A Comparative Study of Deep Learning Architectures for Long Text Classification in Mental Health

Stanford CS224N Custom Project*

Junwei (Ivy) Sun
Department of Statistics
Stanford University
ivysun14@stanford.edu

Siqi Ma
Department of Statistics
Stanford University
siqima@stanford.edu

Yiran Fan
Department of Statistics
Stanford University
yiranf@stanford.edu

Abstract

Document-level classification has long been a target task in the clinical domain; however, there currently exists a gap in handling long conversational transcripts. This project aims to evaluate deep learning systems' efficacy in classifying psychological symptoms from such transcripts with excessive token counts, which is crucial for determining treatments in psychotherapy practices. We fine-tuned various transformer architectures (BERT, RoBERTa, Longformer), as well as trained Support Vector Machine (SVM) models with feature engineering, and performed human annotations using transcripts of token size up to $20,000$. We found that deep learning methods outperform both traditional machine learning method and human baseline while boosted methods seem to have a negligible effect on enhancing psychological symptom prediction outcomes.

1 Introduction

Identifying and classifying psychological symptoms serve as the important first step in psychological interventions. As the prevalence of mental illness continues to rise, there is a corresponding increase in publications focusing on the detection of mental health issues using machine learning algorithms. However, within the domain of psychology, natural language processing (NLP) techniques have predominantly relied on traditional machine learning methods. According to a study conducted by Zhang et al. (2022), 59% of the surveyed mental illness detection papers employed a machine learning pipeline involving Support Vector Machines (SVM), decision trees, AdaBoost, logistic regression, and naive Bayes. Despite the existence of NLP deep learning research in the field, they often utilize short or low-quality web text corpora (e.g., Reddit posts, and tweets) for training purposes (Ji et al., 2021; Salmi et al., 2022). Whether such web text data can address the complexity and subjectivity of mental health diagnostics raises questions about the practical value of proposed ML systems. Moreover, contemporary data consortia in mental health-related contexts frequently involve lengthy clinical textual data, such as electronic health records and conversational transcripts. Effectively harnessing these datasets presents a challenge to existing deep learning methods due to constraints on token size during the training and fine-tuning process.

Our objective for this study is to construct domain knowledge-enriched multi-label classification models for detecting mental health issues and compare their performances in classifying long texts. The overarching strategy involves employing different neural methods capable of processing unstructured English language as input and forecasting the potential psychological and mental health concerns experienced by individuals articulating these statements. Our emphasis will be on two prominent mental health indicators, specifically depression and anxiety, as foundational components for model development. We constructed neural classifiers of various architectures and employed both truncating longer documents to fit models' maximum sequence length and sub-document splitting combined with boosting as two approaches to harness information from our long inputs. Building off

*Mentor: Rohan Taori; external collaborator: Peter Washington, Ph.D., University of Hawaii at Mānoa of this work, we aim to expand the scope with the refined model to incorporate a broader spectrum of psychological states and transition toward a more comprehensive classification framework in the future.

2 Related Work

Reviews of papers in the mental health domain suggest that autoencoders are often used to achieve the classification task (Su et al., 2020). Existing literature often applies such deep learning frameworks on shorter and more accessible social media posts and found processing longer documents challenging (Li et al., 2022). The field of NLP has been actively seeking ways to capture dependency structures and contextual information within a lengthy text corpus. In tasks such as document-level classification, inputs vary from a few tokens to thousands or ten thousand, easily surpassing the popular frameworks' limit. Recent research endeavors have focused on addressing this challenge of learning with long text by exploring various strategies, including text segmentation and truncation, sliding window techniques, and modifications to the training or fine-tuning framework through the incorporation of additional layers, such as convolutional neural networks (Fiok et al., 2021; Park et al., 2022; Zheng et al., 2023). These modifications aim to accommodate pre-trained models for effectively handling long texts. Our objective is to bridge this gap by proposing and benchmarking technical solutions that facilitate the efficient use of long clinical texts within the realm of psychology. Through our work, we primarily aspire to provide insights into the efficient application of deep learning methodologies for the described challenge, while also contributing to the development of a robust mental health detection model that can be applied to various downstream applications, such as therapeutic chatbots.

