Identification of depression and PTSD among Twitter users using pre-trained language model

Anonymous ACL submission

Abstract

Suicide is a global health issue and early diagnosis is necessary for effective treatment. Recent advancements in natural language processing has aided the identification of mental health disorders in social media. This paper investigated the efficacy of pre-trained language model (PLM) in identifying depression and post-traumatic stress disorder (PTSD) with Twitter data. Leveraging the CLPysch 2015 dataset (which constitutes of tweets from users with depression, PTSD and neither condition), we implemented various experimental designs using Long Short Term Memory (LSTM) and attention. The results demonstrate that while detecting specific mental health issues is still difficult, the detection of general mental health conditions improves with the implementation of attention. The results provide insights into the strengths and weaknesses of these models in identifying mental health issues from social media content, with potential implications for improving mental health monitoring.

1 Introduction

Suicide is a global health problem and is the fourth leading cause of death for the 15-44 years demographic globally (World Health Organization, 2021). Mental disorders, including depression and post-traumatic stress disorder (PTSD) have been found to increase the likelihood of suicidal ideation and suicide (Holliday et al., 2021; Busby Grant et al., 2023; Chou et al., 2023; Kratovic et al., 2021). These disorders not only hamper the quality of life for the people who suffer with them but also lessen the quality of life for their families and environment (García-Noguez et al., 2023). Early diagnosis and subsequent treatment can help to lessen the negative impacts that arise from mental health disorders (Beirão et al., 2020; Kearns et al., 2012).

Researchers are leveraging social context to better understand mental health problems and has been an ongoing process. In the past, researchers used Google trends for mental health surveillance (Page et al., 2011), examining depression based chatter on Twitter (Cavazos-Rehg et al., 2016), and implementing machine learning algorithms to classify tweets in terms of stress or relaxation (Doan et al., 2017). Recent advancements in pre-trained language models (PLMs) have been helpful in identifying the mental health disorder traits from textual data (Ji et al., 2021; Vajre et al., 2021).

Prior to PLMs, an early study conducted in 2014 as a part of a hackathon event (Coppersmith et al., 2014) performed a binary classification between the combinations of control, PTSD and depression outcomes based on the tweets gathered via Twitter api (Coppersmith et al., 2015). Following this research, the same dataset has aided other research, for example, interpreting mental health outcomes (Yang et al., 2023) and training new PLMs centric to mental health outcomes (Ji et al., 2021).

The aforementioned studies focused on binary classification to identify the presence or absence of depression among Twitter users. Given how these disorders may affect an individual differently, identification of PTSD and depression separately could influence an individual's journey to recovery (Finch, 2023). Proper diagnosis allows clinicians to recommend therapeutic interventions based on specific conditions (Finch, 2023; Kimberly Holland, Timothy J. Legg, 2019). As such, in this research, we extend the classification to all categories of CLPysch 2015 dataset. Additionally, we will replicate the experiments for detecting depression and further investigate the instances where general disorders are a concern.

2 Methodology

We aim to answer two key questions in this paper: 1. How effective are PLMs for tracking multiple mental health problems? 2. Which method is the most effective for monitoring general mental health conditions?

For this study, we used CLPsych2015 dataset (Details in Appendix: A) to answer these questions.

The experiments were run for all users using Algorithm 1. The number of epochs was set to 10. A single user was taken as their own batch for training because of the choice of model designs. Please refer to Section 3 for the model designs. We used *cardiffnlp/twitter-roberta-base* (Barbieri et al., 2020) as our PLM of choice. More details on Algorithm and PLM can be found in Appendix: B and C.

Algorithm 1 Training CLPysch 2015 dataset

for epochs (e_i) = 1 to e **do**
  for users (u_i) = 1 to u **do**
    Pre-process each tweet removing any punctuation, white space, links, retweets and emoticons
    Pass tweet to tokenizer and pre-trained RoBERTa and extract $[CLS]$ token
    Stack all $[CLS]$ tokens for user u_i
    Perform experiment E (section:3) on stacked $[CLS]$ embedding
    Two layers of neural networks with $tanh()$ and $softmax()$ to compute predicted $\hat{y}$
    Calculate loss and update weight
  end for
  Perform accuracy calculation for epoch e_i
end for

3 Experimental Designs

We describe two classes of implemented network models, each made up of 4 experiments. The first class of models used Long Short Term Memory (LSTM) and the second class of models were based on Attention mechanism, which is the engine of transformer-based models. We trained these model on the top of the PLM as described in Algorithm 1.

