Towards Capturing Changes in Mood and Identifying Suicidality Risk

Sravani Boinepelli, Shivansh Subramanian,
Abhijeeth Singam, Tathagata Raha, and **Vasudeva Varma**
{sravani.boinepelli, shivansh.s,
tathagata.raha}@research.iiit.ac.in
abhijeeth.singam@students.iiit.ac.in
vv@iiit.ac.in
Information Retrieval and Extraction Lab,
IIIT-Hyderabad, Gachibowli, Hyderabad, Telangana, India

Abstract

This paper describes our systems for CLPsych's 2022 Shared Task[1]. Subtask A involves capturing moments of change in an individual's mood over time, while Subtask B asked us to identify the suicidality risk of a user. We explore multiple machine learning and deep learning methods for the same, taking real-life applicability into account while considering the design of the architecture. Our team, IIITH, achieved top results in different categories for both subtasks. Task A was evaluated on a post-level (using macro averaged F1) and on a window-based timeline level (using macro-averaged precision and recall). We scored a post-level F1 of 0.520 and ranked second with a timeline-level recall of 0.646. Task B was a user-level task where we also came in second with a micro F1 of 0.520 and scored third place on the leaderboard with a macro F1 of 0.380.

1 Introduction

Globally, close to 800,000 people die by suicide each year (WHO, 2014). Suicide is the fourth leading cause of death among 15-19 year-olds (WHO, 2021). Though suicide is such a dire issue, a myriad of obstacles such as social stigma, apprehensions about privacy, financial concerns, etc., prevent many from seeking professional help. Over the last couple of years, there has been an influx of suicide and mental health posts on social media, especially from the young users - social media's primary consumers. Anonymous social media platforms such as mental health blogs or Reddit forums have become increasingly popular as they can share their personal stories without judgment. People who face similar issues can share their experiences, give advice, motivate and persuade them to seek counsel from professionals. Therefore, social media has become a valuable source of linguistic cues for work to identify mental health problems from textual data (Cao et al., 2019; Masuda et al., 2013; Choudhury et al., 2016; Pruksachatkun et al., 2019). A challenge in the area of mental illness detection and suicide risk identification on social media is the importance of focusing on the individual and detecting the critical point where intervention is necessary from a batch of posts. This shared task (Tsakalidis et al., 2022a) breaks up the problem into two problem statements:

Subtask A: Given a user's posts over a certain period in time, this task aims to capture those subperiods during which a user's mood deviates from their baseline mood. This is defined as a post-level sequential classification task (Tsakalidis et al., 2022b). It encourages us to identify moments of change in the individual's mood over a timeline of about two months. A moment of change (MOC) is defined as a post/sequence of posts in a timeline indicating that the user's behavior or mental health status is shifted. This is represented in the form of the following labels: IS (indicating a switch in the user's mood), IE (indicating an escalation of the user's mood) and, O (refers to all other cases).

Subtask B: This is a user-level classification problem to predict the degree of suicide risk on Reddit. A user is considered to belong to one of four categories: No Risk (or "None"), Low, Moderate, and Severe Risk based on their posts on r/SuicideWatch (Shing et al., 2018; Zirikly et al., 2019). We present several approaches to tackle both subtasks, keeping in mind real-life application and the temporal aspect of this problem.

In the first subtask, we observed that the post-level classification would be influenced by its context. Hence, due to the longitudinal nature of the task, we use a transformer-based LSTM architecture. Post-level representations are generated using sentence transformer models and passed through an LSTM layer to consider historical context before developing the final output label.

[1] https://clpsych.org/sharedtask2022/

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 245 - 250

Our second task considers the need for detection mechanisms to continuously monitor suicide risk with the introduction of new posts to a user's history. We, therefore, first evaluate on a post-level using finetuned transformer models. We then adopt a majority voting strategy to assign the final label to the user. Our models outperform the baseline and rank in the top 3 submitted models for both subtasks across various categories.