3 Approach

Baseline SVM We utilized radial basis kernel Support Vector Machines (RBF SVMs) with feature engineering as the traditional machine learning (ML) baseline. The feature matrix consists of normalized stemming Bag-of-Words (BoW) and mapped features based on linguistic dictionaries, including the average concreteness score of each sentence and the average value of eight basic emotions and sentiments (Brysbaert et al., 2014; Mohammad and Turney, 2013).

Fine-tuning BERT, RoBERTa, Longformer We implemented a truncation approach wherein either the first 512 or 4096 tokens per sample were utilized to fine-tune multi-label classification tasks using BERT, RoBERTa, and Longformer architectures (Beltagy et al., 2020; Devlin et al., 2018; Liu et al., 2019). Specifically, we chose these pre-trained models to compare with BERT, the classic transformer-based architecture, because RoBERTa is pre-trained more progressively with a byte-level tokenizer that potentially could capture more vocabulary for our data, whereas Longformer employs a mixture of global attention and diluted sliding window attention that scales linearly with sequence length, thus more suitable for long text documents. Classification head with linear, dropout, and tanh layers are attached to the end of the hidden state to obtain logit outputs. Detailed classifier head architectures are specified in Figure 1. The code was written using the PyTorch framework with all pre-trained models loaded from Huggingface `transformers` (Paszke et al., 2019; Wolf et al., 2020).

Boosted Fine-tuning Sub-document slicing and pooling are performed on top of the truncation models to assess if better predictions can be achieved with boosting methodology (Li et al., 2022; Zheng et al., 2023). The procedure consists of each sample being sliced into sub-samples of 512 or 4096 tokens for fine-tuning. Classification results on sub-samples of a single example are pooled at the end and a majority vote is performed to determine the final classification outcome for a sample.

Human Baseline To assess the ceiling performance for this particular task, we evaluated human performance by randomly selecting 100 examples from the dataset, with subsequent annotation to classify whether conversation participants exhibited symptoms indicative of anxiety or depression. Given that the models were fine-tuned using psychotherapy session transcripts, their learning was constrained to a finite set of psychiatry-specific medical knowledge imparted within conversational contexts, such as medications and typical symptoms associated with various mental health states. Thus, our human annotation process was designed to mirror this learning paradigm without having strong psychiatry domain-specific knowledge to ensure comparability in performance metrics.

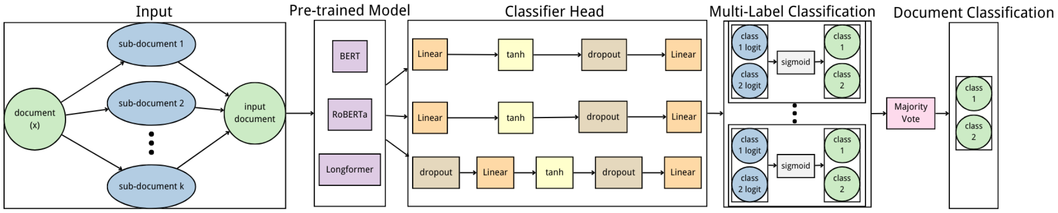


Figure 1: Pipeline for fine-tuning, with sub-document slicing and pooling.

In-Context Learning in LLM Large language models (LLM) like GPT-4 family models are capable of learning in-conversation. We constructed a mood classifier in GPT Store using labeled transcripts as the knowledge base and GPT-API with examples, constrained the output to classes of interest (anxiety, depression, none, or both), and assessed GPT-4's ability to classify mental health labels given the knowledge base samples.

4 Experiments

4.1 Data

This study utilizes the *Alexander Street Press: Counseling and Psychotherapy Transcripts* acquired from the Stanford Library (McNally et al., 2014). The dataset comprises plain-text transcripts of therapy sessions addressing a diverse array of presenting mental health issues with various therapeutic approaches. It is curated under ethical guidelines to ensure anonymity and responsible usage.

After data acquisition, patient texts were distinguished from therapist texts within each transcript. Non-ASCII characters were systematically eliminated, and transcripts with non-UTF encodings were excluded from the analysis. Descriptive elements such as "chuckles" and "laughter" were removed from the dataset. We also removed explicit mentions of the symptom word (i.e., "anxiety," "depression") from the transcript to avoid directly inputting classification labels during training, thus promoting a more robust and unbiased learning process. A verification procedure was conducted on the token size distributions for the examples using the model's pre-trained tokenizers to ensure fair comparison across models. This analysis revealed similar token size distributions, as illustrated in Figure 4 in the appendix. The average token size was approximately 3,800 per sample. Additionally, timestamps were stripped from the transcripts and cataloged separately. Metadata relevant to our analysis, including session identifiers, issue categorization, and therapeutic symptoms, were organized with the processed textual data into JSON format.