3.1 Long Short Term Memory (LSTM)

In our experiment, we implemented LSTM (Hochreiter and Schmidhuber, 1997) as one of the experiment designs. Since the tweets are sequential with each user having up to 1000 tweets and there are a differing number of tweets between the users, LSTM was appropriate as an experimental design. We implemented four LSTM models with variations in the number of layers and direction. The number of hidden layers ranged from 128 to 1024.

3.2 Using attention mechanism

Attention is the core of transformer based models (Vaswani et al., 2017). Since we are using RoBERTa for the base model (Barbieri et al., 2020), we added a multi-headed attention (MHA) layer of heads ranging from 1 to 16 for our second experiment design. This choice was made to attend to various parts of the tweet sequence differently. Four experiments were designed for the attention based models.

3.2.1 Attention

The idea behind this design was that the $[CLS]$ token would attend to a single tweet t_i and the stack of $[CLS]$ tokens from each user $t_n^{u_i}$ would use a cross-attention between tweets. This would determine the presence or absence of some mental health condition (depression or PTSD) for user u_i.

3.2.2 Adding temporal information

In this experiment, we added temporal information in terms of time lapse between the current and previous tweet as a part of the tweet. The first tweet t_1 was converted to $t_1 = "First\ tweet: " + t_1$ and every subsequent tweets were converted to $t_i = "After\ x:," + t_i$, where x was the time lapse between the current tweet t_i and the last tweet t_{i-1}, adding temporal context to the tweets. These were then processed in the same fashion as the attention as described in section 3.2.1.

3.2.3 Two sentence sliding window

For this experiment, we used two sentences appended together before the tokenization i.e. for user u_i, $t_{u_i} = t_1 + t_2, t_2 + t_3, \ldots, t_{n-1} + t_n$. A sliding window meant that there is a information linkage between previous and current tweet, creating a short term attention. The resulting stack of $[CLS]$ tokens would go through MHA layer for long term attention across the tweets, similar to section 3.2.1.

3.2.4 Momentum

In this experiment, we used the concept of momentum (α) for controlling the flow of information between t_i and t_{i-1}. For user u_i, the tweet t_i would convert to $\alpha * t_i + (1 - \alpha) * t_{i-1}$. The value of α ranged between 0 to 1 and it was initialised at 0.2 i.e. 20% of transfer of momentum from the last tweet. There was no particular reason of choosing 0.2 and could be randomised since we trained α while training the weights of the models.

4 Evaluation

Given p_1, p_2 and p_3 are probabilities for control, depression and PTSD respectively, the results were calculated as such for the mentioned three cases and presented in Tables 1, 2 and 3 respectively.

Case A. Multinominal classification: In this case, we performed the identification of control vs depression vs PTSD users based on the highest probability i.e. $max(p_1, p_2, p_3)$.

Case B. Depression vs Control: In this case, we removed the probability p_3 from all experimental results and rescaled the results for p_1 and p_2 and evaluated using the rescaled probabilities. This was done to compare our model results with the baseline models.

Case C: Mental health vs Control: In this case, we added the probability of p_2 and p_3 from all experimental results and evaluated using the new probability. This was done to simulate a scenario for the presence or absence of any general mental health condition.

In our experiments, 4 head two-sentence attention model achieved the best performance in both metrics for case A (Table: 1). Similarly, the same model performed best in F1 score for case B (Table: 2) and case C (Table: 3). However, recall was higher for 2 head Temporal, LSTM 2 layer 512 hidden unit and LSTM 2 layer bidirectional 1024 hidden unit for case B with each of them contributing to 100% recall score (Table: 2). Similar recall values can also be seen for the same models for case C along with single head two sentence (Table: 3). Although these models had a perfect recall score, it would not generalise well to unseen dataset as per their corresponding F1 scores. Since we are dealing with mental health conditions, there may be instances where not identifying mental health users are more costly than identifying false positives of the same, where these models might be useful.

The comparison with previous models was possible only for case B because case A and case C has not been exploited, and to the best of our knowledge, is novel to our study. Comparing to baseline models, our best model did not outperform F1 score of MentalRoBERTa model, but outperformed every other model, including large language models like GPT-4_{FS} and MentaLLaMA-chat-13B (Yang et al., 2024). For a relatively small model compared to some of these models, our model performed relatively well.

Identification of depression and PTSD separately resulted in decreased performance (Table: 1), compared to case B where only depression is identified (Table: 2) and case C, where, general mental health condition is identified (Table: 3), which is to be expected of a multinominal classification. Another possible explanation is the potential overlap of expressions in tweets from users with depression and PTSD. Consequently, the classification between the two groups becomes more challenging compared to the classification of an individual mental disorder from the control group alone. However, when these disorders are combined, the result improves significantly as seen from the results in case C (Table 3). So, if surveillance of general mental health condition is of interest instead of identifying individual conditions, we can achieve up to 75.5% of F1 score.