2 Data

The dataset contains 255 timelines taken from users who have posted on mental health-related subreddits and /r/SuicideWatch. Each timeline consists of 10 to 122 posts each. The data given for this task is taken from 3 separate datasets. The E-Risk dataset (Losada et al., 2020; Losada and Crestani, 2016) is primarily used for Subtask A, while the Reddit datasets, such as the UMD dataset, are used for both subtasks (Shing et al., 2018; Zirikly et al., 2019). The dataset was split into a train and test dataset, with the training dataset having 149 users, 204 timelines, and 5143 posts, and the test dataset having 36 users, 51 timelines, and 1052 posts. Of the 149 users in the training dataset, 61 were labeled as 'Severe', 55 as 'Moderate', 11 as 'Low', and 22 remained unlabelled.

3 Baseline Experiments

We experiment with various popular machine learning, text classification algorithms on a post-level. Majority Voting is then applied for the task B experiments to generate the final label. Count Vectors and tf-idf vectors for different levels of input tokens (words, n-grams, etc.) served as the primary features for most of our baseline models. We used Scikit-learn (Pedregosa et al., 2011) and Keras (Chollet et al., 2015) libraries to develop and evaluate the models.

Logistic regression(LR): The logistic regression model uses tf-idf and n-grams as features for our baseline. Hyperparameter tuning proved the model to work best for ranges of unigrams and bigrams.

Random Forest Model(RF): Decision trees tend to overfit on the training set. Random decision forests with bagging help correct this behavior. We test their performance against tf-idf word-level vector and count vector features.

Xtreme Gradient Boosting Model(Xgb): The boosting algorithm is popularly used to optimize the performance of decision trees by reducing bias and variance. We test the performance of this model against other baselines with count vectors and word-level tf-idf.

4 Final architecture

4.1 Experimental Settings

The HuggingFace transformers library's RoBERTa model was used for finetuning, and all our architectures were implemented in Pytorch (Paszke et al., 2019). Our MOC-LSTM model was run with a learning rate of 2e-06 and a batch size of 8, while our finetuned RoBERTa model was run with batch size 16. The model uses the AdamW optimizer with an initial learning rate of 2e-5 and a linear warm-up schedule.

4.2 Detecting MOCs from a User timeline

To detect switches in a user's mood, our model must retain the knowledge of user history to assign the present post label. Therefore, our model uses an LSTM-based (Gers et al., 2000) architecture to capture the essence of the previous context and generate the labels for the latest posts. We initially convert the content and titles of each Reddit post within a timeline into 384-dimensional embeddings using paraphrase-MiniLM-L6-v2 (Reimers and Gurevych, 2019).

Based on the dataset, the maximum number of posts by a user was 125, so we created a hard limit of 128 posts per user. This required us to pad posts to users who had posted less than 128 posts to make them equal in length but allowed us to do batch-wise computation. We also replace specific posts with no content or title with pad tokens.

Once this preprocessing is complete, we use the sentence transformer to get text embeddings for each post's content and title. We took the truncated text as required by the transformer model. We concatenated the content and title representation to get a single post representation and used that as input to the LSTM layer. A window_size amount of posts is sent at a time to limit the previous posts' influence on the current output. The output of this LSTM layer was then used for the post-wise classification. We experimented with different window_sizes to see the effect of previous information on the quality of predictions. When comparing sizes 4 and 8, we found that window_size = 4 gives us the best results. We added a linear layer and SoftMax activation to get

probability distribution over the three classes. The class with the highest probability was considered the model's prediction. The final loss is computed based on WeightedCrossEntropyLoss to reflect the bias in the dataset. This gave superior results to regular cross entropy loss.

We then changed the embedding model from paraphrase-MiniLM-L6-v2 to robertaSTSB. The embedding dimension for RoBERTa was different since each text gave an output of 768 dimensions. We observed that RoBERTa embeddings performed equal or better in almost all parameters compared to paraphrase-MiniLM. But due to technical issues on our side, we could not use RoBERTa in our final submission. This experiment was performed on a validation set (80-20 split).

