

Privacy Aware Question-Answering System for Online Mental Health Risk Assessment

Prateek Chhikara^{*†}, Ujjwal Pasupulety[†], John Marshall,
Dhiraj Chaurasia and Shweta Kumari

University of Southern California, Los Angeles, USA

{pchhikar, upasupul, jjmarsha, dchauras, shwetaku}@usc.edu

Abstract

Social media platforms have enabled individuals suffering from mental illnesses to share their lived experiences and find the online support necessary to cope. However, many users fail to receive genuine clinical support, thus exacerbating their symptoms. Screening users based on what they post online can aid providers in administering targeted healthcare and minimize false positives. Pre-trained Language Models (LMs) can assess users' social media data and classify them in terms of their mental health risk. We propose a Question-Answering (QA) approach to assess mental health risk using the Unified-QA model on two large mental health datasets. To protect user data, we extend Unified-QA by anonymizing the model training process using differential privacy. Our results demonstrate the effectiveness of modeling risk assessment as a QA task, specifically for mental health use cases. Furthermore, the model's performance decreases by less than 1% with the inclusion of differential privacy. The proposed system's performance is indicative of a promising research direction that will lead to the development of privacy-aware diagnostic systems.

1 Introduction

In recent years, Natural Language Processing (NLP) has emerged as a powerful field of study that focuses on the interaction between human language and computational systems (Singh et al., 2020). Mental health is a crucial aspect of overall well-being, and gaining insights into individuals' mental states has become an increasingly important area of study. NLP techniques have been useful in identifying text markers that indicate an individual's mental well-being (Zhang et al., 2022). Social media websites, such as Twitter and Reddit, provide a wealth of textual data that offers a unique

opportunity to analyze the mental health status of their users at scale, enabling the exploration of patterns, trends, and potential interventions (Skaik and Inkpen, 2020). Assessing users' mental health risk can be reduced to a basic text classification task, where the Transformer architecture (Vaswani et al., 2017) has demonstrated state-of-the-art performance. BERT (Devlin et al., 2019) encodings have been utilized for training a variety of mental health risk detection systems (Jiang et al., 2020; Nisa and Muhammad, 2021; Zeberga et al., 2022). BERT models fine-tuned on social media data (Ji et al., 2022; Murarka et al., 2020) are able to classify at-risk individuals with high accuracy.

However, advances in text classification models have stagnated with the advent of BERT encodings. Posing the risk assessment problem as a QA task is more analogous to consulting a trained clinician (Mutabazi et al., 2021). QA systems built using BERT have been used for public education on topics in mental health (Guo et al., 2021). Nearly 30% of QA healthcare systems focus on mental health applications such as workplace empowerment, screening, effecting behavior change, and reducing smoking/alcohol dependence (Cilar Budler et al., 2023). Multiple-choice QA models demonstrate a promising alternative approach to depression severity estimation even with low amounts of training data (Gabín et al., 2021). Further development of QA models could lead to better autonomous diagnostic systems. This work proposes the use of AllenAI's Unified-QA model (Khashabi et al., 2020) to assess the mental health risk of users from their social media posts. The research objective is to explore whether QA transformer models are better than text classification transformers at assessing the risk to mental health and modeling language markers that are indicative of specific mental illnesses. We compare Unified-QA to state-of-the-art pre-trained language models that perform text classification on the same data.

^{*} Corresponding author

[†] These authors contributed equally to this work

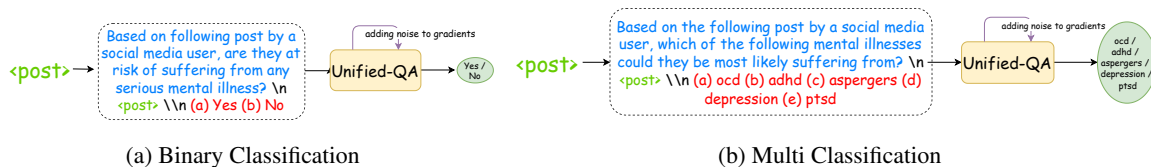


Figure 1: Proposed Pipeline.

Training models on sensitive user data in their raw form makes them non-compliant with data privacy rules which can have serious legal ramifications in the case of unexpected data breaches (Brown et al., 2022). By using Differential Privacy, language models can be trained such that they do not memorize the training data, leading to data security and better model generalization (Basu et al., 2021; Behnia et al., 2022). This work also studies the impact of differential private training on QA model performance. The contributions of the paper are as follows.

