Less is More: Stress Detection Through Condensed Social Media Contents

Zeyad Alghamdi[1], Tharindu Kumarage[1], Garima Agrawal[1], Huan Liu[1], and Russell Bernard[2]
[1]School of Computing and Augmented Intelligence, Arizona State University, Tempe, USA
[2]Institute for Social Science Research, Arizona State University, Tempe, USA

zalgham1@asu.edu
kskumara@asu.edu
garima.agrawal@asu.edu
huanliu@asu.edu
asuruss@asu.edu

Abstract: In the digital age, social media has been a go-to platform for stress-related discussions, yielding valuable data to advance the understanding and detection of stress. Swift identification of stress indicators in these online conversations is essential in enabling immediate support and helping to avert subsequent severe mental and physical health issues, especially during global crises such as pandemics and conflicts. Detecting stress in social media posts automatically poses a formidable challenge. While techniques such as supervised Pretrained Language Models (PLMs) and zero-shot Large Language Models (LLMs) based classifiers have demonstrated significant performance, they exhibit limitations, especially on platforms like Reddit. For example, on Reddit, users tend to write lengthy, expressive posts, which causes these methods to often fail to consider the entire context, leading to incomplete or inaccurate assessments of a user's mental health or stress status. To overcome these limitations, we present a new approach to identifying and classifying stress-related discourse on social media. Our approach involves analyzing condensed versions of user posts, such as user-provided summaries or the "Too Long Didn't Read" (TLDR) portion of the original post. We question whether these abridged texts can yield a more accurate classification of stress. In this paper, we make the following contributions. First, we investigate the relationship between the performance of the model's perceived textual context and the length of social media posts. Second, we present a novel approach to use the summarized texts for stress detection. We experiment with different classifiers to evaluate their performance on stress detection accuracy using summarized versus full-length posts. Furthermore, by examining the emotional and linguistic features of the original posts and their summaries, we suggest improvements to current state-of-the-art LLM-based stress classifier prompts, thereby enhancing stress detection capabilities. Finally, when user summaries are absent, we synthetically generate meaningful user post summaries by incorporating the power of LLMs. Our results show that the stress detection performance deteriorates for longer posts, and utilizing the TLDR and summaries improves classification outcomes. We also provide augmented datasets containing human and AI-generated summaries for future research in stress detection on social media.

Keywords: Mental health, Stress detection, Social media, Large language models (LLMs), Linguistic features analysis, Text summarization

1. Introduction

Recent global crises have escalated stress levels, profoundly impacting mental health worldwide. Data from the American Psychological Association (APA, 2023) reveals a significant increase in chronic illnesses and mental health diagnoses since the COVID-19 pandemic, with adults aged 18 to 34 reporting the highest rates. In this digital era, social media platforms have emerged as crucial forums for mental health discourse and support, offering both anonymity and empathy (Sowles, S.J. et al., 2018; Sher, L., 2020). More importantly, these mental health discourses on social media provide a wealth of textual data that can be leveraged for early stress detection. This, in turn, could play a pivotal role in mitigating and addressing severe mental health challenges.

In the field of stress detection on social media, supervised Pretrained Language Models (PLMs) have demonstrated state-of-the-art performance across various platforms, including Twitter and Reddit (Lin, H. et al.,2017; Nijhawan, T. et al., 2022). With the emergence of large language models (LLMs), recent research has presented compelling cases for LLMs serving as superior zero-shot stress classifiers capable of performing effectively across multiple social media platforms(Xu, X., et al., 2023; Lamichhane, B., 2023), alleviating the need for additional fine-tuning, as typically required by PLM-based stress classifiers. However, both fine-tuned PLM and zero-shot LLM-based stress classifiers encounter challenges when confronted with lengthy content. Yang, K., et al. (2023) have noted that models like ChatGPT face difficulties in effectively addressing long contextual posts. Moreover, as shown by Ji, S., et al. (2023), transformer-based models like BERT exhibit inherent limitations in processing long text content, constraining their effective range to a mere 512 tokens. This limitation is significant in the context of social media platforms like Reddit, where posts often exceed this length.

Our research addresses this problem by proposing an innovative approach distinct from direct long-document analysis. We hypothesize that integrating summaries could effectively mitigate the identified limitation for two primary reasons: **1) Conciseness:** Summaries distill the essence of longer posts while preserving crucial information; **2) Availability:** On platforms like Reddit, the prevalent use of "Too-Long-Didn't-Read" (TLDR) sections in lengthy posts presents a readily accessible summary, aiding stress detection. Consequently, our study enhances the established Dreaddit stress dataset (Turcan, E., and Kathleen M. 2019) by incorporating summaries derived from these TLDRs (in our paper, we use 'TLDR' and 'summary' interchangeably). Subsequently, we assess the effectiveness of summaries as an alternative unit of analysis for stress classification in lengthy posts. Furthermore, through a detailed analysis of the psychological, linguistic, and emotional features between these user-generated summaries and the complete posts, we propose modifications to current LLM-based classification prompts to boost stress detection efficacy further.