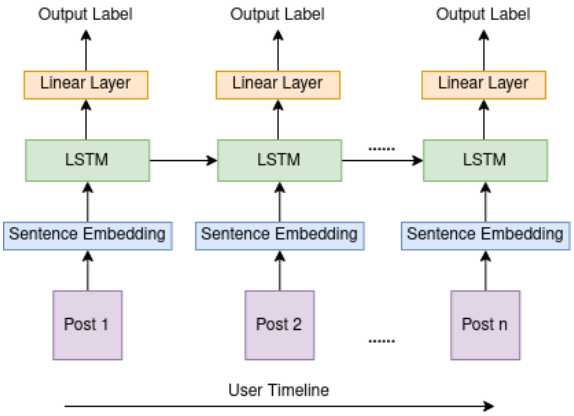

Figure 1: Framework to detect MOCs from a user timeline

4.3 Finetuned RoBERTa model for Assessing Suicidality Risk

Transformers and transfer learning architectures have previously achieved SOTA results in multiple datasets. They combine positional encoding and non-sequential single input processing to achieve better long-range dependencies than LSTMs and RNNs. The attention mechanisms are targeted toward such sequential data. Therefore, our final architecture uses RoBERTa for predicting suicide risk for a given text by fine-tuning it for classification on the given dataset. We also consider using sampling techniques to correct the imbalance in the dataset. However, models like RoBERTa are known to perform better on imbalanced data sets rather than on oversampled augmented datasets (Tayyar Madabushi et al., 2019). Weighted Random Sampling was preferred over other under-sampling techniques based on the results of our experiments.

The cleaned dataset is passed to the RoBERTa tokenizer. The length of each input is fixed to be 256 tokens. Only the first embedding produced by the RoBERTa for Sequence Classification Model is used for classification. This embedding is then passed to a linear layer and produces logits used to predict the post-level labels. Backward propagation of the loss is performed, and the weights of the linear layer and the model are updated. This fine-tuning process helps the model learn the unique domain related to the problem.

We now have the labels for each post made by the user. However, it is difficult to determine the number of posts (with a certain level of severity) required to assign a final label to the user. Because the span of each timeline is about two months, the most straightforward approach would be to simply take the most occurring label as the final assigned user label. This is called majority voting. Given that the posts we are considering represent the user's 'n' most relevant posts, we postulate that the 'degree' of suicide risk of the user can be ascertained by simply taking the mode of the outputted 'n' post labels. This approach performed well on the leaderboard for this task. Our team came in second with a micro F1 of 0.520 and scored third place with a macro F1 of 0.380.

4.4 Finetuning MOC-LSTM

We also tried to leverage our results from Task A to improve performance on Task B. As a preprocessing step, we assumed that the user's risk level is similarly reflected in their timeline's risk level. Once we do that, the task becomes a timeline-level classification task, and we determine the user's final label based on majority voting.

We used transfer learning to classify the entire timeline and give us better results. Our initial model was trained on task A (as described in MOC-LSTM) for the post-level classification task. We had to take special care of the window_size in this approach since our primary goal is a timeline-level classification. We used a window_size = 128, implying that all the posts are considered simultaneously. Hence, the output of the pre-trained model was a probability distribution over the three classes: IS, IE, and O, for 128 posts. We then utilize the pre-trained model, learned on task A, to finetune the task B

dataset. We added another linear layer followed by SoftMax to the output of the task A model to combine the post-level classification into a final timeline-level classification. Hence, we combined the post-level probability distribution to get the timeline-level probability distribution. Once we got out probability distribution, we used CrossEntropyLoss to train our model and Adam optimizer. Though the model performs well, we were unable to officially submit it to the shared task due to technical issues from our side.

4.5 Evaluation metrics

For Subtask A, the post-level results are calculated using macro F1 scores (represented as 'M-F1' in Table 2. The coverage and window-based results are evaluated using Precision('P') and Recall('R') oriented scores as specified by the shared task organizers. The details for calculating these evaluation metrics may be found in their overview of the shared task (Tsakalidis et al., 2022a).