1. We approached the text classification task for mental health posts using a QA framework. Specifically, we transformed our input data according to the Unified-QA model and subsequently fine-tuned and evaluated it against one binary dataset and one multi-class dataset.
2. We add noise to the model gradients to prevent the QA model from memorizing the input text, thereby making it differentially private.
3. Our model’s performance is compared against various text classification models, including both traditional classifiers and pre-trained language models with varying degrees of scale, spanning from low to high-parameter models.

2 Proposed Methodology

The proposed work consists of fine-tuning the Unified-QA model on two large mental health datasets in a differentially private manner.

2.1 Mental Health Assessment as a QA task

Treating a text classification task as a seq2seq QA task involves converting input text-label pairs into a Question-Answer format. This approach can be beneficial for tasks such as sentiment analysis, topic classification, and other text classification tasks as it provides a more structured and interpretable output (Bakır and Aktas, 2022). Additionally, it allows for using pre-trained QA transformer models, which can improve performance without requiring large amounts of training data

(Han et al., 2021). In the proposed work, we have utilized Unified-QA, a single pre-trained QA model that performs exceptionally well across 17 diverse QA datasets with different formats. It demonstrates strong generalization even on unseen datasets, outperforming specialized models trained on individual datasets. Fine-tuning the Unified QA model leads to state-of-the-art performance on six datasets, making it a robust foundation for developing QA systems.

The Unified-QA model is prompted as shown in Figure 1, where each post-label pair from the training set is pre-processed by creating a task-specific question prompt to which the post is appended. For outputs, each training sample contains answer options whose values are equal to the set of class labels for each task. This approach enables a more nuanced analysis of mental health-related content, providing a clear rationale for the predicted mental health status. The input and output formatting is consistent for both training and testing.

2.2 Incorporating Differential Privacy

Mental health datasets are highly sensitive, containing personal and often confidential information about individuals’ private lives. Naïvely training language models on such data poses a high risk of input-memorization, allowing adversaries to reconstruct samples using the models and their weights (Fredrikson et al., 2015; Carlini et al., 2020). This work enhances the training process of the Unified-QA model by incorporating differential privacy (Dwork, 2006). By adding controlled noise to the gradients during training (Neelakantan et al., 2015; Chhikara et al., 2023), a balance between model accuracy and privacy preservation is achieved. Consequently, individuals’ data remains secure within the model, even if an adversary gains access to the model’s weights. The proposed approach makes a powerful general-purpose QA system suitable for use in applications that involve sensitive data.

Table 1: Samples from each dataset with their corresponding label.

RedditMH	SMK
<p>Positive: Fed up with my crazy mother. First she was just angry all the time, now she's angry and mad(schizophrenic) all the time, it's frustrating to not get a proper father in the house and add to that not even a stable mother, unstable is not even half the things she is. My god. She thinks my father is cooking up dark magic portions all the time. It's infuriating to have family like this, especially when half the things in my life doesn't seem to work properly or s**t always happens with me. honestly fed up with always trying to improve my life, exercising, meditating, everything under the sun, and they definitely have effect but it's not satisfactory and I always wonder what'd happen if I just didn't have such a mother, and actually have a father in the house.</p> <p>Negative: Is it possible to lose fat without losing weight? I used to workout all the time and was pretty muscular (80kg) when I was 17. I quit working out like about 3 months ago because I was working 14h everyday to sustain my s**t and had problems back then. I'm now 75ish kg and I'm slim but have high body fat. I found a pretty good job that is pretty good in terms of hours and finally have time to working out. I'll start rucking with 20kg some 5 miles starting Monday. I have gotten a pretty fat face, and some love handles. My question is, can I lose fat but still maintaining my weight?</p>	<p>ADHD: Please, please help me. The past 2-3 months I have been doing very well managing my symptoms and I have been overall, happy. About 2 weeks ago I noticed myself becoming angry just by my fiancé's presence... it makes me feel crazy as he is literally my favorite person. I have been SO irrationally angry with him about nothing. I can reason with myself, but can't help my anger. Not currently on any medications as none I have tried have worked. Been off for 3ish months. I'm also upset because Christmas is my FAVORITE and being p***y is ruining it for me. Please help me 😞😞😞😞😞</p> <p>Depression: Does anyone else still manage fully function society's eyes despite severely depressed? I fulltime job I'm never late for, I act happy people believe it, I go gym regularly, I try best maintain friendships even though they're seeming one-sided. To society, I'm fine. What see crying eyes whenever I drive own, I constantly think suicide self harm. I feel like I'm wasting life. When I get home work, I'll sit phone wait go sleep. I interest passions I feel trying push something I used love drains me. But society, I'm fine.</p>