Additionally, our study investigates the potential of LLMs to produce summaries analogous to human-written TLDRs. The primary aim is to assess whether AI-generated TLDRs could serve as a viable substitute for stress classification in lengthy posts lacking human-authored TLDRs. To this end, we further supplement our dataset with TLDRs generated by ChatGPT. We then benchmark the stress detection performance on AI-generated TLDRs against those written by users. Our comparative analysis reveals that both human and AI-generated summaries significantly improve classification accuracy and F1 scores, particularly in the context of longer posts. In summary, our key contributions are as follows:

- We propose a novel approach for classifying stress using user-provided summary texts as an alternative to analyzing full-length posts.
- We augment the existing LLM-based stress classification prompts by conducting a psychological, linguistic, and emotional feature analysis on user-written posts against their user-written summaries.
- We study the effectiveness of LLM-written summaries for stress classification on longer posts that lack human-authored summaries.
- We release an augmented dataset comprising the original extended post texts, the user-provided summaries, summaries generated by LLMs, and associated features, making it available for other researchers in the field. The code and data are available in our GitHub repository: *(https://github.com/Zeyad-o/TLDR-AISummarizeStress/)*

2. Related Work

2.1 Evolution of Stress Detection Techniques in Social Media

The field of stress detection on social media has evolved substantially over time. Beginning with traditional text analysis methods, such as rule-based systems (Thelwall, M., 2017) and Latent Dirichlet Allocation (LDA) (Khan, A., Ali, R. (2020); Nijhawan, T. et al. (2022)). The scope of stress detection expanded when a new dimension of integrating multimodal data, including images and social network information was proposed by Lin, H., et al. (2014, 2017). Furthermore, Turcan, E., et al. (2021) and Alghamdi Z., et al. (2023), have delved into the role of emotions in stress detection from the poster's perspective. More recent studies have explored emotional disparity in social media comments as a novel approach for stress detection (Alghamdi, Z. et al., (2023)). The advent of LLMs marked a new era, Lamichhane, B. (2023) demonstrated ChatGPT's effectiveness in mental health classification tasks, achieving notable F1 scores. This study highlights ChatGPT's potential for mental health classification roles that are typically reserved for domain-specific models. In parallel, Xu, X., et al. (2023) evaluated multiple LLMs, including GPT-3.5, across various mental health tasks, showing promising results with zero-shot and few-shot prompting despite limitations related to the different LLMs used context window sizes compared to GPT-3.5. Furthermore, Yang, K., et al. (2023) advance this research by focusing on interpretable mental health analysis using LLMs. They address low interpretability of traditional methods by exploring different prompting strategies and the generation of explanations that are close to human performance. Most importantly, they highlight that ChatGPT shows strong in-context learning abilities, but it still falls short of advanced task-specific methods, indicating a need for careful prompt engineering.

2.2 Summarization in Mental Health

The application of summarization in mental health is increasingly being recognized as crucial. It has gained significant recognition for its ability to aid healthcare professionals. Manas, G., et al. (2021) demonstrate the importance of creating semantically relevant summaries from clinical diagnostic interviews. Gao, Y., et al. (2022) investigated how summaries of medical problems can help healthcare stakeholders accurately grasp patient conditions, easing their workload and reducing cognitive biases. Furthering the application of technology in this

domain, Li, Hao, et al. (2023) used LLMs to generate concise lists summarizing patients' problems, showcasing the value of LLMs in enhancing the efficiency of patient care. Kim, T., et al. (2023) applied LLMs to summarize psychiatric patients' experiences for clinician dashboards, enhancing patient monitoring. Syed, S., et al. (2023) used multiple LLMs, including GPT3.5 and GPT4, for summarizing extensive social media discussions, focusing on comments. This approach aids in navigating and analyzing complex social media content, demonstrating the practicality of LLMs in mental health and social media contexts. The growing interest in LLMs for text summarization is highlighted by studies from Zhang, Tianyi, et al. (2023), Pu, Xiao, et al. (2023), and Laban, P., et al. (2023). These studies focus on comparing human-generated and LLM-generated summaries, revealing the complexities and potential for further research in this area. While these studies underscore the growing validity of LLM-generated summaries in various contexts, our research focuses on the potential of human-generated summaries, specifically TLDRs, in the realm of stress detection on social media.

3. Methodology

In this section, we first discuss the process used to extract and prepare the data, followed by how the different text sizes were handled. Then, we present our approach for the analysis and classification of the features. For brevity, only the critical aspects of the method are presented here. The overall design and methodology are shown in Figure 1.

