DDP Stage 2 Presentation: Mental Disorder Identification through Temporal Representation of Text

Raja Kumar [190110070]

Centre for Digital Health
IIT Bombay

22nd June 2024

External Examiner: Prof. Diptesh Kanojia
Internal Examiner: Prof. Raj Dabre

Guide: Prof. Pushpak Bhattacharyya

Outline

- Problem Statement
- Background and Motivation
- Literature Survey
- Contributions
- Datasets
- Technique
- Experiments and Results
- Analysis
- Summary
- Conclusion
- Future Work
- EMNLP Submission Under Review

Problem Statement

Mental Disorder Identification through Temporal Representation of Text

Input: Social media posts in chronological order

Output: Presence or Absence of the Disorders: Binary Classification

Mental Disorders under consideration

- Anorexia
- Depression
- Self-Harm

Background

Types of Mental Disorders

- Mood disorders (such as depression and bipolar disorder)
- Anxiety disorders
- Personality disorders
- Psychotic disorders (such as schizophrenia)
- Eating disorders
- Trauma-related disorders (such as post-traumatic stress disorder)
- Substance abuse disorders

- Self-harming behavior is not a mental disorder in itself, but it is often a symptom of an underlying mental health issue.

Motivation: Scarcity of Mental Health Professionals

- About 970 million mental or neural disorders
- 14.3% deaths (approximately 8 million) worldwide are identified as mental-health originated
- 1:100000- Psychiatrist : Patients
- A study co-led by researchers from Harvard Medical School and the University of Queensland further reveals that over 50% of individuals worldwide experience a mental health disorder during their lifetime

Motivation: Personal Pronouns and emotional words can reveal aspects of Psychological State

- Interestingly, heightened usage of first-person singular pronouns is associated with feelings of grief, depression, or thoughts of suicide (Rude et al. (2004); Boals and Klein (2005); Eichstaedt et al. (2018)).

- Positive and negative emotional words are linked to mental health, with a excess of negative emotional words and a scarcity of positive ones reflecting an unhealthy mental state (Pennebaker et al. (2003); Kahn et al. (2007)).

- Language related to suicidal ideation contains approximately 30% more absolutist words than language associated with anxiety and depression and roughly 80% more than language reflective of normal mental health (Al-Mosaiwi and Johnstone et al. (2018)).

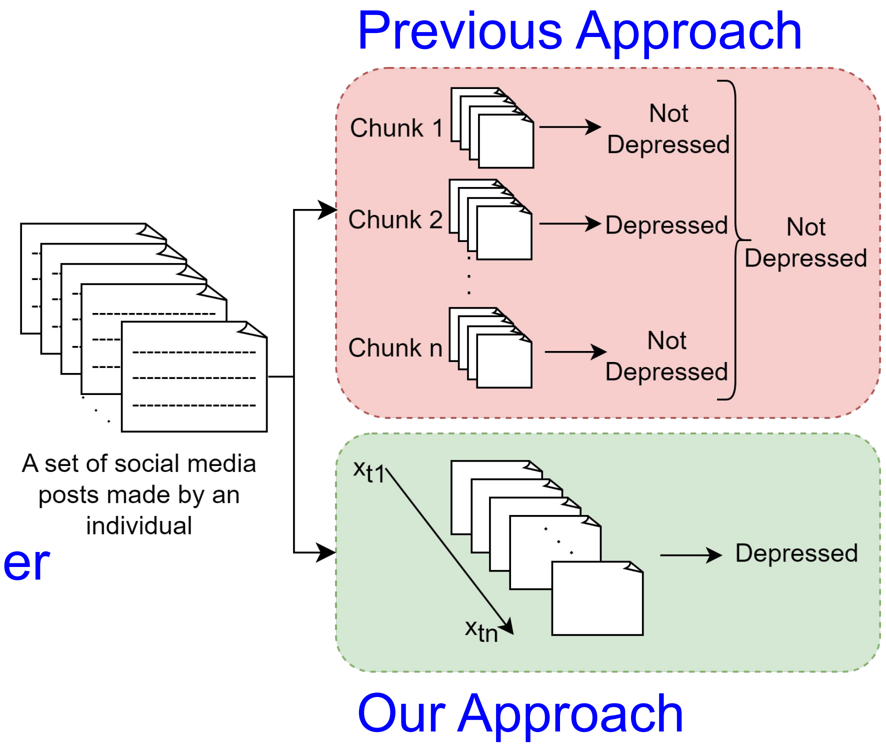
Motivation: Temporality and Computation

- Current approaches for the automatic detection of mental disorders need to be made aware of the temporal nuances of textual data.

- LLM-based approaches are too computationally expensive to train and perform poorly in low-resource settings like mental health intervention, where data is limited.

- Loss of temporal information
- Lack of global view
- Semantic noise

- Preserve the post-identity and order
- Global classification



Literature Survey

Related Work

- **Simms et al., 2017** used LIWC features and applied machine learning to detect disorders such as anxiety, anorexia, and depression.

- **Guntuku et al., 2018** used a skip-gram model to learn word embeddings and then trained ML models along with LIWC features to predict anxiety.

- **Gaur et al., 2021** developed an architecture stacking CNN over LSTM to predict user-level suicidality.

- **Reece et al., 2017** was the first to employ state-space temporal analysis for depression detection, but a significant limitation was their reliance on low-level features.

MentalBERT(Ji et al. 2022) and DisorBERT(Aragon et al. 2023)

- **MentalBERT** is a pre-trained language model designed for mental healthcare. The training corpus comprised a total of 13,671,785 sentences from Reddit.

- **MentalBERT** was trained on approximately 13,671,785 sentences over the course of eight days, utilizing four Tesla V100 GPUs.