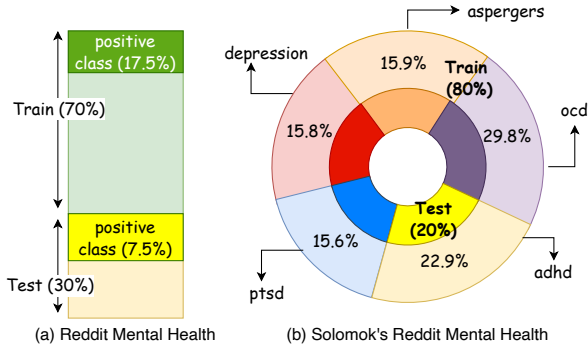


Figure 2: Dataset details.

2.3 Datasets

The following public domain datasets were used to fine-tune and evaluate the language models listed in Section 2.4. As the datasets are openly available for research purposes, no additional user consent was required to procure them. Table 1 shows a snapshot from each dataset and Figure 2 shows the train-test splits and class distributions.

Reddit Mental Health Dataset (RedditMH):

This dataset (Low et al., 2020) contains posts from 28 subreddits collected from 2018-2020, including 15 specific mental health support groups, two broad mental health subreddits, and 11 non-mental health subreddits. For the purpose of this study, instances from broad mental-health (45K) and non-mental health (137K) subreddits are chosen. This was done since LMs perform better at answering clinical questions when pre-trained on open-domain corpora (Soni and Roberts, 2020).

Solomonk's Reddit Mental Health (SMK): This dataset¹ is comprised of 151K Reddit posts exclusively from five mental health-related subreddits. After cleaning, approximately 88k (depression: 13.8K, OCD: 26.3K, Aspergers: 14.1K, PTSD: 13.7K, and ADHD: 20.3K) entries were considered for experimentation.

2.4 Models

The Unified-QA model is fine-tuned and evaluated using the aforementioned datasets. To assess its performance, we compare its results against the following model classes.

Baseline classifiers: The following traditional algorithms were used to define baselines which the LMs improve upon: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), AdaBoost, Stochastic Gradient Descent (SGD), Multilayer Perceptron (MLP), ExtRA trees (xTrees) & Multinomial Naive Bayes (MNb).

Low-parameter LMs: The following transformers (10-25M parameters) were evaluated: (1) *roberta-small*², a distilled version of RoBERTa that can be trained on a single GPU. (2) Rohanian et al. (2023a) developed a suite of lightweight transformers specifically for clinical NLP applications. Of these, *tiny-clinicalbert*, *clinical-mobilebert*,

¹https://huggingface.co/datasets/solomonk/reddit_mental_health_posts

²<https://huggingface.co/smallbenchnlp/roberta-small>

bio-mobilebert and *tiny-biobert* (Rohanian et al., 2023b) are evaluated.

High-parameter LMs: The following large transformers (>100M parameters) were evaluated on the datasets : (1) *PsychBERT* (Vajre et al., 2021), a model pre-trained on clinical and biomedical literature. (2) *MentalBERT* and *MentalRoBERTa*, state-of-the-art models which are pre-trained on data from mental health-related subreddits.

3 Experimental Setup

3.1 Model Training

The baseline classifiers were implemented by training vectorized inputs (TF-IDF, Count, Hash) using scikit-learn on consumer-grade PCs without any hyperparameter tuning or GPU acceleration. The language models were fine-tuned as text-classifiers using Nvidia A10G GPUs with 24 GB GDDR6 memory. They were fine-tuned for five epochs with varying batch sizes using the Hugging Face API with default hyperparameters. Since these models are pre-trained on relevant clinical or social media data, fine-tuning on the evaluation datasets for too many epochs hurts model performance.

The proposed QA models were trained for 20 epochs. The small and base models were trained with 128 and 64 batch sizes, respectively. The input text token length is set to 200. The *Adam* optimizer ($\text{lr}=10^{-3}$) with weight decay and a linear learning rate scheduler are used.