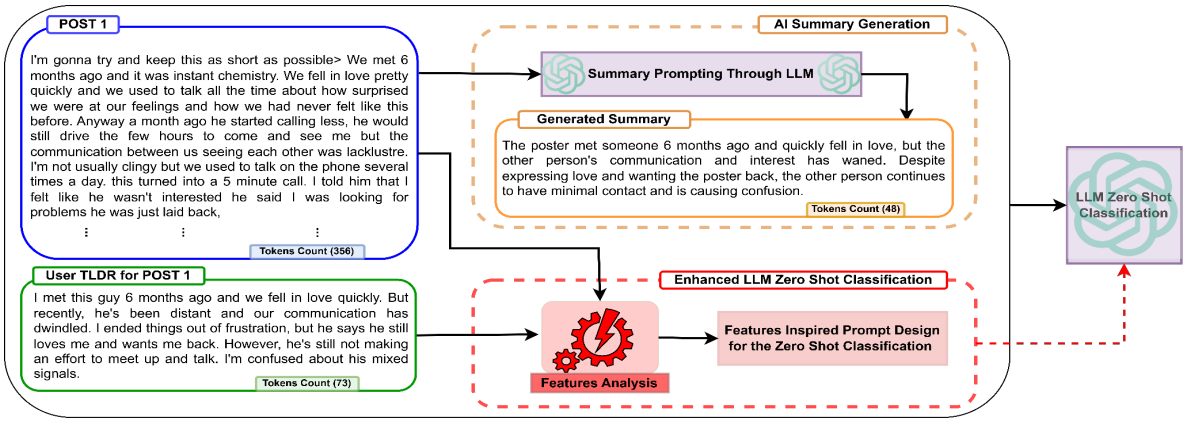

Figure 1: Methodology Pipeline - illustrates the flow of our approach with a sample from our dataset

3.1 Dataset

3.1.1 Dataset extraction

For our analysis, we utilized the Dreaddit dataset, a comprehensive and manually annotated collection of social media posts, specifically from Reddit. This dataset, made publicly available by Turcan, E., and Kathleen M. (2019), is widely recognized for its application in stress detection research. Spanning from January 1, 2017, to November 19, 2018, Dreaddit offers a diverse range of lengthy posts, encompassing various subreddits related to mental health issues, including abuse, anxiety, financial stress, PTSD, and social challenges.

Dreaddit's distinctive quality lies in its focus on everyday stress experiences, as opposed to strictly clinical scenarios, and its inclusion of extensive Reddit posts, which provide a rich insight into the multifaceted nature of stress expression in social media contexts. We utilized the PRAW API (Reddit, 2023) to extract these posts based on their unique IDs. Although the original Dreaddit dataset comprises 2,750 posts with individual classifications, our extraction yielded 1,984 posts, with the reduction mainly due to deletions by users or removals by moderators.

3.1.2 TLDRs extraction

TLDRs are commonly employed by users to encapsulate the essence of their posts. We employed regular expressions to methodically extract TLDRs and user-written summaries, ensuring they were explicit and identified by key phrases such as 'long story short,' 'basically,' 'short story,' among others, or as integral components of the post narratives. Our extraction process prioritized original content, deliberately excluding any edits influenced by subsequent user comments to maintain the authenticity of the posts' sentiment and intended message. This approach was crucial as our objective was to pinpoint early indicators of stress, and incorporating edits that might reflect positive feedback could potentially distort the original context and skew the results. From the initial dataset comprising 1,984 posts, we successfully narrowed it down to a subset of 527

samples that included both the original post text and user-written summaries. It is important to note, however, that not all TLDRs strictly conform to the traditional definition of summaries. Some users tend to list key points or pose questions, aiming to attract readers interested in the core message, who might then choose to engage with the full post or respond based on these highlighted elements.

3.1.3 Augmented dataset

In this work, we developed an augmented dataset that contains the cleaned original social media posts along with both user-generated and AI-generated summaries. This enriched dataset is designed not only for feature extraction and emotional analysis but also to augment the reproducibility of our research. By providing these diverse data elements in a consolidated form, we aim to facilitate future studies in the field of mental health and stress detection.

3.2 Features Extraction

In our study, we used the Linguistic Inquiry and Word Count (LIWC) software (LIWC-22; Boyd, R.L., et al., 2022), a text analysis program. LIWC, known for quantifying psychological and linguistic attributes in text, has been validated for stress detection in previous studies, including Turcan, E. and Kathleen M,.(2019). We employed LIWC's latest version, which extracts up to 118 features from unsegmented posts to maintain contextual integrity. It categorizes words into emotional, cognitive, and structural components, aiding in a comprehensive analysis of psychological constructs in the text. LIWC's ability to assess emotional and cognitive states in social media posts is crucial for detecting early stress indicators in our study.

3.3 Summarization

In this study, we utilized LLMs for summarization with a simple prompt:

> *"Summarize the following social media post: [POST]".*

This approach leverages the model's broad training, enabling it to adapt to various topics efficiently. The use of a straightforward prompt accentuates its capability in zero-shot learning, producing coherent and contextually appropriate summaries. This demonstrates the models' potential to rapidly process diverse mental health topics, and highlight their utility.