Following the preprocessing steps, the dataset consisted of 3,503 session records. Subsequently, we divided the 3,503 psychotherapy sessions into an 80% training set and a 20% evaluation set for evaluation. The training set comprises 2,802 examples, while the evaluation set comprises 701 examples. The training set was exclusively designated for fine-tuning, whereas the evaluation set was reserved for performance assessments. Because neural network-based models usually do not overfit the evaluation set and given that our main objective is to compare across various methods, we did not have a test set result to report (Recht et al., 2018).

4.2 Evaluation method

In our analyses, we employed three evaluation metrics across all models: accuracy score, F1 score, and ROC AUC score. Accuracy is defined as

$$\frac{\textit{number of correct predictions}}{\textit{total number of samples}}$$

F1 is defined as

$$\frac{2 \cdot Precision \times Recall}{Precision + Recall}$$

where precision $= TP/(TP + FP)$ and recall $= TP/(TP + FN)$. The ROC AUC score is a summary statistic quantifying the area under the receiver operating characteristic curve, offering insights into the model's discrimination ability across various threshold settings. These metrics were selected to comprehensively assess each model's proficiency in correctly identifying both the presence and absence of the symptoms. F1 score and ROC AUC score are computed with different averaging methods, including 1) globally counting the total TP, FN, FP, 2) calculating metrics for each label and finding their weighted average by the number of true instances for each label, and 3) calculating the metrics for each sample and finding the unweighted averages. Since the classes are imbalanced in our data where both labels have more non-symptomatic samples than symptomatic samples, the F1 and ROC AUC scores that weight each label using its support are more representative of the model performance. The reported metrics are computed using the weighted averaging method. All evaluation metrics are implemented with `sklearn.metrics` library (Buitinck et al., 2013).

4.3 Experimental details

We fine-tuned all models with multi-label tasks and utilized binary cross entropy loss with sigmoid layer `torch.nn.BCEWithLogitsLoss` as the loss function. We used AdamW optimizer and cosine scheduler for the learning rate with all fine-tuning processes. Other aspects of model configurations are listed below. We tuned a range of learning rates (1e-2 to 1e-6) and training time (up until around 15 epochs) for all truncation and boosted models. Performance metrics across the training time for the best learning rate are provided in the result section in Figure 2.

Pre-trained model	Max sequence length	Sub-document sample size	Batch size
bert-base-uncased	512	23216	16
roberta-base	512	27607	16
allenai/longformer-base-4096	4096	5044	16*

Table 1: Model configuration details.

* Trained with gradient accumulation. Batch size for boosted model is 32.

4.4 Results

We present the detailed performance of our best models below. Firstly, we observed that deep learning approaches outperform the traditional ML approach, as evidenced by performance metrics such as accuracy, F1 score, and AUROC. These slight improvements in these metrics are expected, as neural networks convert words into word embeddings, enabling them to better capture word-level patterns. Additionally, attention mechanisms allow the model to capture a richer input context compared to the BoW methods employed in the RBF SVM models. Secondly, compared to the human baseline, our models slightly outperform the accuracy and F1 score while the best model outperforms the human baseline ROC AUC score by around 0.1. This performance was expected as we found it challenging to distinguish the psychological state with conversational text when we were conducting human baseline evaluations.

Model	Learning rate	Epoch	Accuracy	F1	AUROC
Human baseline	-	-	0.490	0.529	0.656
SVM (Depression)*	-	-	0.818	0.349	0.613
SVM (Anxiety)*	-	-	0.676	0.528	0.638
BERT	5e-5	5	0.503	0.508	0.694
Boosted BERT	5e-5	2	0.530	0.351	0.686
RoBERTa	1e-5	7	0.561	0.517	0.713
Boosted RoBERTa	1e-5	4	**0.566**	0.542	**0.756**
Longformer	1e-5	8	0.549	**0.565**	0.681
Boosted Longformer	2e-5	7	0.514	0.319	0.568

Table 2: Performance of best fine-tuned models compared to SVM and human baselines.