Subtask B was evaluated using Macro and Micro averaged F1 scores. We look at the range per timeline and the distribution of labels. Since the range for each timeline in the dataset is about two months, we propose a majority voting approach. This performed well, and our model ranked in the top three with both macro F1 and micro F1 scores. However, this may fall short for more extended time periods, at which point it becomes increasingly imperative to adopt a more longitudinal and temporal approach to calibrate the level of a user's suicide risk.

5 Results and Analysis

The comparison between the baseline models and our main models can be found in Table 1. LogisticRegression with CountVectorizer was the best baseline for both tasks with respect to Macro-F1. Our model MOC-LSTM beats the baselines of Task A comfortably with a 0.05 increase in Macro-F1. For task B, the finetuned MOC-LSTM has a slight edge over the baselines, whereas the finetuned RoBERTa model scores significantly better with a 0.08 F1 over the baseline models.

Results on the unseen test set for our submitted models can be found in Table 2 as provided to us by the workshop organizers. The 'Baseline' results belong to the Logistic Regression model trained on tf-idf features supplied by the organizers. Values in bold are amongst the top 3 ranked by the shared task. Our final submission included the 'MOC-LSTM', and Finetuned RoBERTa models. Our MOC-LSTM model scores a post-level F1 of 0.520 and ranks second with a timeline-level recall of 0.646. Our Finetuned-RoBERTa model ranks second with a micro F1 of 0.520 and scored third place with a macro F1 of 0.380.

Task	Model	Macro-F1
Task A	RF, Count	0.48
	RF, tf-idf, Word lvl	0.48
	Xgb, Count	0.50
	Xgb, tf-idf, Word lvl	0.49
	LR, Count	**0.54**
	LR, tf-idf, Word lvl	0.47
	LR, tf-idf, N-Gram	0.47
	MOC-LSTM	**0.59**
Task B	RF, Count	0.43
	RF, tf-idf, Word lvl	0.42
	Xgb, Count	0.40
	Xgb, tf-idf, Word lvl	0.40
	LR, Count	**0.46**
	LR, tf-idf, Word lvl	0.45
	LR, tf-idf, N-Gram	0.44
	Finetuned MOC-LSTM	**0.51**
	Finetuned RoBERTa	**0.54**

Table 1: Comparing baseline results with the final models for Task A and B on an 80/20 split of the training data.

Eval_type	Model	P	R	M-F1
Task A, Window	Baseline	0.496	0.539	-
	IIITH	0.530	**0.646**	-
Task A, Coverage	Baseline	0.377	0.424	-
	IIITH	0.346	0.405	-
Task A Post-level	Baseline	0.545	0.495	0.492
	IIITH	0.520	0.600	0.520
Task B, Macro-avg	Baseline	0.302	0.338	0.295
	IIITH	**0.396**	0.407	**0.380**
Task B, Micro-avg	Baseline	0.412	0.468	0.406
	IIITH	**0.538**	**0.562**	**0.520**

Table 2: CLPsych 2022 Official Results on the test set.

6 Conclusion

In this shared task, we have worked towards detecting moments of change and the suicidality risk of a user based on their post history. Our MOC-LSTM model allows us to determine post-level information and timeline level classification, enabling us to better understand mental health by identifying

the specific moments of change in emotions. The finetuned RoBERTa model works to identify at-risk users based on their post timeline to better care for them. In the future, we plan to consider multi-modal approaches for different social media platforms. This helps give a better picture of the user's mental health since people use different media for different purposes. We also plan to build more efficient models that work on longer timelines to provide these warnings in a real-time platform-agnostic manner and help identify at-risk users.

7 Ethical 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).

8 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.

References

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1718–1728, Hong Kong, China. Association for Computational Linguistics.

François Chollet et al. 2015. Keras. `https://keras.io`.

Munmun De Choudhury, Emre Kıcıman, Mark Dredze, Glen A. Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12:2451–2471.

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.

Naoki Masuda, Issei Kurahashi, and Hiroko Onari. 2013. Suicide ideation of individuals in online social networks. *PLoS One*, 8(4):e62262.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China. 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.

WHO. 2014. Preventing suicide: A global imperative.

WHO. 2021. Suicide.

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.