3.4 LLM Zero-Shot Classification

The design of our prompts was an integral part of our experiment. Drawing from previous research, the prompts were specifically crafted to align with our task's unique characteristics. This included incorporating expressions of current negative stress as identified in the work of Turcan and McKeown (2019) and Turcan, E., et al. (2021). Based on our dataset text types, we have the following: either post or (post TLDR or summary). Therefore, we use the following zero-shot prompt:

> *"Is the following [Text Type] indicative of current negative stress or not? Just answer in Yes or No. Don't provide explanations. [POST]".*

4. Experimental Setup

Here, we describe the experimental settings used to validate our approach, including the bucketization of the dataset, LLM usage, and stress detection baselines, to support reproducibility.

4.1 Bucketization

To investigate the relationship between stress classification performance and the textual data length, we divided the dataset into three segments into buckets or Quartiles 1, 2, and 3 (Q1, Q2, and Q3) based on the ascending order of token counts in the posts. Each quartile contains approximately one-third of the total dataset. The tokenization of each post was done using the OpenAI tokenizer (Turbo 3.5). Our analysis revealed the following average token counts per quartile:

- Q1 average is 178 tokens (minimum 80, maximum 268).
- Q2 average is 387 tokens (minimum 269, maximum 554).
- Q3 average is 982 tokens (minimum 555, maximum 3,833).

The overall dataset average is 515 tokens. Figure 2 shows the frequency distribution of token count in each quartile.

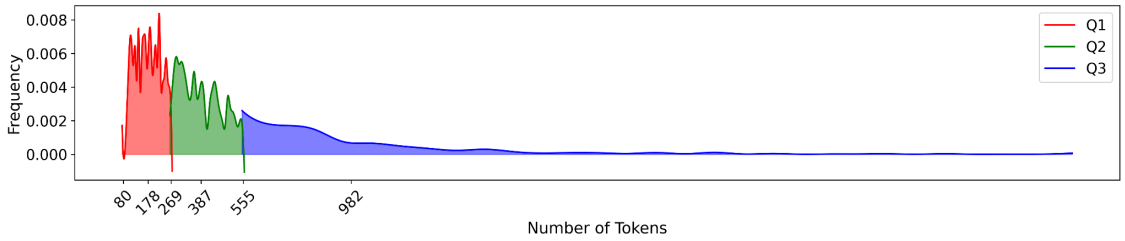

Figure 2: Post-token distribution across the buckets Q1, Q2, and Q3

4.2 LLM Settings

We selected ChatGPT (GPT-3.5) for its balance of capability and cost-effectiveness in the GPT series. With 175 billion parameters, it's refined through reinforcement learning from human feedback. Notably, this model has received appraisals for its performance in studies by Xu, X., et al. (2023) and Lamichhane, B. (2023). As a closed-source model available via OpenAI's API, we integrated the OpenAI library (version 0.28.1) with LangChain for zero-shot LLM prompting. We specifically used *'gpt-3.5-turbo-1106'* to ensure reproducibility. For deterministic classification tasks, we set the temperature to 0, and 0.7 for creative variability in summary generation. We employed LangChain's 'response schema' feature to direct LLM outputs based on input descriptions, using precise instructions for classification and summarization tasks.

4.3 Baseline Stress Classification

As a baseline for our comparative analysis, we used a state-of-the-art PLM-based stress classifier. To be specific, we used the BERT-base model fine-tuned on classifying stress data *('bert-base-cased')*. To ensure the robustness and reliability of our findings, we stratified the dataset by label, implementing an 80/20 split for fine-tuning and testing purposes. The performance metrics were calculated as an average of over 10 distinct runs, enabling a thorough and comprehensive evaluation of the PLM's effectiveness. It is important to note that PLMs have been a cornerstone in stress detection research, with numerous studies leveraging them to understand and classify stress-related content.

5. Experiments and Results

Here, we provide the experiments and analyze the results. We begin with a preliminary analysis of the effect of post-length on stress classification performance, then examine the three research questions in detail, where the role of human and AI-generated summaries in enhancing classification accuracy is discussed, along with a comprehensive performance benchmark. In summary, we address the following key research questions:

- RQ1: Can human-written summaries improve classification?
- RQ2: Are AI-generated summaries effective?
- RQ3: How do summarization approaches compare to baselines?

5.1 Preliminary Analysis - Post Size vs. Stress Classification Performance

We use the Zero-Shot LLM classifier on all the posts in the dataset and focus on evaluating the classification performance, specifically accuracy and F1 score, following the evaluation metrics outlined in related works. We assessed not only the overall performance but also the performance within each defined quartile. As depicted in Figure 3, distinct performance variations across the quartiles were observed. In quartile 1 (Q1), containing shorter posts, the classification performance was notably higher than in quartiles 2 (Q2) and 3 (Q3). Interestingly, Q2 consistently displayed a performance level near the overall average, while Q3, comprising longer posts, showed the lowest performance metrics. The accuracy difference between Q1 and Q3 in zero-shot classification was significant at 22.2%, and a similar trend was noted in the F1 score, with a difference of 15.3%.