- In **DisorBERT**, the concept involves initially instructing BERT on the broad language patterns found in a large-scale social media platform like Reddit.

- Following that, the model is fine-tuned to cater to the language specific to users with mental disorders.

Contributions

- A first-of-its-kind representation method which transforms textual data from social media posts into a time series format to capture the time-dependent patterns of a patient. This provides a compressed representation of the textual data while reducing the floating point operation by at least 330 times compared to SOTA.

- A novel framework incorporating temporal data for mental disorder identification via foundational deep learning models (LSTM and 1D CNN) which surpasses the performance of BERT-based approaches by 5% in the F1 score on three different mental conditions: Depression, Self-harm, and Anorexia.

- A cross-domain study of the three disorders to understand the commonality across disorders. We investigate the possibility of cross-domain data usage, which can further benefit the identification of low-resource mental disorders.

Dataset: RMHD

Reddit Mental Health Dataset

- This dataset includes posts from 15 mental health support groups, centered around discussions on various mental health issues.

- Specifically, we focused on posts from three subreddits: r/EDAnonymous, r/depression, and r/suicidewatch, with each post labeled based on self-reported diagnoses.

	r/ED	r/depr	r/suicide
total # posts	9535	58089	41354
avg # words	129.59	190.69	171.25

Dataset Statistics: eRisk

- The datasets contain Reddit users' post histories.

	Training		Validation		Test		Total
	Condition	Control	Condition	Control	Condition	Control	
Anorexia							
#subjects	45	332	14	81	73	742	1307
avg # posts	404.7	552.3	411.9	560.9	241.4	745.1	639.44
avg # words	36.2	21.1	39.6	20.9	37.2	21.7	23.10
Depression							
#subjects	173	1195	44	298	40	49	1799
avg # posts	444.9	663.4	436.7	658.2	493.0	543.7	629.31
avg # words	24.2	20.55	29.8	24.77	39.2	45.6	22.91
Self-Harm							
#subjects	29	243	12	56	104	319	763
avg # posts	172.0	543.9	167.8	549.7	112.4	285.6	357.52
avg # words	22.4	17.5	26.8	19.7	21.4	11.9	16.17

- User classification into the "Condition" group was based on responses to a depression questionnaire or self-reported clinical diagnoses on Reddit.

Data Instance Example

23 OCT 2019- 18:18:43

Interesting how you did the letter grades... I personally would have represented each letter with ten percent (A would be 90-100, B would be 80-89, etc.) until F, which would be 60.

22 OCT 2019- 23:11:01 - No but seriously guys my friends keep talking about banging a guy named Joe when I'm around I need help quickly it's getting out of hand

21 OCT 2019- 23:52:27 - Im not questioning the validity in that, but how?

20 OCT 2019- 00:54:43 - ^its ^satire ^we ^don't ^actually ^want ^you ^to ^die

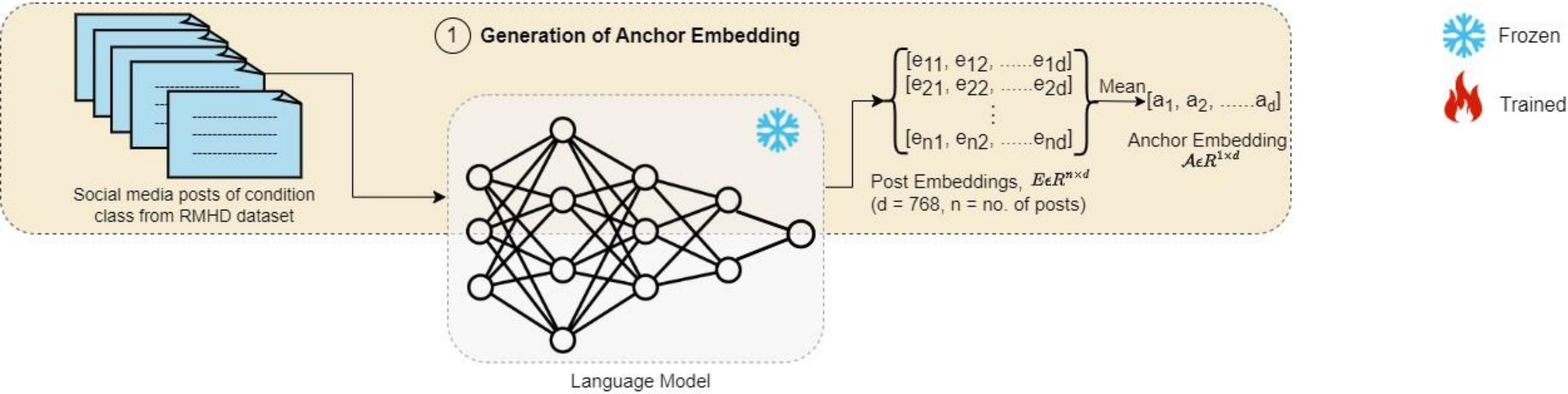
19 OCT 2019- 00:48:29 - Well, I guess you are pardoned.