To answer the key questions of this research, PLMs may not be effective to identify individual mental health conditions, with our best model achieving only a maximum F1 score of 63.5%. However, when the conditions are combined for monitoring general mental health traits, the F1 score increases to 75.5%. This means in general, more than 3 out of 4 mental health patients could be diagnosed using their social media presence (eg: tweets). We found appending two tweets and passing through attention layer can help achieve this.

5 Conclusion

This study demonstrates the potential of pre-trained language models in detecting a range of mental health disorders, including depression and PTSD, from textual data on social media platforms like Twitter. Through various experimental models, including LSTM-based and attention based mechanisms, we were able to assess the effectiveness of these models in classifying specific and general mental health conditions. Our results reveal that attention-based models, particularly two-sentence sliding window tend to outperform other methods. The ability to classify specific and general mental health conditions separately could be crucial for more accurate diagnosis and treatment recommendations and there is a need for advanced monitoring systems that enable this. Further, research and optimization of these models could contribute significantly to early mental health diagnosis strategies.

Table 1: Performance metrics for control vs depression vs PTSD in multinominal classification setting (Case A)

Heads → / **Models ↓**	**F1 Score**					**Recall**				
	1	**2**	**4**	**8**	**16**	**1**	**2**	**4**	**8**	**16**
Attention	0.411	0.438	0.595	0.362	0.362	0.451	0.46	0.59	0.397	0.397
Temporal	0.29	0.31	0.606	0.554	0.461	0.363	0.378	0.603	0.562	0.495
Two sentence	0.222	0.434	**0.635**	0.457	0.542	0.333	0.449	**0.637**	0.494	0.553
Momemtum	0.312	0.333	0.333	0.333	0.333	0.361	0.369	0.369	0.369	0.369

Hidden units → / **Models ↓**	**F1 Score**				**Recall**			
	128	**256**	**512**	**1024**	**128**	**256**	**512**	**1024**
LSTM 1 layer	0.532	0.533	0.44	0.345	0.535	0.541	0.451	0.382
LSTM 2 layer	0.496	0.474	0.226	0.227	0.489	0.475	0.334	0.336
LSTM 1 layer bidirectional	0.523	0.49	0.478	0.437	0.519	0.494	0.479	0.434
LSTM 2 layer bidirectional	0.514	0.5	0.231	0.259	0.513	0.515	0.338	0.338

Table 2: Performance metrics for control vs depression in binary classification setting (Case B)

Heads → / **Models ↓**	**F1 Score**					**Recall**				
	1	**2**	**4**	**8**	**16**	**1**	**2**	**4**	**8**	**16**
Attention	0.37	0.43	0.616	0.303	0.303	0.3	0.407	0.78	0.213	0.213
Temporal	0.157	0.501	0.62	0.589	0.604	0.087	**1.0**	0.68	0.593	0.787
Two sentence	0	0.512	**0.655**	0.537	0.534	0	0.7	0.76	0.82	0.66
Momemtum	0.24	0.314	0.314	0.314	0.314	0.167	0.253	0.253	0.253	0.253
BERT-base	0.628					0.647				
MentalBERT	0.626					0.647				
MentalRoBERTa	**0.697**					0.703				
GPT-4$_{FS}$	0.62					-				
MentaLLaMA-chat-13B	0.526					-				

Hidden units → / **Models ↓**	**F1 Score**				**Recall**			
	128	**256**	**512**	**1024**	**128**	**256**	**512**	**1024**
LSTM 1 layer	0.589	0.589	0.484	0.333	0.627	0.687	0.52	0.26
LSTM 2 layer	0.502	0.503	0.501	0.116	0.507	0.48	**1.0**	0.093
LSTM 1 layer bidirectional	0.545	0.533	0.497	0.396	0.547	0.56	0.48	0.373
LSTM 2 layer bidirectional	0.568	0.573	0.026	0.502	0.6	0.693	0.013	**1.0**

Table 3: Performance metrics for control vs general mental health condition (depression and PTSD) in binary classification setting (Case C)

Heads → / **Models ↓**	**F1 Score**					**Recall**				
	1	**2**	**4**	**8**	**16**	**1**	**2**	**4**	**8**	**16**
Attention	0.63	0.593	0.724	0.403	0.403	0.587	0.547	0.83	0.277	0.277
Temporal	0.212	0.667	0.716	0.732	0.676	0.12	**1.0**	0.723	0.793	0.743
Two sentence	0.667	0.654	**0.755**	0.672	0.729	**1.0**	0.883	0.797	0.86	0.767
Momemtum	0.347	0.449	0.449	0.449	0.449	0.257	0.37	0.37	0.37	0.37