* Metrics reported on binary classification tasks for each psychological symptom.

Overall, we observed comparable or marginally diminished performance of boosting when compared to the truncation baseline in both BERT and Longformer architectures. However, we noted a converse trend for the RoBERTa model specifically. Notably, across all models, boosted RoBERTa exhibited superior performance, achieving the highest scores in two out of three performance metrics.

When compared to the model's truncation baseline, we observed an increase in AUROC score of approximately 0.04, alongside an increase in F1 of about 0.03. Interestingly, we did not observe superior performance in Longformer across 3 truncation baselines, despite the fact that the average token size for our documents is around 3800 so a portion of the documents can be consumed completely in one input by the Longformer model.

To delve deeper into the factors influencing the degraded performance of boosted models in BERT and Longformer, we conducted supplementary fine-tuning experiments on BERT due to its relatively shorter training time. Our aim was to investigate whether the diminished performance stemmed from the incorporation of more data and potentially some noise from all sections of the document. To assess this, we randomly selected an adjacent section of 512 tokens from each sample and conducted fine-tuning using these inputs. Additionally, in the supplementary boosted BERT experiment, we altered the pooling mechanism such that if any sub-document was labeled as true, the entire document was classified as true (OR construction). The results of these experiments are detailed in Figure 3. The performances were similar across the four variants of BERT models.

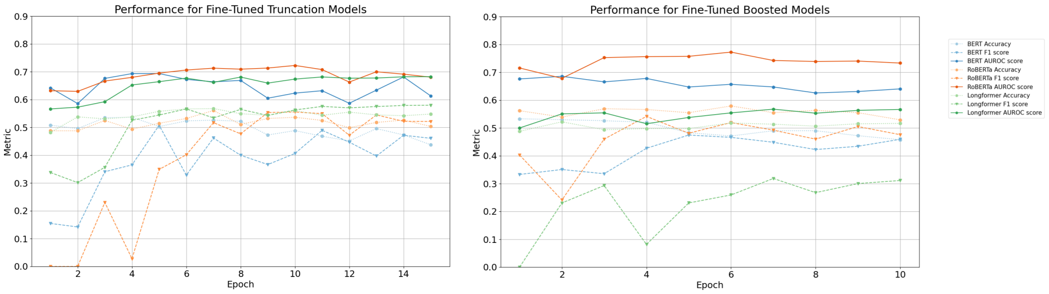


Figure 2: Performance metrics over epoch for truncation and boosted models.

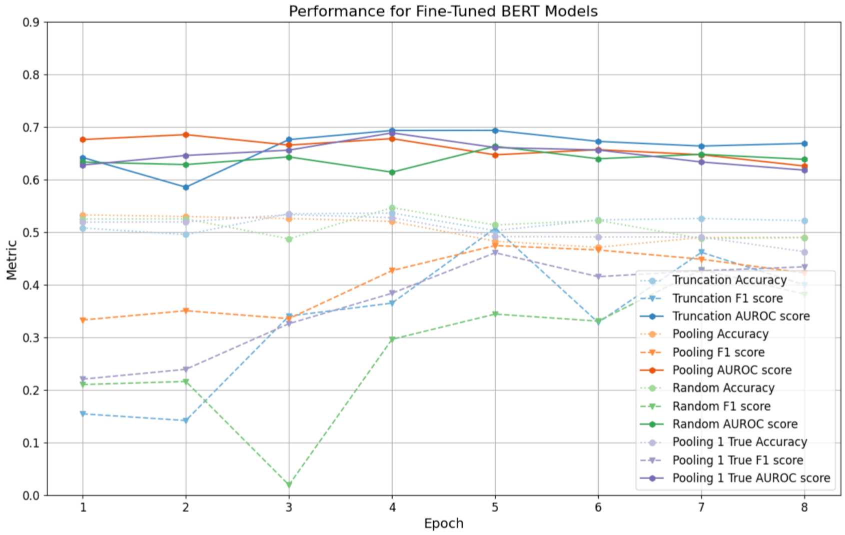


Figure 3: Performance metrics over epoch for BERT-based methods.