…

Architecture

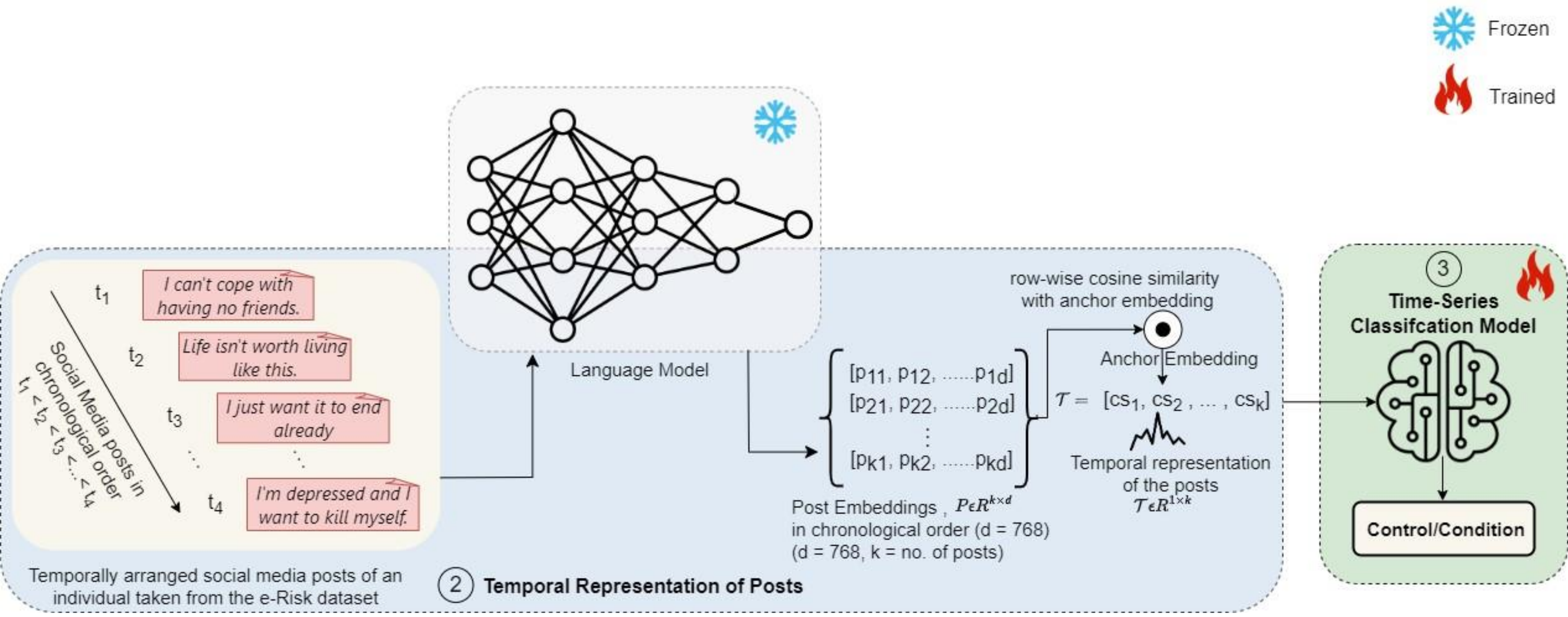
Framework: Anchor Embedding

First we took the posts from RMHD and generate a post wise embedding representation (Model: all-mpnet-base-v2). These embedding representations were averaged by mean operation to generate a final “anchor embedding”