Hidden units → / **Models ↓**	**F1 Score**				**Recall**			
	128	**256**	**512**	**1024**	**128**	**256**	**512**	**1024**
LSTM 1 layer	0.684	0.689	0.59	0.431	0.707	0.733	0.567	0.337
LSTM 2 layer	0.608	0.587	0.667	0.138	0.6	0.53	**1.0**	0.093
LSTM 1 layer bidirectional	0.664	0.636	0.593	0.588	0.653	0.64	0.56	0.573
LSTM 2 layer bidirectional	0.667	0.663	0.039	0.668	0.697	0.733	0.02	**1.0**

Limitations

One of the limitations of the study is that only last 1000 tweets (if more than 1000 tweets present) per user were considered for this research due to computational restraints. For most of the experiments, we used a single A100 80GB GPU to train the models. Each experiment took 10-12 days (on average) to complete (approximately one epoch per day). While a second A100 80GB GPU was obtained at the tail end of training the models, most of the training was done using only a single A100 80GB GPU. Further, the GPU server was shared between various projects as well as the lack of resources to add more GPU servers meant that not all tweets could be processed. The reliance on the processing of tweets sequentially further meant that each epoch was much longer, since batching was not possible. This caused each model to run around 10-12 days, hence resulting in limited number of experiments. Further, only a single dataset was used, which could bias the results. In addition, the tweets were extracted a decade ago, which means the newer tweets would not have been collected. The vocabulary in which humans express sentiments perhaps changed in the last decade and those were not captured. Additionally, the collected tweets are only a sub-sample of the much larger cohort of mental health users who are not considered in this study. Even while focusing on this cohort itself, there is a lack of evidence to affirm the presence or absence of mental health conditions between the Twitter users. Finally, our study aims to develop a model for assisting researchers and clinicians for detection of mental health conditions using social context for non-clinical use. However, it does not replace clinical diagnoses which is essential for the detection and treatment of mental health issues.

Ethics Statement

The ethics was approved in accordance to Human Research Ethics Committee (HREC) approval number H15559. The data was already de-identified when it was received from Department of Computer Science, John Hopkins University.

Acknowledgements

We would like to thank Professor Mark Dredze for providing data and Department of Computer, Data and Mathematical Sciences, Western Sydney University for allowing to use the GPU clusters which allowed for the processing of the data.

References

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *arXiv preprint*. ArXiv:2010.12421 [cs].

Diogo Beirão, Helena Monte, Marta Amaral, Alice Longras, Carla Matos, and Francisca Villas-Boas. 2020. Depression in adolescence: a review. *Middle East Current Psychiatry*, 27(1):50.

Janie Busby Grant, Philip J. Batterham, Sonia M. McCallum, Aliza Werner-Seidler, and Alison L. Calear. 2023. Specific anxiety and depression symptoms are risk factors for the onset of suicidal ideation and suicide attempts in youth. *Journal of Affective Disorders*, 327:299–305.

Patricia A. Cavazos-Rehg, Melissa J. Krauss, Shaina Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J. Bierut. 2016. A content analysis of depression-related tweets. *Computers in Human Behavior*, 54:351–357.

Po-Han Chou, Shao-Cheng Wang, Chi-Shin Wu, and Masaya Ito. 2023. Trauma-related guilt as a mediator between post-traumatic stress disorder and suicidal ideation. *Frontiers in Psychiatry*, 14. Publisher: Frontiers.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

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*. Type: Journal article.

Son Doan, Amanda Ritchart, Nicholas Perry, Juan D. Chaparro, and Mike Conway. 2017. How Do You #relax When You're #stressed? A Content Analysis and Infodemiology Study of Stress-Related Tweets. *JMIR Public Health and Surveillance*, 3(2):e5939.

Company: JMIR Public Health and Surveillance Distributor: JMIR Public Health and Surveillance Institution: JMIR Public Health and Surveillance Label: JMIR Public Health and Surveillance Publisher: JMIR Publications Inc., Toronto, Canada.

Jon Finch. 2023. The Difference Between Depression and PTSD.

Luis Roberto García-Noguez, Saúl Tovar-Arriaga, Wilfrido Jacobo Paredes-García, Juan Manuel Ramos-Arreguín, and Marco Antonio Aceves-Fernandez. 2023. Automatic classification of depressive users on Twitter including temporal analysis. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 12(1):38.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. Type: Journal article.