3.2 Differential Privacy

Fine-tuning a small portion of layers on a subset of the training data can enhance pre-trained model performance on specific tasks and reduce the risk of overfitting. Fine-tuning saves time and resources compared to training a model from scratch on the same dataset, as pre-trained models have already learned multiple language features that are pertinent. Noise, denoted by noise_{std} is sampled from the standard normal distribution ($\mu = 0, \sigma = 1$) and added during the training phase. The encoder and decoder of the Unified-QA model are frozen while the remaining layers are fine-tuned with 10% of the evaluation datasets. Clipping the gradients if their norm exceeds a specified maximum value prevents them from exploding during training, which mitigates numerical instability and improves convergence time (Zhang et al., 2020).

3.3 Evaluating Model Safety

In order to assess whether the proposed model effectively implements differential privacy, ϵ and δ values are computed. ϵ measures the strength of the privacy guarantee, and δ is a parameter that bounds the probability of failing to provide the privacy guarantee (Abadi et al., 2016). S is the sensitivity of the loss function with respect to individual training examples. The clip_{norm} parameter is the maximum L2-norm of the gradient before adding noise. The standard deviation of the maximum noise is computed using Equation 1.

$$\begin{aligned} \text{maxNoise}_{std} \leftarrow & \text{clip}_{norm} * \sqrt{\frac{2 * \text{total}_{\epsilon}}{n}} \\ & + S * \sqrt{\frac{2 * \log\left(\frac{1.25}{\delta}\right)}{\epsilon}} \end{aligned} \quad (1)$$

where total_{ϵ} is 2ϵ and n represents the number of instances used in the computation. If the standard deviation of the Gaussian noise added (noise_{std}) to gradients during training process is less than maxNoise_{std} , then the model is differential private.

4 Results and Analysis

4.1 Unified-QA model

Table 2 shows the metrics calculated with respect to the positive class (label-1/answer-yes) for the RedditMH (binary choice) dataset. Table 3 shows the weighted average results for the SMK (multiple choice) dataset. The Unified-QA model clearly outperforms the top three baselines and low parameter language models. It also performs competitively against state-of-the-art language models, being only slightly outperformed by PsychBERT. The Appendix contains the complete evaluation results for traditional classifiers.

One possible explanation for the superior performance of the QA approach is that it is better suited to the mental health-related text classification task, which requires a deeper understanding of the context and nuances of the text. By transforming the input text into a question, Unified-QA can better capture the semantic relationships and dependencies between the words in the text. Its ability to handle long and complex sentences lends more accurate mental health risk assessments.

Table 2: Comparison of models on RedditMH dataset.

Model	Precision	Recall	F1
MLP (TF-IDF)	87.939	88.831	88.382
MLP (Count)	87.541	87.412	87.476
LR (TF-IDF)	90.309	82.294	86.115
roberta-small	93.930	94.000	93.965
tiny-clinicalbert	94.521	93.742	94.130
tiny-biobert	94.624	94.095	94.359
clinical-mobilebert	94.036	92.522	93.273
bio-mobilebert	93.745	92.580	93.159
PsychBERT	95.911	95.389	95.649
MentalBERT	94.604	92.191	93.382
MentalRoBERTa	95.779	95.772	95.775
Unified-QA-small	94.947	95.564	95.258
Unified-QA-base	95.431	94.765	95.097

4.2 Differential Private model

Training the Unified-QA model using differential privacy makes it an ideal system for mental health applications, given the sensitive nature of the training data. The final model’s privacy is validated using Equation 1. The F1 score of the Unified-QA-small differential private model dropped by **0.47%** and **0.82%** (absolute values) on the RedditMH and SMK dataset, respectively. The additional complexity of a multi-class problem challenges the model to make a privacy-performance trade-off even further, which could explain the higher drop in F1. These findings suggest that while differential privacy techniques can help protect user privacy during model training, their impact on model performance is not insignificant. The promise of safeguarding user data while being able to triage at-risk users accurately offers a promising direction of research to reduce the gap in performance.

5 Conclusion & Future Work

This work aimed to explore the effectiveness of Question-Answering models for mental health risk assessment. The Unified-QA model performed $\sim 2\%$ better than low-parameter pre-trained language models and produced competitive results when compared to larger state-of-the-art language models. With differential privacy, the model only suffers a sub-one percent drop in performance. This study contributes to the growing body of research on mental health analysis using social media data. It highlights the potential of artificial intelligence techniques to improve our understanding of mental

Table 3: Comparison of models on SMK dataset.