5 Analysis

Traditional ML vs. Deep Learning Our results suggest that the performance disparities between traditional ML and deep learning approaches are not substantial, indicating that traditional ML models with meticulous feature engineering may still achieve comparable results to large neural networks, particularly when dealing with input texts of excessive token size in mental health. However, the characteristics of lengthy texts potentially render the classification task challenging for neural network-based models, as these models must infer a close-to-binary answer from a vast amount of information that may contain noise. Additionally, the label provided by experts may also contain noise, especially for problems in psychology and mental health, which adds additional challenges. Existing works suggest that noisy datasets need additional handling, such as trailing the loss function, applying a probabilistic view, or employing some robust architecture (Song et al., 2020; Zhao and Gomes, 2021).

Human Baseline vs. Deep Learning Establishing a comparable, non-psychiatric knowledge-enriched human performance baseline enables us to comprehend the inherent difficulty of this task when learning exclusively from language input. The observation that deep learning models marginally surpass non-expert human annotations might imply that these sophisticated architectures possess the capability to glean more information and discern patterns from the data compared to non-experts in the context of psychological symptom classifications. However, it is essential to note that human annotations are conducted on a subset of randomly selected samples due to labor and time constraints. Thus, while indicative of ceiling performance, we do not treat human annotations as a strict benchmark in this analysis.

Truncation vs. Boosting In both BERT and Longformer models, slicing the documents into smaller chunks and aggregating prediction outcomes through majority voting did not yield the anticipated performance improvement. For the Longformer model, we examined the distribution of token sizes as depicted in Figure 4 and hypothesized that the pooling approach led to suboptimal performance possibly due to conflicting information in the majority voting process. When documents were partitioned into segments of a maximum length of 4096 tokens, they typically bifurcated into two subdocuments, potentially resulting in an even split between positive and negative predictions, thereby undermining the efficacy of the majority voting mechanism. Further investigation into the BERT model involved randomly selecting contiguous sections of 512 tokens from each sample and modifying the pooling mechanism with OR construction. These experiments yield similar results and hint at the possibility that information is uniformly scattered throughout the transcript.

On the contrary, a boosted RoBERTa model exhibited enhancements over the baseline RoBERTa metrics. We formulated several hypotheses to explain this phenomenon based on insights from our inputs and token size distributions. First, RoBERTa generated the highest number of sub-samples following document segmentation, potentially leading to the fragmentation of key information into smaller segments compared to the procedures applied in boosted BERT and Longformer models. Leveraging a byte-pair encoder, RoBERTa is adept at capturing misinformation arising from typos and abbreviations with greater comprehensiveness. The combination of these two advantages may partially account for the observed performance improvement in the boosted RoBERTa model. Nevertheless, when rounded to two decimal points, all metrics demonstrate minimal fluctuations, suggesting that overall, our three truncation and three boosted models exhibit comparable performance in long text classification tasks.

Our result is consistent with previous literature by Kamran et al. (2023) and Tanzia Parvin and Hoque (2021) on emotion classification that shows limited improvement of ensemble methods on deep neural network classifiers. Simple ensemble method such as majority vote with deep neural networks only helps to digest longer text without providing a significant improvement in model performance.

Evaluations While we conducted evaluation metrics using three different methods of averaging, we decided to compare and report using the weighted F1 score and ROC AUC score, which calculate metrics for each label and derive their weighted average based on the number of true instances for each label. Although other averaging methods yielded higher scores in certain models, we deemed weighted calculation to best reflect performance, as it adjusts the label-wise metric according to the prevalence of true labels in each category. Given the imbalanced classes in our dataset, where both labels have more than half of non-symptomatic samples, weighted F1 and ROC AUC scores offer a more objective assessment of performance.

Moreover, we did not select the best model based on evaluation loss due to its relatively minor fluctuation magnitude. During the fine-tuning of both truncation and pooling versions of BERT and Longformer, we observed a slight overfitting of loss despite continuous improvement in metrics. This observation provides another rationale for considering the pooled RoBERTa model as the optimal choice in our experiment. We hypothesized that this counterintuitive pattern between loss and evaluation metrics could stem from the fact that loss is a representation of logits before the sigmoid layer, which may fluctuate and influence the averaged loss over epochs despite consistent output labels.

In existing research studies, fine-tuned Clinical-Longformer from Li et al. (2022), which is tailored to long document classification tasks, achieved an F1-score of 0.484 and an AUROC score of 0.762 in predicting acute kidney injury using electronic health records. Our boosted RoBERTa model achieved a similar performance in our experiment. While we utilize this as a soft benchmark, our methodology differs due to the notably longer size of our dataset. Additionally, we acknowledge the discrepancy in data sources, with their emphasis on electronic health records contrasting with our utilization of conversational clinical transcripts. Therefore, we remain aware of the potential transferability of the method across varied clinical data types for classification tasks.