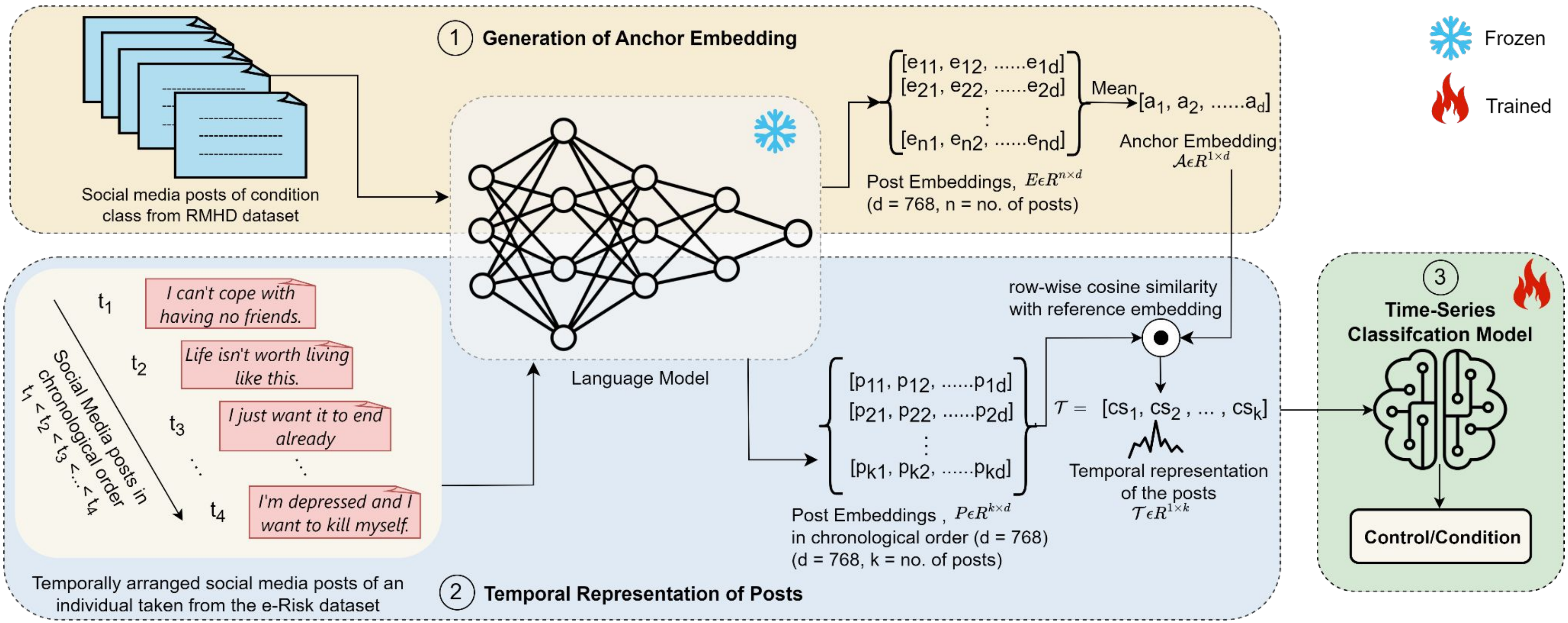


Framework: Temporal Representation

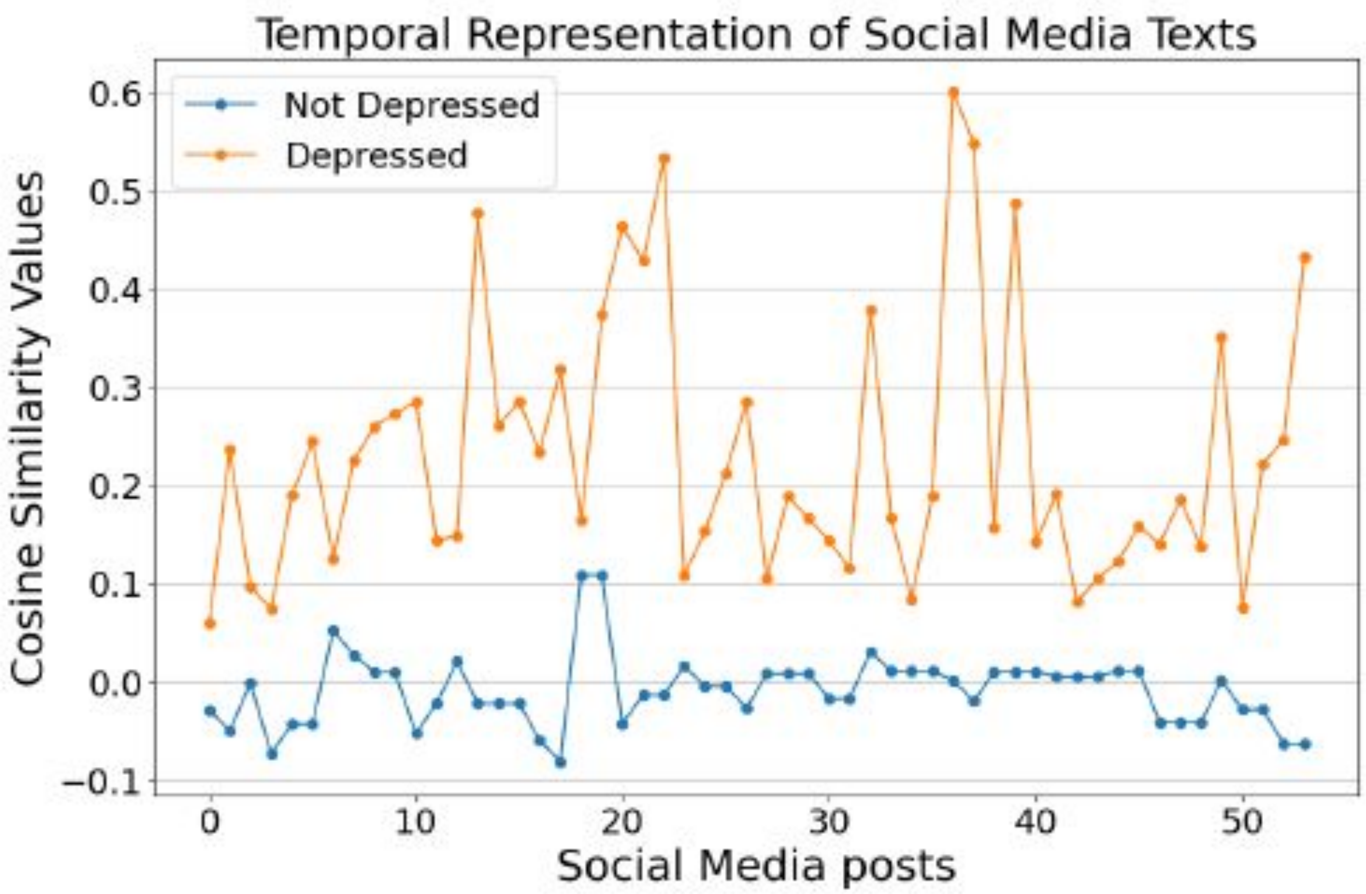
Taking mean vector as anchor embedding, we calculated time series of cosine similarities using sentence embedding of sequential posts. We performed time series classification using feature based approaches and representation learning methods.



Combined Framework



Temporal Representation Example



A distinction between depressed and non depressed classes in temporal representation

Results

Making Baseline Results Stronger

Impact of Different Language Models (LMs):

- The use of different LMs (e.g., MPnet vs. BERT) introduces a variable.
- **Raises the question**: Is performance improvement due to the superior LM or the post-level information?

Incorporating RMHD Data:

- Baseline models should incorporate RMHD data.
- Fine-tuning to access and utilize pertinent information from RMHD data.

Additional Baseline Experiments:

- Conducted baselines with various models fine-tuned on eRisk data and eRisk + RMHD data.
- **Models used for fine-tuning**: MPnet, Roberta and Deberta

Baseline Results: BERT, MentalBERT, DisorBERT and LLMs

- For each user, researchers divided their post history into N=35 segments.
- In the testing phase, each segment is labeled either 1 or 0 subsequently if the majority of these segments contain positive labels.