Model	Precision	Recall	F1
LR (TF-IDF)	82.564	87.526	84.973
SGD (TF-IDF)	82.665	87.147	84.847
SGD (Count)	82.943	86.654	84.758
roberta-small	87.520	87.590	87.500
tiny-clinicalbert	88.770	88.790	88.740
tiny-biobert	88.076	87.834	87.917
clinical-mobilebert	86.965	87.062	86.935
bio-mobilebert	86.963	87.045	86.954
PsychBERT	90.45	90.47	90.42
MentalBERT	88.76	88.82	88.75
MentalRoBERTa	87.29	87.39	87.32
Unified-QA-small	89.455	89.525	89.419
Unified-QA-base	89.605	89.654	89.533

health and support the development of effective interventions.

Some limitations in the design of this study offer directions for further work to improve performance. Since the RedditMH dataset uses posts from general mental health subreddits, the binary-choice QA model was not exposed to finer-grained data from illness-specific subreddits, which could lead to better assessments. Also, the multiple-choice QA model forces an answer from the list of five diseases even if there are no distinct language markers of any disease. Future work will combine binary and multi-choice QA models to filter at-risk individuals. Two-stage QA models will serve as effective screening tools to ultimately provide care to individuals who need it the most.

Acknowledgements

We want to express our gratitude to *beam.cloud*³, an ML Infrastructure startup for generously providing the necessary compute resources to train and evaluate our models. We greatly appreciate their commitment to advancing the field of natural language processing and their willingness to support our research endeavors.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC con-*

³<https://www.beam.cloud>

- ference on computer and communications security*, pages 308–318.
- Dilan Bakır and Mehmet S. Aktas. 2022. [A systematic literature review of question answering: Research trends, datasets, methods](#). In *Computational Science and Its Applications – ICCSA 2022 Workshops: Malaga, Spain, July 4–7, 2022, Proceedings, Part I*, page 47–62, Berlin, Heidelberg. Springer-Verlag.
- Priya Basu, Tiasa Singha Roy, Rakshit Naidu, Zumur Muftuoglu, Sahib Singh, and Fatemehsadat Mireshghallah. 2021. Benchmarking differential privacy and federated learning for bert models. *ArXiv*, abs/2106.13973.
- Rouzbeh Behnia, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan. 2022. [EW-tune: A framework for privately fine-tuning large language models with differential privacy](#). In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE.
- Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. [What does it mean for a language model to preserve privacy?](#) In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 2280–2292, New York, NY, USA. Association for Computing Machinery.
- Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. In *USENIX Security Symposium*.
- Prateek Chhikara, Rajkumar Tekchandani, and Neeraj Kumar. 2023. Adaptive federated learning scheme for recognition of malicious attacks in an iot network. *Computing*, pages 1–16.
- Leona Cilar Budler, Lucija Gosak, and Gregor Stiglic. 2023. [Review of artificial intelligence-based question-answering systems in healthcare](#). *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.
- Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333.
- Jorge Gabín, Anxo Pérez, and Javier Parapar. 2021. [Multiple-choice question answering models for automatic depression severity estimation](#). *Engineering Proceedings*, 7(1).
- Chaohui Guo, Shaofu Lin, Zhisheng Huang, and Yaohong Yao. 2021. [Mental health question and answering system based on bert model and knowledge graph technology](#). In *ISAIMS 2021*, ACM International Conference Proceeding Series, pages 472–476. Association for Computing Machinery. Publisher Copyright: © 2021 ACM.; 2nd International Symposium on Artificial Intelligence for Medicine Sciences, ISAIMS 2021 ; Conference date: 29-10-2021 Through 31-10-2021.
- Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. [Pre-trained models: Past, present and future](#). *AI Open*, 2:225–250.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. [Mental-BERT: 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.
- Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. [Detection of mental health from Reddit via deep contextualized representations](#). In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156, Online. Association for Computational Linguistics.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hananeh Hajishirzi. 2020. [UNIFIEDQA: Crossing format boundaries with a single QA system](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.
- Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S Ghosh. 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.
- Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2020. Classification of mental illnesses on social media using roberta. In *International Workshop on Health Text Mining and Information Analysis*.
- Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. 2021. [A review on medical textual](#)