In-Context Learning The last piece of related experiments we conducted involved using both OpenAI's API interface and the GPT store to perform in-context learning. Beginning with naive prompting, we provided two positive and two negative samples from the prompt window. In the GPT store, we furnished examples by presenting formatted queries as a knowledge base. Both models exhibited inconsistency in the classification result with the exact same input in a new prompt. Even after employing a pooling strategy for 10 prompts of the same input, the overall accuracy remained similar to a random guess. This suggests that current user-facing language models may not be capable of this task, indicating the necessity for a more domain-focused NLP model for this specific task.

6 Conclusion

In summary, we found that fine-tuned pre-trained neural networks outperform traditional machine learning models and naive human baseline in classifying long, conversational therapeutic transcripts into mental state labels. Boosting through a majority vote or an OR construction does not improve model performance significantly. Among the three neural network models, RoBERTa appears to exhibit the best performance, likely due to its aggressive pre-training and tokenization scheme. Longformer, which takes in longer context length, does not outperform RoBERTa as we expected, suggesting the possibility that model input size itself is not a dominating factor in determining performance. In boosting, percentage overlap among sub-documents is considered as a hyper-parameter, but we observed no impact of overlap between sub-samples on model performance. We further fine-tuned the BERT model on a random segment of adjacent 512 tokens of each document, and the results suggest no difference compared to the truncation baseline using the first 512 tokens. All these experiments suggest that information in a therapy session may be uniformly scattered throughout the conversation. Lastly, we experimented with GPT-4 family models and observed that although in-context learning allows LLM to access labeled therapy transcripts as its knowledge base and make predictions based on such knowledge, they are highly unstable in predicting mental labels, and their results lack reproducibility.

To further advance the findings of this study, several promising directions for future research can be explored. Firstly, employing cutting-edge models such as Mistral with QLoRA (Dettmers et al., 2023; Jiang et al., 2023) can enhance our understanding of the relationship between input length and prediction performance. By utilizing these advanced architectures to construct a multi-label classifier, we can assess whether longer inputs correlate positively with improved prediction accuracy. This investigation holds significant promise in shedding light on the complexities inherent in mental health classification tasks, which often challenge even human judgment. Additionally, it can elucidate whether augmenting the model architecture alone is adequate for achieving more accurate document-level predictions. A separate, expert human baseline can also be established for our dataset, as suggested by a similar approach in Van Veen et al. (2024), to investigate the current performance cap in human annotations.

Expanding our framework to encompass a broader spectrum of mental health labels presents another compelling direction for future research. The original dataset comprises a diverse array of over 60 mental health indicators, ranging from suicidal intent and sleep disturbances to hallucinations and mania, among others. By leveraging our existing framework, we aim to develop a more robust, symptom-rich multi-label classification system for psychological states. Furthermore, as psychological issues often exhibit interdependencies, incorporating a wider range of labels holds the potential to yield superior predictive performance and deepen our understanding of the intricate relationships between various mental health manifestations.

Team member contributions
Ivy: RoBERTa fine-tuning, GPT-prompting, human annotation, writeup
Siqi: BERT fine-tuning, GPT store, human annotation, writeup
Yiran: Longformer fine-tuning, figures, human annotation, writup

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Krzysztof Fiok, Waldemar Karwowski, Edgar Gutierrez-Franco, Mohammad Reza Davahli, Maciej Wilamowski, Tareq Ahram, Awad Al-Juaid, and Jozef Zurada. 2021. Text guide: improving the quality of long text classification by a text selection method based on feature importance. *IEEE Access*, 9:105439–105450.

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*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Sara Kamran, Raziyeh Zall, Saeid Hosseini, MohamadReza Kangavari, Sana Rahmani, and Wen Hua. 2023. Emodnn: understanding emotions from short texts through a deep neural network ensemble. *Neural Computing and Applications*, 35.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

A McNally et al. 2014. Counseling and psychotherapy transcripts, volume ii.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. *arXiv preprint arXiv:2203.11258*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2018. Do cifar-10 classifiers generalize to cifar-10?