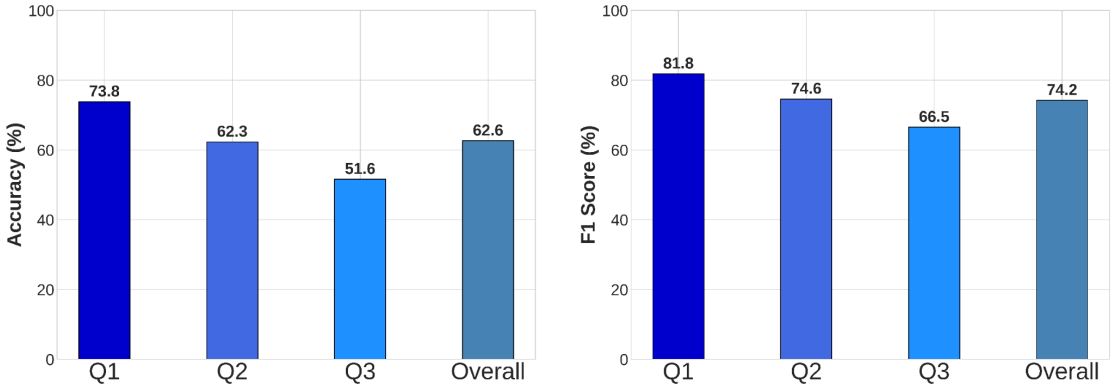

Figure 3: Quartile-Specific and the Overall Performance of Zero-Shot LLM classification

These results indicate that post length impacts classification effectiveness, with performance inversely related to input size. This finding underscores the need for alternative strategies to counteract performance declines associated with longer posts. Consequently, we look into our RQ1: "Can human-written summaries improve classification accuracy?". This question explores whether utilizing human-authored summaries can mitigate the performance drop observed with longer posts. The following section details our approach and findings in addressing RQ1.

5.2 Analysis of Human Written Summaries

Now, we present an analysis of a specific subset of the data comprising social media posts accompanied by user-generated summaries or "TLDRs." This subset consisted of 527 posts, with label distribution at 49% stress and 51% non-stress and an overall average token count of approximately 715 tokens. Notably, this subset predominantly featured longer texts, with 59% falling into Quartile 3 (Q3), 32% in Quartile 2 (Q2), and only 9% in Quartile 1 (Q1). The average token count for the TLDR segments was 44, constituting about 6.16% of the average total length of the original posts.

5.2.1 Stress Classification Performance

As shown in Figure 4, our analysis compared the classification performance using either the full posts or the user-written summaries. Employing the LLM zero-shot classifier, we observed that the summaries outperformed full post classifications, showing a 6.31% increase in accuracy and a 2% increase in F1 score. It suggests that user-generated short summaries could serve as an effective alternative for classifying stress in lengthy social media posts.

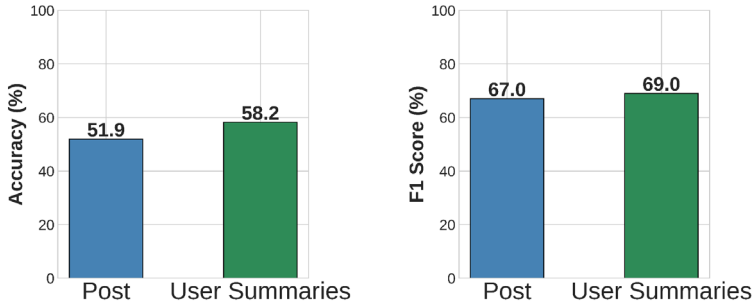

Figure 4: Comparative Performance of Full Posts vs. User-Written Summaries (TLDRs)

5.2.2 Feature Analysis

To further understand these results, it is imperative to understand the dominant characteristics and differences between the original posts and user-generated summaries. To achieve this understanding, we embarked on a comprehensive correlation analysis encompassing 118 features from the Linguistic Inquiry and Word Count (LIWC) tool. We focused on identifying features that exhibited a high correlation to the stress label, that are statistically significant ($P<0.05$) between the two types of texts, summaries, and original posts. Figure 5 shows a Venn diagram illustrating the top 10 features shared and unique to posts and user-written summaries (TLDRs).

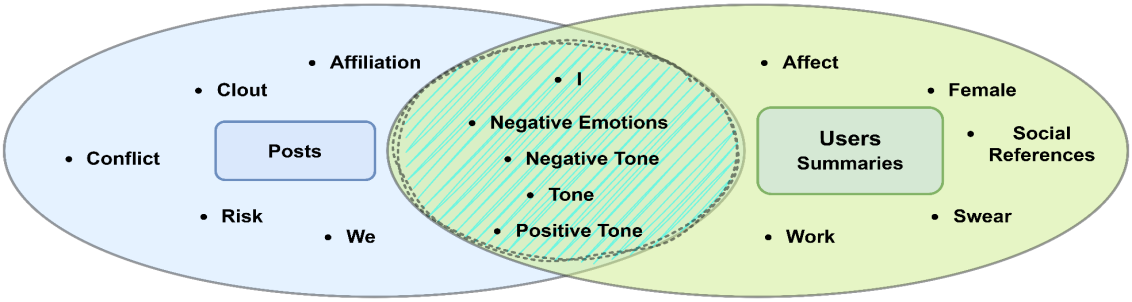

Figure 5: Venn Diagram of the Top 10 Features in Posts vs. User-Written Summaries (TLDRs)