- question answering systems based on deep learning approaches. *Applied Sciences*, 11(12):5456.
- Arvind Neelakantan, Luke Vilnis, Quoc Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. 2015. Adding gradient noise improves learning for very deep networks.
- Qamar Un Nisa and Rafi Muhammad. 2021. Towards transfer learning using bert for early detection of self-harm of social media users. In *Conference and Labs of the Evaluation Forum*.
- Omid Rohanian, Mohammadmahdi Nouriborji, Hannah Jauncey, Samaneh Kouchaki, ISARIC Clinical Characterisation Group, Lei Clifton, Laura Merson, and David A. Clifton. 2023a. [Lightweight transformers for clinical natural language processing](#).
- Omid Rohanian, Mohammadmahdi Nouriborji, Samaneh Kouchaki, and David A Clifton. 2023b. On the effectiveness of compact biomedical transformers. *Bioinformatics*, 39(3):btad103.
- Prabhjot Singh, Prateek Chhikara, and Jasmeet Singh. 2020. An ensemble approach for extractive text summarization. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–7. IEEE.
- Ruba Skaik and Diana Inkpen. 2020. Using social media for mental health surveillance. *ACM Computing Surveys (CSUR)*, 53:1 – 31.
- Sarvesh Soni and Kirk Roberts. 2020. [Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5532–5538, Marseille, France. European Language Resources Association.
- V. Vajre, M. Naylor, U. Kamath, and A. 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, Los Alamitos, CA, USA. IEEE Computer Society.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Kamil Zeberga, Muhammad Attique, Babar Shah, Farman Ali, Yalew Zelalem Jembre, and Tae-Sun Chung. 2022. [A novel text mining approach for mental health prediction using bi-lstm and bert model](#). *Computational Intelligence and Neuroscience*, 2022:7893775.
- Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jabbaie. 2020. [Why gradient clipping accelerates training: A theoretical justification for adaptivity](#). In *International Conference on Learning Representations*.
- T. Zhang, A Schoene, and S. Ananiadou. 2022. [Natural language processing applied to mental illness detection: A narrative review](#). *n p j Digital Medicine*, 5:46.

Appendix

Comparison on Traditional ML Classifiers

Table 4: Reddit MH: Precision, recall, and F1-score for 9 models on 3 vectorization techniques. x - did not converge within a reasonable time period

Model	Count Vectorizer			TF-IDF Vectorizer			Hashing Vectorizer		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Logistic Regression	91.05	78.993	84.594	90.309	82.294	86.115	87.337	78.147	82.487
Decision Tree	74.179	75.897	75.027	74.95	76.471	75.702	69.265	71.021	70.133
Random Forest	91.875	77.662	84.173	91.229	78.853	84.591	92.841	69.896	79.752
Ada Boost	88.097	69.662	77.802	88.179	72.235	79.415	83.667	69.006	75.634
Gradient Boosting	90.417	68.618	78.023	90.436	70.919	79.497	88.841	67.088	76.447
Extra Trees	92.768	74.985	82.934	93.417	74.713	83.025	96.423	58.087	72.5
Multinomial NB	72.013	86.765	78.703	93.003	64.22	75.978	x	x	x
SGD Classifier	91.476	79.463	85.048	90.592	80.506	85.253	88.352	74.684	80.945
Multi-Layer Perceptron	87.541	87.412	87.476	87.939	88.831	88.382	x	x	x

Table 5: SMK: Precision, recall, and F1-score for 9 models on 3 vectorization techniques. x - did not converge within a reasonable time period

Model	TF-IDF Vectorizer			Count Vectorizer			Hashing Vectorizer		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Logistic Regression	82.564	87.526	84.973	80.325	89.005	84.442	82.178	85.403	83.759
SVC	x	x	x	75.667	90.313	82.344	x	x	x
Decision Tree	74.832	78.237	76.497	73.74	77.934	75.779	72.061	75.64	73.807
Random Forest	76.999	88.91	82.527	77.838	87.223	82.264	74.641	89.668	81.467
Ada Boost	82.766	62.275	71.073	80.932	67.829	73.804	77.396	69.194	73.066
Gradient Boosting	83.677	81.63	82.64	79.119	86.123	82.473	79.252	85.953	82.466
Extra Trees	73.5	91.014	81.325	73.688	90.256	81.135	66.807	93.213	77.831
SGD Classifier	82.665	87.147	84.847	82.943	86.654	84.758	82.994	84.284	83.634
Multi-Layer Perceptron	81.376	86.806	84.003	79.078	86.483	82.615	80.964	84.095	82.5