Salim Salmi, Saskia Mérelle, Renske Gilissen, Rob van der Mei, and Sandjai Bhulai. 2022. Detecting changes in help seeker conversations on a suicide prevention helpline during the covid- 19 pandemic: in-depth analysis using encoder representations from transformers. *BMC public health*, 22(1):530.

Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. 2020. Learning from noisy labels with deep neural networks: A survey. *CoRR*, abs/2007.08199.

Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. 2020. Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10(1).

Omar Sharif Tanzia Parvin and Mohammed Moshiul Hoque. 2021. Multi-class textual emotion categorization using ensemble of convolutional and recurrent neural network. *SN Computer Science*, 3.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, pages 1–9.

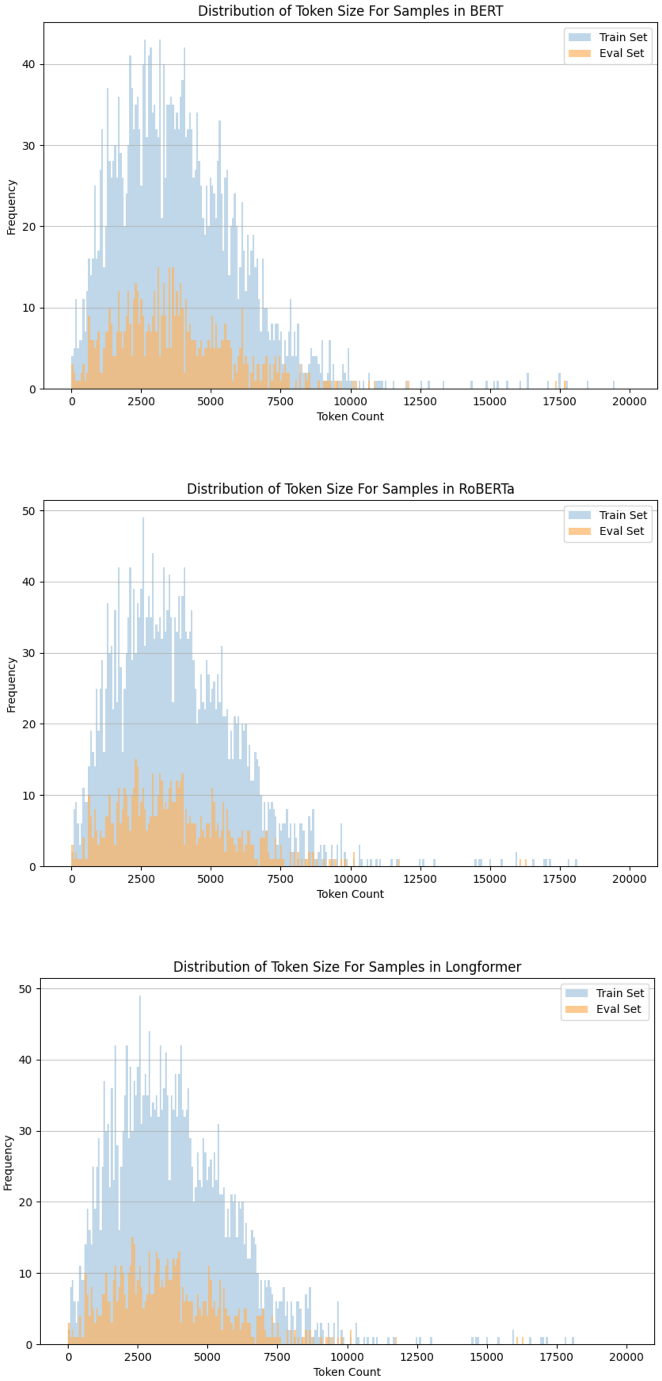
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.

Wenting Zhao and Carla P. Gomes. 2021. Evaluating multi-label classifiers with noisy labels. *CoRR*, abs/2102.08427.

Yuan Zheng, Rihui Cai, Maihemuti Maimaiti, and Kahaerjiang Abiderexiti. 2023. Chunk-bert: Boosted keyword extraction for long scientific literature via bert with chunking capabilities. In *2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 385–392.

A Appendix



Distribution of Token Size For Samples in BERT

Distribution of Token Size For Samples in RoBERTa

Distribution of Token Size For Samples in Longformer

Figure 4: Distribution of tokenized sample length for BERT, RoBERTa, Longformer.