Method		Anorexia			Depression			Self-Harm		
	Masking	F1	P	R	F1	P	R	F1	P	R
Baselines										
BERT	Random	0.77	0.70	0.85	0.62	0.55	0.72	0.60	0.44	0.94
MentalBERT	Random	0.76	0.66	0.89	0.67	0.57	0.80	0.71	0.62	0.84
BERTw/Reddit	Random	0.81	0.75	0.88	0.66	0.56	0.80	0.71	0.66	0.76
BERTw/Reddit	Guided	0.82	0.82	0.82	0.68	0.55	0.90	0.72	0.65	0.82
BERTw/Health	Random	0.80	0.77	0.84	0.67	0.53	0.93	0.69	0.60	0.82
BERTw/Health	Guided	0.82	0.81	0.84	0.68	0.57	0.85	0.74	0.72	0.76
DisorBERT	Random	0.82	0.83	0.81	0.68	0.54	0.93	0.72	0.65	0.80
DisorBERT	Guided	**0.83**	0.82	0.85	0.69	0.56	0.89	0.72	0.73	0.71
MPNetv2 (ZS)*	-	0.16	0.09	1.00	0.62	0.45	1.00	0.40	0.25	1.00
MPNetv2 (FT [eRisk])*	-	0.71	0.60	0.89	0.62	0.57	0.68	0.48	0.89	0.33
MPNetv2 (FT [eRisk+RMHD])*	-	0.78	0.73	0.85	0.62	0.45	1.00	0.42	0.27	0.98
GPT-3.5-turbo*	-	0.05	1.00	0.03	0.37	1.00	0.23	0.22	0.93	0.12
MentalLLaMA-chat-13B*	-	0.08	1.00	0.04	0.05	1.00	0.03	0.02	0.50	0.01

The first eight baseline values were taken from Aragon et al. (2023), and the baselines marked with * were trained by us

Additional Baselines Results

Results of additional experiments on MPNet-base, RoBERTa-base and DeBERTa-base models.

Model	Anorexia			Depression			Self-Harm		
	F1	P	R	F1	P	R	F1	P	R
MPNet (ZS)	0.16	0.09	1.00	0.62	0.45	1.00	0.40	0.25	1.00
MPNet (FT on eRisk)	0.53	0.37	0.90	0.71	0.58	0.93	0.67	0.82	0.57
MPNet (FT on eRisk+RMHD)	0.20	0.11	0.99	0.61	0.44	0.97	0.42	0.27	0.98
DeBERTa (ZS)	0.16	0.09	1.00	0.62	0.45	1.00	0.40	0.25	1.00
DeBERTa (FT on eRisk)	0.73	0.62	0.89	0.59	0.61	0.57	0.36	0.86	0.23
DeBERTa (FT on eRisk+RMHD)	0.23	0.13	0.99	0.61	0.44	0.97	0.42	0.27	0.98
RoBERTa (ZS)	0.16	0.09	1.00	0.62	0.45	1.00	0.40	0.25	1.00
RoBERTa (FT on eRisk)	0.68	0.55	0.89	0.63	0.59	0.68	0.36	0.96	0.22
RoBERTa (FT on eRisk+RMHD)	0.23	0.13	0.99	0.61	0.44	0.97	0.42	0.27	0.97

F1, precision (P), and recall (R) values are reported over the condition class in three e-Risk tasks: Anorexia, Depression and Self-Harm. ZS and FT refer to Zero-Shot and Fine-Tuned experiments respectively.

Our Results in terms of P, R, F1-score of Disorder Class

Method		Anorexia			Depression			Self-Harm		
	Masking	F1	P	R	F1	P	R	F1	P	R
Baselines										
BERT	Random	0.77	0.70	0.85	0.62	0.55	0.72	0.60	0.44	0.94
MentalBERT	Random	0.76	0.66	0.89	0.67	0.57	0.80	0.71	0.62	0.84
BERTw/Reddit	Random	0.81	0.75	0.88	0.66	0.56	0.80	0.71	0.66	0.76
BERTw/Reddit	Guided	0.82	0.82	0.82	0.68	0.55	0.90	0.72	0.65	0.82
BERTw/Health	Random	0.80	0.77	0.84	0.67	0.53	0.93	0.69	0.60	0.82
BERTw/Health	Guided	0.82	0.81	0.84	0.68	0.57	0.85	0.74	0.72	0.76
DisorBERT	Random	0.82	0.83	0.81	0.68	0.54	0.93	0.72	0.65	0.80
DisorBERT	Guided	**0.83**	0.82	0.85	0.69	0.56	0.89	0.72	0.73	0.71
MPNetv2 (ZS)*	-	0.16	0.09	1.00	0.62	0.45	1.00	0.40	0.25	1.00
MPNetv2 (FT [eRisk])*	-	0.71	0.60	0.89	0.62	0.57	0.68	0.48	0.89	0.33
MPNetv2 (FT [eRisk+RMHD])*	-	0.78	0.73	0.85	0.62	0.45	1.00	0.42	0.27	0.98
GPT-3.5-turbo*	-	0.05	1.00	0.03	0.37	1.00	0.23	0.22	0.93	0.12
MentalLLaMA-chat-13B*	-	0.08	1.00	0.04	0.05	1.00	0.03	0.02	0.50	0.01
Our Methods										
Feedforward Network	-	**0.83**	0.87	0.79	0.71	0.83	0.59	0.81	0.84	0.78
1D-CNN	-	0.82	0.86	0.78	0.70	0.77	0.65	**0.83**	0.85	0.81
LSTM	-	0.79	0.84	0.74	**0.75**	0.79	0.71	**0.83**	0.93	0.75
Transformer	-	0.82	0.85	0.79	0.71	0.83	0.61	0.74	0.81	0.67

F1, precision (P), and recall (R) values over the condition class in three eRisk tasks: anorexia, depression and self-harm.

Results using ML methods

This approach involved extracting statistical features from the time series data. We then perform feature selection based on Gini impurity (Yuan et al., 2021) criteria to get top 30 features by incorporating a Random Forest classifier.