Moreover, analyzing the intersection of features from posts and user-written summaries proves vital. Shared features like 'general Tone', 'Negative Emotions and Tone', ' lack of Positive Tone', and the first person singular pronouns such as ('I') in both posts and summaries are particularly telling. These elements, indicative of negative emotions and a self-centric narrative, are potent stress markers. Furthermore, each text type offers unique insights. Posts often feature elements like 'We', 'Affiliation,' 'Clout,' 'Conflict,' and 'Risk,' which delve into personal dynamics, interpersonal tensions, and perceived threats. On the other hand, the TLDRs, known for their brevity and focus, highlight specific aspects such as 'Female,' 'Work,' and 'Swear,' 'Affect,' and 'Social References'. These features shed light on gender reference, workplace stress, and emotional intensity more directly and succinctly.

5.2.3 Enhanced LLM-Classification

We have leveraged the insights gained from our feature analysis to refine the prompt used in existing LLM-based stress classifiers. Specifically, we have emphasized capturing first-person perspectives and emotional intensity, key elements highlighted by the dominant features shown in Figure 5. Consequently, we have revised the classification prompt, which we denote as the **'enhanced prompt'**:

> *"Given the following social media text (can be either post or post summary or post TLDR), looking from the poster's perspective, only classify if it is indicative of current very severe negative stress as 'Yes' otherwise 'No'. Just answer in 'Yes' or 'No'. Don't provide explanations. Text:[Text Type]"*

Figure 6 shows the enhanced prompt's results, demonstrating an improvement in post-classification accuracy from 51.9% to 57.6% and an increase in performance on human summaries from 58.2% to 60.0%, compared to the generic prompt results in Figure 4.

5.3 Analysis of AI-Generated Summaries

While the above findings are promising, they underscore a practical challenge: not all posts include user-generated summaries. This challenge compels us to explore the feasibility of employing AI-generated summaries as a potential alternative. Here, we try to answer the RQ2: "Can we use AI-generated summaries instead of human-generated summaries?". Figure 6 shows the performance comparison of full posts, human summaries, and AI-generated summaries using enhanced LLM classification.

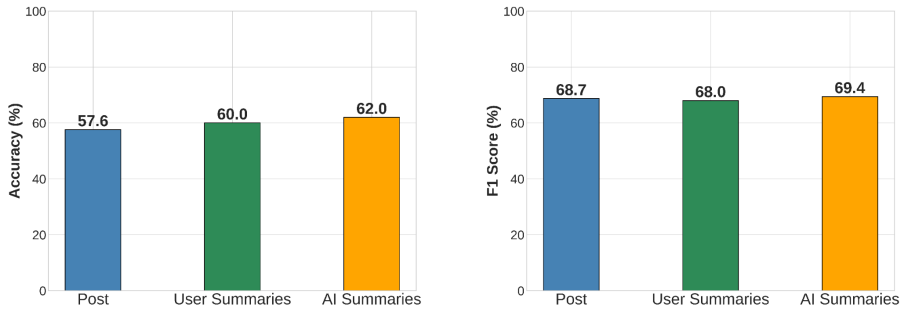

Figure 6: Performance Comparison of Subset Text Types with Zero Shot Enhanced Prompt

These results revealed that while human-written summaries outperformed full posts by 2.4% in classification accuracy, the AI-generated summaries further improved on this by an additional 2%.

Regarding the F1 score, human summaries and posts had similar performance, but AI-generated summaries showed a slightly better F1 score.

To evaluate the resource efficiency of summaries in stress classification, we analyzed token counts within the subset, including human summaries. The average counts were 715 tokens for original posts, 44 for human summaries (6.2% of the total), and 65 for AI-generated summaries (9.1%), as shown in Figure 7.

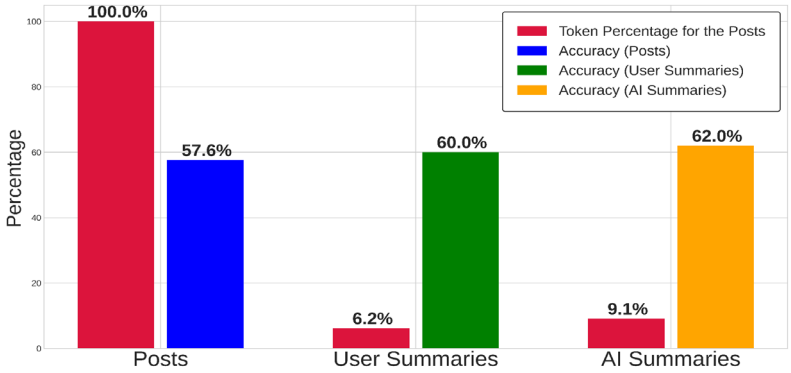

Figure 7: Token Percentage and Performance Comparison for Different Text Types

This analysis yielded two significant insights. First, human-generated summaries, with only 6.2% of the total token count, improved classification accuracy over full posts by 2.4%. More importantly, AI-generated summaries, constituting just 9.1% of the total tokens, outperformed both full posts and human summaries. The marginal increase in token count (2.9%) correlated with a notable 2% accuracy gain compared to human summaries. Thus, AI-generated summaries emerge not only as a valid approach for stress classification but also demonstrate that a slight increase in length can significantly enhance performance.

