of estimators = [200,300,400,500... 2000]; max features = [‘auto’, ‘sqrt’]; max depth = [10,20,30,...110]; min samples split = [2, 5, 10]; min samples leaf = [1, 2, 4]; bootstrap = [True, False].` Best performing parameters are reported in section 4.3.

4 Results and Evaluation

Results from the all models in both tasks on the blind test set are shown in Table 2. Baseline models (as reported by the shared task organisers) are Majority, TFIDF-LR, and BERT-Talklife-focal.

4.1 Evaluation metrics

As per (Tsakalidis et al., 2022b), the evaluation is carried out using two types of metrics. The first one is post-level metrics, which assesses the model's performance using precision, recall, and F1 score. The second type is coverage-level, these are the same metrics (precision, recall, and F1 score) but assessing the performance at the timeline level to assure that the model captures the sequence of mood changes overtime.

4.2 Subtask A Results

For this task we used three models, LSTM with SpaCy embeddings (LSTM-SpaCy), LSTM with GloVe embeddings (LSTM-GloVe), and BERT. All models are trained on data after undersampling. BERT performed the best in all the evaluation metrics, we think the reason behind that is BERT was fine-tuned on the dataset while LSTM models used pre-trained embeddings. Results for "IS" are the lowest as the class is underrepresented. For LSTM, the best hyperparameters are as follows: `batch size = 32, epochs=40, learning rate=0.05, optimiser = Adam`. For BERT, the best results were obtained for a model with batch size = 8 and number of epochs = 8.

4.3 Subtask B Results

For this task, we tried three types of random forests models.

The best results were obtained with the following settings: max depth=60, max features='sqrt', min samples leaf=2, min samples split=10, no. of estimators=600, random state=3. Surprisingly, RF3 where we used only the extra features as input (without using embeddings) outperforms the other RF models by 17% in accuracy. The reason behind this could be that the high similarity of words presented in all classes negatively affected prediction, and using only suicidal words and sentiments provided better context inference. This indicates that extracting additional meaningful features from text can enhance classification results.

5 Discussion

For subTask A, as shown in table 2, our BERT model outperformed the baseline BERT model - where LSTM over BERT embeddings is used - for macro-average results in both coverage based metrics and post-level metrics evaluation. Our proposed model with undersampling also scored the highest in precision and recall for the least presented class "IS". The reason behind this could be that the model was able to learn the features of this class after undersampling. On the other hand, the model performed less well in detecting "0" class. It could be an effect of undersampling, or that because other models were trained on the severely imbalanced dataset, they were biased toward predicting "0", and thus scoring higher precision and recall values.

When it comes to all participants in this year's CLPsych shared task, our BERT model ranked the third best performing model for post-level metrics evaluation. This emphasizes the feasibility and usefulness of the undersampling technique used.

For subTask B, compared to baseline models, RF3 was the only model able to predict the class "low". A possible explanation is that using the suicidal words count feature helped in identifying "low" suicidal risk users. On the other hand, results for "Moderate" and "Severe" classes were less compared to baseline models, this might be because we did not normalize the number of suicidal words to the number of posts per user and thus the model was inflated for users with more posts.

It is essential that the limitations of this study are considered in future studies. Firstly, the suicidal words list is collected from different sources and from analysis and manual inspection of the dataset. It could be expanded and validated to include more suicide related words. Another limitation is that we did not fine-tune any embedding model to our dataset (except for BERT). We used general pretrained embeddings such as GloVe. Also, we aimed to try oversampling techniques to address data imbalance but we could not achieve that due to time constraints.

6 Conclusions and Future Work

In this paper, we presented our system description for CLPsych shared task (Tsakalidis et al., 2022a). The task consists of two subtasks. Subtask A aims to detect moments of mood change for posts in a timeline. For this, first we undersample the dataset to address the severe imbalance in dataset by filtering out the posts with positive sentiment irrelevant to mood changes. BERT without explicit modelling of the post sequence outperforms other models. Subtask B aims to classify a user to correspondent suicide risk-level. For this task, we extracted additional features and performed random forests. The proposed model succeeded in detecting the least represented class. In future, we aim to perform oversampling using GPT-3 to balance the dataset. We also aim to expand the suicidal words list and extract additional features from the text that could enhance obtained results.

Ethics Statement

Secure access to the shared task dataset was provided with IRB approval under University of Maryland, College Park protocol 1642625 and approval by the Biomedical and Scientific Research Ethics Committee (BSREC) at the University of Warwick (ethical application reference BSREC 40/19-20).