Method	Anorexia			Depression			Self-Harm		
	F1	P	R	F1	P	R	F1	P	R
Decision Tree	0.66	0.66	0.66	0.54	0.74	0.42	0.69	0.71	0.67
XGBoost	0.74	0.83	0.67	0.55	0.77	0.42	0.74	0.86	0.64
Adaboost	0.74	0.81	0.68	0.50	0.88	0.35	0.78	0.87	0.71
Random Forest	0.75	0.86	0.67	0.57	0.85	0.42	0.78	0.84	0.72
LightGBM	0.81	0.86	0.77	0.53	0.88	0.38	0.83	0.90	0.77

F1, precision (P), and recall (R) result over the condition class in three eRisk tasks.

Analysis

Efficiency Analysis

Objective: To study the computational efficiency of our framework, we report the number of floating point operations (Flos) in a single forward pass (Kaplan et al.,2020)

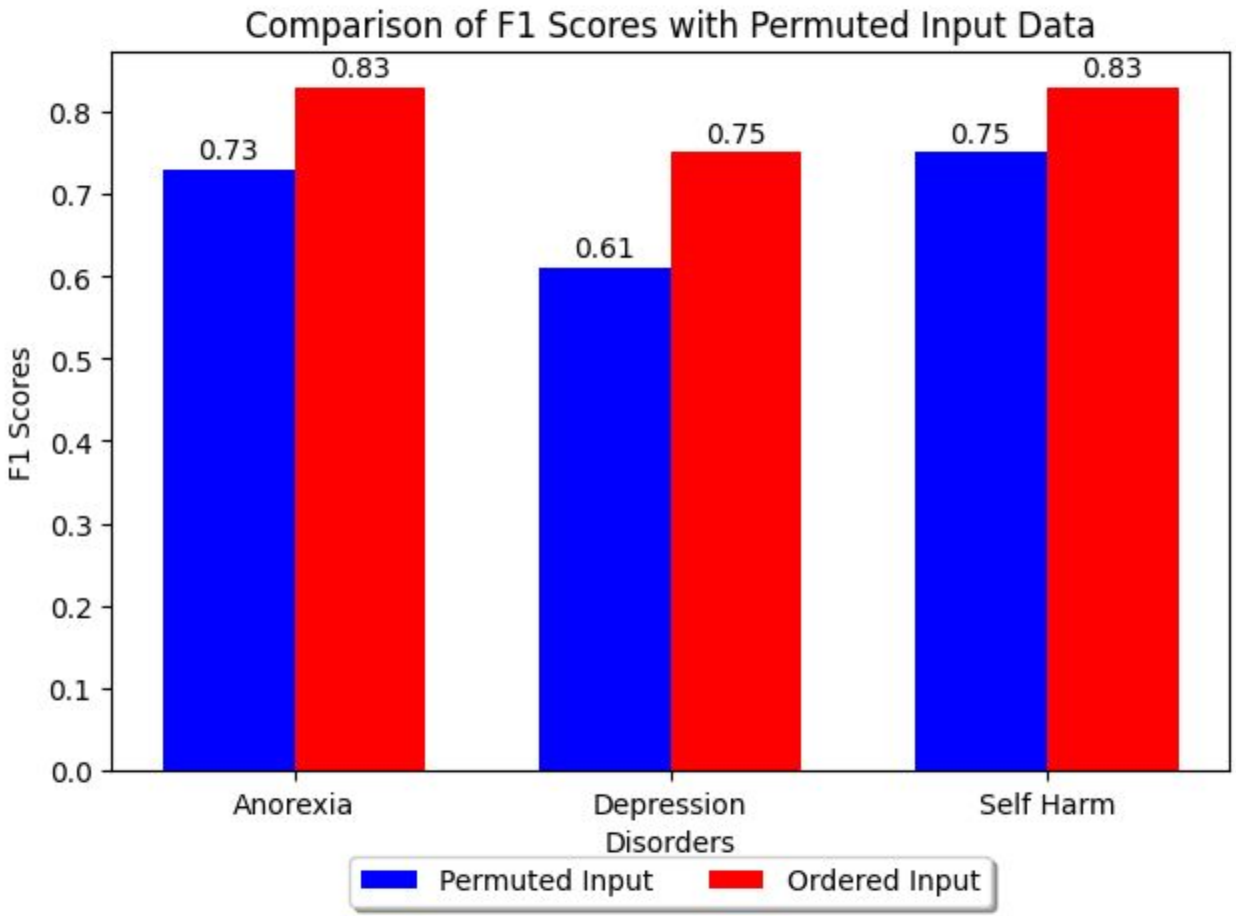
Findings: Reduces the total number of floating point operations by 330 times in the worst-case scenario

Model	Anor	Depr	SH
Feedforward	2.47 K	2.51 K	1.89 K
LSTM	4.18 K	4.23 K	4.30 K
1D-CNN	25.44 M	5.87 M	5.85 M
Transformer	14.61 M	14.62 M	14.56 M
BERT*	8.42 B	8.42 B	8.42 B
MPNet*	8.42 B	8.42 B	8.42 B
MLLaMA13B*	1.75 T	1.75 T	1.75 T

Total number of Floating points operation required for a single forward pass. Here, "Anor", "Depr", and "SH" stand for anorexia, depression, and self-harm. Models marked with * are baselines.

Temporal Analysis

Objective: To understand the impact of temporal order on performance, we train the model after permuting the input data five times in random order.