Ryan Holliday, Claire A. Hoffmire, W. Blake Martin, Rani A. Hoff, and Lindsey L. Monteith. 2021. Associations between justice involvement and PTSD and depressive symptoms, suicidal ideation, and suicide attempt among post-9/11 veterans. *Psychological Trauma: Theory, Research, Practice, and Policy*, 13(7):730–739. Place: US Publisher: Educational Publishing Foundation.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. *arXiv preprint*. ArXiv:2110.15621 [cs].

Megan C. Kearns, Kerry J. Ressler, Doug Zatzick, and Barbara Olasov Rothbaum. 2012. Early Interventions for Ptsd: A Review. *Depression and Anxiety*, 29(10):833–842. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.21997.

Kimberly Holland, Timothy J. Legg. 2019. PTSD and Depression: Similarities, Differences & What If You Have Both.

Layla Kratovic, Lia J. Smith, and Anka A. Vujanovic. 2021. PTSD Symptoms, Suicidal Ideation, and Suicide Risk in University Students: The Role of Distress Tolerance. *Journal of Aggression, Maltreatment & Trauma*, 30(1):82–100. Publisher: Routledge _eprint: https://doi.org/10.1080/10926771.2019.1709594.

Andrew Page, Shu-Sen Chang, and David Gunnell. 2011. Surveillance of Australian Suicidal Behaviour Using the Internet? *Australian & New Zealand Journal of Psychiatry*, 45(12):1020–1022. Publisher: SAGE Publications Ltd.

Vedant Vajre, Mitch Naylor, Uday Kamath, and Amarda Shehu. 2021. PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1077–1082.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30. Type: Journal article.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

World Health Organization. 2021. Suicide worldwide in 2019: Global health estimates. *World Health Organization*. Type: Journal article.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards Interpretable Mental Health Analysis with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077. ArXiv:2304.03347 [cs].

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pages 4489–4500, New York, NY, USA. Association for Computing Machinery.

A CLPysch 2015 shared dataset

The CLPysch 2015 shared dataset contains publicly available tweets collected from the Twitter api over the period 2008 to 2013. The tweets were posted by users with PTSD, depression and a control group who did not have any identified mental health conditions as per tweets (Coppersmith et al., 2014). In total, there are 1145 training set and 599 testing set of anonymous users. Please note that the numbers may not match the original set due to the exclusion of users whose conditions were not recorded.
For this study, we used all available users and their subsequent tweets to identify their category of mental health condition, if present. Since the number of control (572 training, 299 testing) users were higher than depression (327 training, 150 testing) and PTSD (246 training, 150 testing) users, we used weighted cross entropy function for calculation of loss. However, the number of tweets was reduced to a maximum of last 1000 tweets per user out of a possible maximum of 3000 tweets per user due to computational constraints.

B Algorithm

The tweets went through pre-processing phase where the textual content was cleaned removing any white spaces, retweets, mentions, URLs, punctuations and emoticons. For each user u_i, their individual tweets $t_1, t_2, \ldots, t_n$ were tokenized and passed through a pre-trained RoBERTa model. The output was a tensor containing the embedding of the tweet t_i. The 768 dimension $[CLS]$ token, which contains the classification information of the entire sentence (Devlin et al., 2018), was extracted for each tweet. For each user, these $[CLS]$ tokens were then stacked to form the tensor of shape $t_{n^{u_i}} \times 768$, where $t_{n^{u_i}}$ was the number of tweets for user u_i. Further experiments were performed using these stacked tensors as explained in section 3. The output of each experiment was then connected to two fully connected layers, with $tanh()$ as the activation function on both layers. The first layer converted the output from 768 dimensions to 100 dimensions and the second layer converted from 100 dimensions to 3 dimensions. The output of the second fully connected layer was passed to softmax function, given by, $\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$, to convert the results into probabilities. The final output was the category (control, depression or PTSD) with the highest probability i.e. $max(\sigma(x_i))$.

C RoBERTa for base embeddings

We used a Twitter-based fine-tuned model of RoBERTa called *cardiffnlp/twitter-roberta-base* (Barbieri et al., 2020) for the base embeddings as our PLM. The embeddings were extracted using *transformer* library (Wolf et al., 2019). The small memory size of RoBERTa and its pre-training on Twitter data made it an appropriate choice for this study. There was an expectancy that the Twitter vocabulary was present in the PLM of choice since it was trained on Twitter data, thus providing appropriate token embeddings. For this study, since we were interested in embedding rather than the sentiment analysis, which was the intended use of this PLM, the $[CLS]$ token of each tweet's embedding was extracted using this PLM.