Acknowledgements

The authors are particularly grateful to the anonymous users of Reddit whose data feature in this year's shared task dataset, to the annotators of the data for Task A, to the clinical experts from Bar-Ilan University who annotated the data for Task B, the American Association of Suicidology, to NORC for creating and administering the secure infrastructure and providing researcher support and to UKRI for providing funding to the CLPsych 2022 shared task organisers.

Table 2: Results of all models for subTask A and the best variant of RF for subTask B on the official test set. We boldface the best results.

SubTask A												
1- Coverage Based Metrics												
	Macro- Average			IS			IE			0		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Majority	nan	0.141	-	nan	0	-	nan	0	-	0.489	0.426	-
TFIDF-LR	**0.378**	0.424	-	0.111	0.008	-	**0.284**	**0.504**	-	**0.738**	**0.762**	-
BERT-Talklife-focal	0.260	0.204	-	0.025	0.007	-	0.226	0.093	-	0.530	0.513	-
LSTM-SpaCy	0.220	0.186	0.202	0.016	0.013	0.014	0.134	0.049	0.072	0.509	0.496	0.503
LSTM-GloVe	0.260	0.205	0.229	0.123	0.053	0.074	0.138	0.071	0.094	0.518	0.492	0.505
BERT	0.375	**0.440**	**0.405**	**0.253**	**0.372**	**0.301**	0.193	0.243	**0.215**	0.680	0.705	**0.692**
2- Post-level Metrics												
Majority	nan	0.333	0.280	nan	0	0	nan	0	0	0.724	1	0.840
TFIDF-LR	0.545	0.495	0.492	0.222	0.0243	0.044	0.569	**0.514**	**0.540**	0.844	0.947	**0.893**
BERT-Talklife-focal	0.522	0.386	0.380	0.090	0.012	0.022	**0.723**	0.163	0.266	0.753	**0.983**	0.853
LSTM-SpaCy	0.353	0.336	0.305	0.055	0.024	0.033	0.272	0.028	0.052	0.733	0.956	0.830
LSTM-GloVe	0.376	0.343	0.316	0.1	0.061	0.075	0.3	0.0288	0.052	0.729	0.939	0.821
BERT	**0.552**	**0.534**	**0.523**	**0.165**	**0.353**	**0.225**	0.609	0.389	0.475	**0.881**	0.860	0.871
SubTask B												
	Macro-Average			Low			Moderate			Severe		
Majority	0.156	0.333	0.213	nan	0	0	nan	0	0	0.469	**1**	**0.638**
TFIDF-LR	0.303	0.338	0.295	0	0	0	**0.428**	**0.214**	**0.286**	0.48	0.8	0.6
RF3	**0.305**	**0.423**	**0.297**	**0.166**	**0.666**	**0.266**	0.25	0.071	0.111	**0.5**	0.533	0.516

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*.

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Haitao Cai, Yucheng Ruan, Karen O'Connor, Gonzalez-Hernandez Graciela, Jeanmarie Perrone, and Abeed Sarker. 2021. Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC Medical Informatics and Decision Making*, 21(1):27.

Falwah AlHamed and Aljohara AlGwaiz. 2020. A Hybrid Social Mining Approach for Companies Current Reputation Analysis. In *Recent Advances on Soft Computing and Data Mining*, pages 429–438, Cham. Springer International Publishing.

Gérard Biau and Erwan Scornet. 2016. A random forest guided tour. *TEST*, 25(2):197–227.

Jonnathan Carvalho and Alexandre Plastino. 2021. On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis. *Artificial Intelligence Review*, 54(3):1887–1936.

Chun Yueh Chiu, Hsien Yuan Lane, Jia Ling Koh, and Arbee L.P. Chen. 2021. Multimodal depression detection on instagram considering time interval of posts. *Journal of Intelligent Information Systems*, 56(1):25–47.

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

David E. Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, volume 9822 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings*, volume 12260 of *Lecture Notes in Computer Science*, pages 272–287. Springer.

Bahman Mirheidari and Supervisor Heidi Christensen. 2019. Detecting early signs of dementia in conversation. (March).

Tomas Pranckevivius and Virginijus Marcinkevicius. 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Balt. J. Mod. Comput.*, 5.

Yada Pruksachatkun, Sachin R. Pendse, and Amit Sharma. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. pages 7685–7697.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022a. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

Bing Xiang Yang, Pan Chen, Xin Yi Li, Fang Yang, Zhisheng Huang, Guanghui Fu, Dan Luo, Xiao Qin Wang, Wentian Li, Li Wen, et al. 2022. Characteristics of high suicide risk messages from users of a social network—sina weibo “tree hole”. *Frontiers in psychiatry*, 13.

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*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.