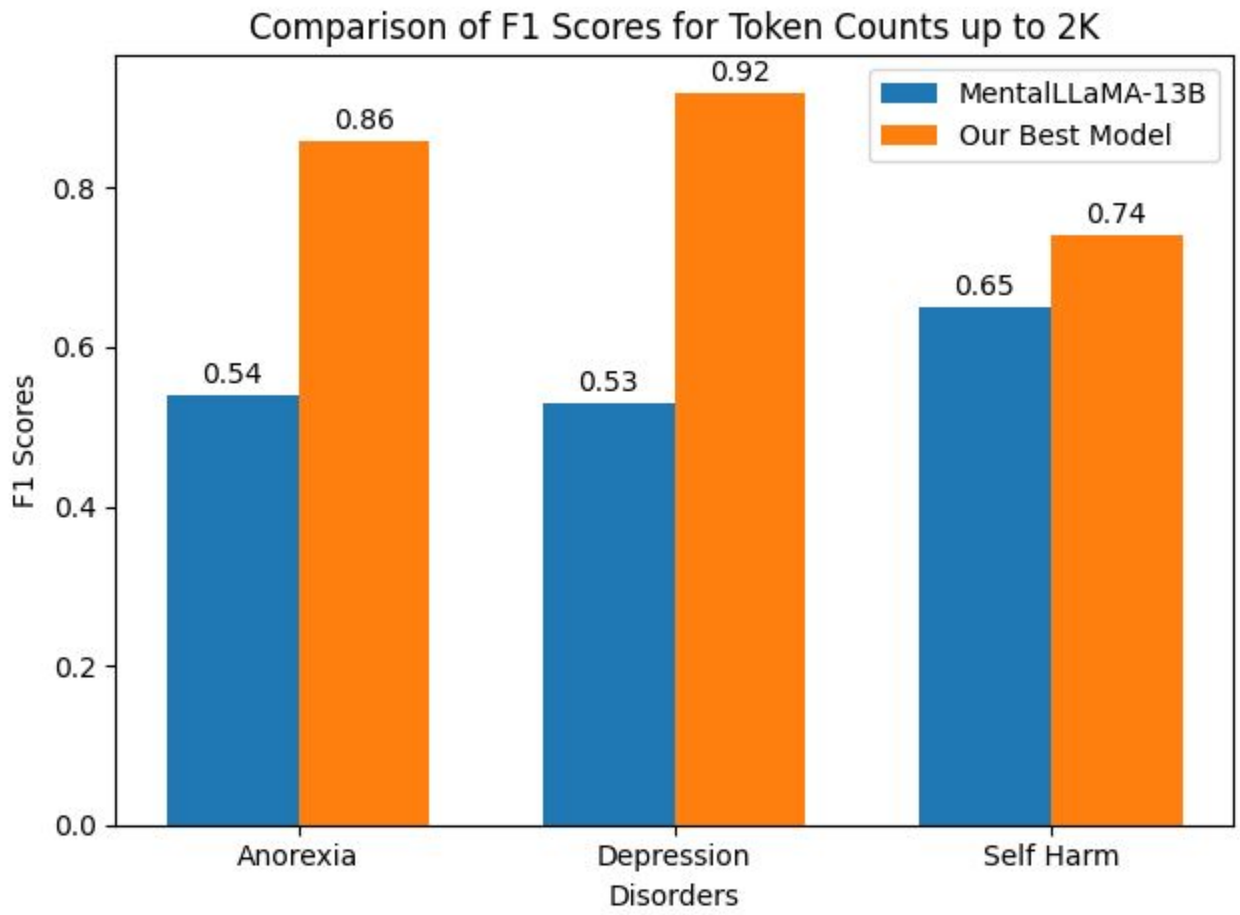


Findings:
We observe a significant dip in performance as compared to our original setup.

Results for temporal analysis: F1 scores comparison between the permuted input data and the ordered input data for three disorders

Full Context Analysis

Sub-optimal results aligns with recent studies like Liu et al. (2024), which demonstrates the inability of LLMs to utilize long context inputs



Data Distribution

Anorexia: 117
Depression: 15
Self-Harm: 177

F1 scores over the condition class in three eRisk tasks by considering up to 2k context length.

Cross Domain Study: Results

- Anorexia and Self-Harm show good cross-domain F1 scores, whereas the other pairs involving Depression show sub-optimal results as compared to SOTA.
- Overall, this indicates that linguistic cues essential for classifying one disorder may be present in others, hinting at the potential of leveraging data for other domains.

Model	Train+Val	Test	F1	P	R
Anorexia (A)					
DisorBERT	A	A	**0.83**	0.82	**0.85**
LSTM	A	A	0.79	0.84	0.74
LSTM	D	A	0.75	0.68	0.83
LSTM	SH	A	0.80	**0.85**	0.75
Depression (D)					
DisorBERT	D	D	0.69	0.56	**0.89**
LSTM	D	D	**0.75**	0.79	0.71
LSTM	A	D	0.63	**0.86**	0.50
LSTM	SH	D	0.63	0.73	0.56
Self-Harm (SH)					
DisorBERT	SH	SH	0.74	0.72	0.76
LSTM	SH	SH	**0.83**	**0.93**	0.75
LSTM	A	SH	0.78	0.85	0.72
LSTM	D	SH	0.69	0.65	**0.77**

Results for cross-domain evaluations for all six combinations of disorders. Here, ‘A’ is anorexia, ‘D’ is depression and ‘SH’ is self-harm.

Error Analysis: Out of context Inputs

Found in all three scenarios of Depression, Self-Harm and Anorexia.

- For example:
  - “'I did something really similar to this with some friends in Joshua tree, but it was odd because you could still see all of the stars.”
  - “opinion on these flip finz things? These caught my eye in an ad on a youtube video and it reminded me of when i used to flip.”

Error Analysis: Out-of-context Instance Example

Example of an out-of-context in social media posts. The word cloud shows the word distribution of a depressed person being predicted as non-depressed. The larger words in the cloud indicate higher word frequency, and notably, there are no words related to depression symptoms, contributing to the misclassification.