5.4 Benchmarking Summarization Impact on Classification

To address RQ3: 'How does the summarization approach compare against baseline methods?', we broaden our analysis to encompass the entire dataset. This comparison is critical to understanding the efficacy of our approach across different data buckets or quartiles and the overall performance, as illustrated in Table 1.

We observed that the PLM-based classifier (BERT) was better for the Q2 and Q3 quartiles, though the F1 scores are low. In contrast, the enhanced zero-shot LLM classifier gives the best classification accuracy. The improvement is notable, with a 67.5% accuracy and a 76.4% F1 score.

Table 1: Performance of Posts vs. Generic Summaries in LLMs and PLMs

Classifier / Metric		Q1		Q2		Q3		Overall	
		ACC%	F1%	ACC%	F1%	ACC%	F1%	ACC%	F1%
Post	BERT	73.5	71.3	67.5	63.6	62.0	57.0	66.9	63.6
	Generic Prompt	73.8	81.8	62.3	74.4	51.6	66.5	62.6	74.3
	Enhanced Prompt	78.3	**84.0**	69.2	77.5	54.9	67.9	67.5	76.4
AI Generated Summary	BERT	74.9	73.8	68.2	65.9	61.9	57.7	67.4	64.5
	Generic Prompt	75.3	82.5	65.0	75.6	54.0	67.9	64.8	75.2
	Enhanced Prompt	**78.5**	83.4	**71.5**	**78.1**	**62.4**	**70.3**	**70.8**	**77.4**

Shifting our focus to AI-generated summaries, the application of the zero-shot prompt for the AI-generated summaries we observe the following: i) notable improvement of 2.2% in accuracy and a 0.9% rise in F1 score, and ii) summaries are at least on par with, if not superior to full post classifications. Moreover, the PLM-based baseline on these summaries improved classification in Q1 and Q2, and overall (an increase of 1.7% in accuracy and 1.9% in F1 score) compared to the PLM full post-performance, albeit with a slight dip in Q3 performance. Our analysis concludes that applying the enhanced prompt to AI-generated summaries significantly enhances classification performance. It outperforms the generic zero-shot prompts (8.2% higher accuracy, 3.1% better F1 score) and PLM-based classifications (3.9% more accurate, 13.8% higher F1 score).

6. Limitations and Challenges

Our work, distinct from traditional long document classification, offers an alternative approach where summaries, rather than full-length posts, are utilized for efficient stress analysis. This contribution is particularly crucial considering the complexity and volume of content on social media platforms.

One challenge we encounter is the "black-box" nature of LLMs, which complicates obtaining optimal prompts for classification and summarization tasks. We have attempted to approximate the necessary prompt parameters through feature analysis, but this remains an area that needs further exploration and refinement. We also acknowledge potential limitations in our approach, mainly when dealing with very long texts that may need to be adequately summarized. Our findings showed that the quality of summaries, especially for longer texts (Q3 data), appeared to affect the classification performance negatively when using BERT. This phenomenon suggests that the effectiveness of summarization may vary depending on the text length and warrants additional research.

Another challenge we encounter is the well-known hallucination problem in LLMs (Agrawal, Garima, et al. (2023)). To avoid prompts that can generate hallucinations or unrelated output, we inspected the output generated by the summarization prompts. Multiple graduate students were asked to review around 50-100 randomly selected AI-generated summaries to check the original post's reflection, accuracy, and consistency in the LLMs' summary.

7. Conclusion and Future Work

This research contributes to the growing body of work in mental health analysis on social media, particularly in the context of stress detection. By focusing on the utility of summaries for stress classification, we have demonstrated a practical approach that mitigates some of the challenges posed by lengthy social media content. Our findings indicate that human and AI-generated summaries can serve as viable alternatives to full-length posts for stress detection, with AI-generated summaries showing particular promise due to their scalability and efficiency. This suggests a significant potential for the application of LLMs in mental health monitoring and support systems through social media data. Which complements the usage of summarization of mental health data through medical records and reports.

Our research opens avenues for further investigation into the optimization of LLMs for mental health applications. Future studies could focus on refining summarization techniques to capture the essence of lengthy posts more accurately and expanding the dataset to include a broader range of social media platforms and languages, which could provide a more comprehensive understanding of stress expression in diverse digital environments.