Error Analysis: Incomplete Context

- Found in all three scenarios of Depression, Self-Harm and Anorexia.
- People with mental disorders may have had some posts removed due to NSFW content.

Examples:

- **28 OCT 2020 23:52:27 -** Your post was removed for breaking [**rule 2g**]. Instead consider posting it in the...
- **2 MAY 2021 2:59:17 -** Your post was removed for breaking [**rule 2d**]. Instead consider posting it in the..
- **17 NOV 2018 12:25:25 -** Your post was removed for breaking [**rule 2g**]. Instead consider posting it in the..

NSFW: Not Safe For Work

Summary

- Proposed a novel framework for incorporating temporal representation of textual data to identify Anorexia, Depression, and Self-harm from social media content.
- Achieved superior performance compared to state-of-the-art Language Model (LM)-based baselines by integrating foundational deep-learning architecture while significantly reducing computational resource requirements.
- Our methodology utilizes fundamental deep-learning architecture and surpasses LLM-based baselines by accounting for temporality and the full context of the input data.
- Our cross-domain analysis highlights the overlapping linguistic cues among the disorders and hints at the possibility of leveraging data from different mental disorders.

Conclusion

- Despite the widespread use of large language models (LLMs), essential NLP tasks, like mental disorder identification, face substantial accuracy challenges, especially in low-resource settings.

- Fine-tuning more compact models remains crucial to achieve significantly better performance in such demanding tasks.

- A promising approach to enhance mental disorder identification accuracy will be leveraging the benefits of transfer learning by fine-tuning LLMs that are pre-trained using extensive domain-related datasets.

Future Work

- To explore a more diverse linguistic landscape on understudied and complex mental disorders such as schizophrenia, personality disorders, and bipolar disorder.
- To incorporate audio and visual signals alongside textual data to gain valuable insights into behavioral patterns exhibited by condition subjects.
- To explore federated learning to facilitate data aggregation and model training while respecting data ownership and privacy concerns.

Submission Under Review

Raja Kumar*, Kishan Maharaj*, Ashita Saxena, and Pushpak Bhattacharyya. 2024. **Mental Disorder Identification through Temporal Representation of Text**. *(Submitted to June ARR 2024).*

*Equal Contributions

References

Mario Aragon, Adrian Pastor Lopez Monroy, Luis Gonzalez, David E. Losada, and Manuel Montes. 2023. **DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media.** *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15305–15318,* Toronto, Canada. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. **MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare.** *In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 7184–7190*, Marseille, France. European Language Resources Association.

Low, D. M., Rumker, L., Torous, J., Cecchi, G., Ghosh, S. S., & Talkar, T. (2020). **Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study.** *Journal of medical Internet research, 22(10), e22635.*

Cohn, M. A., Mehl, M. R., Pennebaker, J. W., 2004. **Linguistic markers of psychological change surrounding september 11, 2001.** *Psychological science 15 (10), 687–693*

Simmons, R. A., Chambless, D. L., Gordon, P. C., 2008. **How do hostile and emotionally over involved relatives view relationships?: What relatives' pronoun use tells us**. *Family Process 47 (3), 405–419*

References

Rude, S., Gortner, E.-M., Pennebaker, J., 2004. **Language use of depressed and depression vulnerable college students.** *Cognition & Emotion 18 (8), 1121–1133*

Boals, A., Klein, K., 2005. **Word use in emotional narratives about failed romantic relationships and subsequent mental health**. *Journal of Language and Social Psychology 24 (3), 252–268*

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., Schwartz, H. A., 2018. **Facebook language predicts depression in medical records.** *Proceedings of the National Academy of Sciences 115 (44), 11203–11208.*

Pennebaker, J. W., Mehl, M. R., Niederhoffer, K. G., 2003. **Psychological aspects of natural language use: Our words, our selves**. *Annual review of psychology 54 (1), 547–577*

Kahn, J. H., Tobin, R. M., Massey, A. E., Anderson, J. A., 2007. **Measuring emotional expression with the linguistic inquiry and word count.** *The American journal of psychology 120 (2), 263–286*

Al-Mosaiwi, M., Johnstone, T., 2018. **In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation**. *Clinical Psychological Science 6 (4), 529–542*

References

Savekar, A., Tarai, S., Singh, M., 2019. **Linguistic markers in individuals with symptoms of depression in bi-multilingual context. In: Early Detection of Neurological Disorders Using Machine Learning Systems.** *IGI Global, pp. 216–240*

Simms, T., Ramstedt, C., Rich, M., Richards, M., Martinez, T., Giraud-Carrier, C., 2017. **Detecting cognitive distortions through machine learning text analytics.** *In: 2017 IEEE international conference on healthcare informatics (ICHI). IEEE, pp. 508–512*

Guntuku, S. C., Giorgi, S., Ungar, L., 2018. **Current and future psychological health prediction using language and socio-demographics of children for the clpysch 2018 shared task.** *In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. pp. 98–106*

Gaur, M., Aribandi, V., Alambo, A., Kursuncu, U., Thirunarayan, K., Beich, J., Pathak, J., Sheth, A., 2021. **Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-ssrs**. *PloS one 16 (5), e0250448*

Reece, A. G., Reagan, A. J., Lix, K. L., Dodds, P. S., Danforth, C. M., Langer, E. J., 2017. **Forecasting the onset and course of mental illness with twitter data**. *Scientific reports 7 (1), 13006*

Acknowledgement

I wish to record a deep sense of gratitude to **Prof. Pushpak Bhattacharyya** for his valuable guidance and constant support at all stages of my Dual Degree Project and related research

Thank You!