Acknowledgement

This work was supported by the Office of Naval Research under Award No. N00014-21-1-4002. We also appreciate the valuable guidance and support provided by Wildflower Primary Care & Wellness Center. Interpretations, conclusions, and recommendations within this article are solely those of the authors.

References

Agrawal, G., et al. (2023) "Can Knowledge Graphs Reduce Hallucinations in LLMs?: A Survey", arXiv preprint arXiv:2311.07914.

Alghamdi, Z., et al. (2023). "Code RED: Reactive Emotion Difference for Stress Detection on Social Media". No. 10659. EasyChair.

Alghamdi, Z., et al. (2023) "Studying the Influence of Toxicity and Emotion Features for Stress Detection on Social Media", ECSM 2023 10th European Conference on Social Media, Academic Conferences, and Publishing Limited.

American Psychological Association. (2023) "Stress in America." [Online] Available at: https://www.apa.org/news/press/releases/stress

Boyd, R.L., et al. (2022) "The Development and Psychometric Properties of LIWC-22", Austin, TX: University of Texas at Austin.

Gao, Y., et al. (2022)"Summarizing Patients' Problems from Hospital Progress Notes Using Pre-trained Sequence-to-Sequence Models." Proceedings of COLING. International Conference on Computational Linguistics. Vol. 2022. NIH Public Access.

Ji, S., et al. (2023)"Domain-specific Continued Pretraining of Language Models for Capturing Long Context in Mental Health." arXiv preprint arXiv:2304.10447.

Khan, A. & Ali, R. (2020) "Stress Detection from Twitter Posts Using LDA", International Journal of High Performance Computing and Networking, 16(2-3), pp. 137-147.

Kim, T., et al. (2023) "MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling", arXiv preprint arXiv:2310.05231.

Laban, P., et al. (2023) "SummEdits: Measuring LLM Ability at Factual Reasoning Through The Lens of Summarization", Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.

Lamichhane, B. (2023) "Evaluation of chatgpt for nlp-based mental health applications", arXiv preprint arXiv:2303.15727.

Li, H., et al. (2023) "Team: PULSAR at ProbSum 2023: PULSAR: Pre-training with Extracted Healthcare Terms for Summarising Patients" Problems and Data Augmentation with Black-box Large Language Models', The 61st Annual Meeting Of The Association For Computational Linguistics.

Lin, H., et al. (2014) "User-level Psychological Stress Detection from Social Media Using Deep Neural Network", Proceedings of the 22nd ACM International Conference on Multimedia, pp. 507-516.

Lin, H., et al. (2017) "Detecting Stress Based on Social Interactions in Social Networks", IEEE Transactions on Knowledge and Data Engineering, 29(9), pp. 1820-1833.

Manas, G., et al. (2021) "Knowledge-infused Abstractive Summarization of Clinical Diagnostic Interviews: Framework Development Study", JMIR Mental Health, 8(5), e20865.

Nijhawan, T., Attigeri, G., & Ananthakrishna, T. (2022) "Stress Detection Using Natural Language Processing and Machine Learning Over Social Interactions", Journal of Big Data, 9(1), pp. 1-24.

Pu, X., Mingqi, G., & Xiaojun, W. (2023) "Summarization is (almost) Dead", arXiv preprint arXiv:2309.09558.

Reddit. (2023) "PRAW: The Python Reddit API Wrapper", Available from: https://praw.readthedocs.io/.

Syed, S., et al. (2023) "Indicative Summarization of Long Discussions", The Conference on Empirical Methods in Natural Language Processing, pp. 2752–2788, Singapore, Association for Computational Linguistics.

Sher, L. (2020) "The Impact of the COVID-19 Pandemic on Suicide Rates", QJM: An International Journal of Medicine, 113(10), pp. 707-712.

Sowles, S.J., et al. (2018) 'A Content Analysis of an Online Pro-eating Disorder Community on Reddit', Body Image, 24, pp. 137-144.

Thelwall, M. (2017) "TensiStrength: Stress and Relaxation Magnitude Detection for Social Media Texts", Information Processing & Management, 53(1), pp. 106-121.

Turcan, E., Smaranda, M., & Kathleen, K. (2021) "Emotion-infused Models for Explainable Psychological Stress Detection", Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Turcan, E., & Kathleen, M. (2019) "Dreaddit: A Reddit Dataset for Stress Analysis in Social Media", Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019).

World Health Organization. (2023) 'World Mental Health Report', [Online] Available from: https://www.who.int/teams/mental-health-and-substance-use/world-mental-health-report.

Xu, X., et al. (2023) "Leveraging Large Language Models for Mental Health Prediction via Online Text Data", arXiv preprint arXiv:2307.14385.

Yang, K., et al. (2023) "On the Evaluations of ChatGPT and Emotion-Enhanced Prompting for Mental Health Analysis", arXiv preprint arXiv:2304.03347.

Yang, K., et al. (2023) "Towards Interpretable Mental Health Analysis with Large Language Models", Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.

Zhang, T., et al. (2023) "Benchmarking Large Language Models for News Summarization", arXiv preprint arXiv:2301.13848.