Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
4(2): 234–252.
DOI: 10.3934/aci.2024014
Received: 14 November 2024
Revised: 09 December 2024
Accepted: 12 December 2024
Published: 18 December 2024

Research article

A transformer-driven framework for multi-label behavioral health classification in police narratives

Francis Nweke[1,*], **Abm Adnan Azmee**[1], **Md Abdullah Al Hafiz Khan**[1], **Yong Pei**[1], **Dominic Thomas**[2] **and Monica Nandan**[3]

[1] Department of Computer Science, Kennesaw State University, Marietta, GA 30060, USA

[2] Department of Information Systems and Security, Kennesaw State University, Kennesaw, GA 30144, USA

[3] Department of Social Work and Human Services, Kennesaw State University, Kennesaw, GA 30144, USA

* **Correspondence:** Email: fnweke@students.kennesaw.edu.

Academic Editor: Jidong Yang

Abstract: Transformer-based models have shown to be highly effective for dealing with complex tasks in a wide range of areas due to their robust and flexible architecture. However, their generic nature frequently limits their effectiveness for domain-specific tasks unless significantly fine-tuned. We understand that behavioral health plays a vital role in individual well-being and community safety, as it influences interpersonal interactions and can significantly impact public safety. As a result, identifying and classifying these cases demands the use of an effective tool, such as a framework, that has been fine-tuned to context-specific behavioral health issues. In this work, we demonstrated a trainable lightweight approach for addressing behavioral health analysis utilizing feature embeddings generated from transformer-based models. To facilitate in domain adaptation, we created instruction sets based on annotations by subject matter experts, enabling for targeted fine-tuning of the large language model (LLM) for behavioral health applications. Our experiments demonstrated that parameter-frozen transformer-based models can capture high-quality feature representations that allowed for the integration of a lightweight framework, making them especially useful in resource-constrained settings.

Keywords: behavioral/mental health; transformers; large language model; neural networks

1. Introduction

An individual's behavioral health is crucial because it determines how they interact with others in their community. The substance abuse and mental health services administration (SAMHSA) [9] estimates that 53 million adults had a mental disorder in 2020. It is critical to understand that behavioral health includes not just mental health, but also substance use, lifestyle choices, and habits. Behavioral health disorders affect millions of people, from mild to severe cases [10].

The rising number of behavioral health-related incidents has put an immense pressure on community resources. With over 240 million 911 calls made yearly, law enforcement personnel spend a significant percentage of their time responding to service calls [26]. The standard procedure for each reported incident is for first responders to arrive, provide required assistance to the involved parties, and to file an incident report. Each month, these reports are manually evaluated to identify cases of behavioral health issues, which can often result in incorrect identification [12]. This process is not only time-consuming, but also error-prone due to a lack of expertise. Early detection of behavioral and mental health issues is essential for preventing harm to both individuals and others [5].

To address this issue, machine learning approaches have been utilized for identifying behavioral and mental health issues. However, these efforts have predominantly focused on identifying specific subclasses of behavioral health, such as schizophrenia, bipolar disorder, anxiety, depression, post-traumatic stress disorder, and other mental health conditions [7, 16, 18–20, 36], and detecting mental health issues from social media data [3, 20, 33]. LLMs are widely used in several tasks, such as vulnerability detection [31], mental health detection via online text [37], and domain-specific machine translation [41].

Since its inception in 2018, the bidirectional encoder representations from transformers (BERT) model [8] was developed as a pretrained transformer trained on a large corpus of text, enabling fine-tuning for specific natural language processing (NLP) tasks. BERT's use of stacking transformers for bi-directional processing led to various variants of the original model, including the robustly optimized BERT pretraining approach (RoBERTa) [22], a generalized autoregressive pretraining method (XLNet) [39], decoding-enhanced BERT with disentangled attention (DeBERTa) [13], and a lite BERT for self-supervised learning of language representations (ALBERT) [21]. Transformer-based LLMs, such as generative pre-trained transformer 4 (GPT-4) [1], Mistral [17], and large language model meta artificial intelligence (LLaMA) [34], consist of millions or even billions of parameters, making them highly robust and powerful. The LLaMA [34] family includes models with different parameter numbers and configurations. While the family's architecture is similar, each model is unique in its configuration, providing a wide range of capabilities. Currently, the LLaMA model family has just a handful but rising number of models [34]. However, pretrained LLMs struggle when assigned a domain-specific task. To achieve the best results on a particular task, LLMs must be fine-tuned using domain-specific data. For our work, we collaborated with subject matter experts (SMEs) to extract ground data from police narrative reports. In the first configuration, we utilized a parameter-frozen RoBERTa model to generate feature embeddings from the dataset, which we then fed into trainable fully connected layers for classification. In the second configuration, we fine-tuned a pretrained LLaMA-3.2-3B model using instruction sets and integrated its feature embeddings with trainable fully connected layers for multi-label classification. We summarize our contributions as follows:

- We analyzed deep learning models, including long short-term memory (LSTM) [14], bi-directional long short-term memory (BiLSTM) [30], and convolutional neural network (CNN)+LSTM (a hybrid model), for behavioral health multi-label classification in various configurations, utilizing a number of word embeddings such as model-generated embeddings, Word2Vec [24], GloVe [29], and fastText [4].

- We propose a trainable framework with a configurable architecture, providing the following capabilities:
    - Classifies police narrative reports into one or more behavioral health classes through multi-label classification.
    - Integrates transformer-based models to generate semantic feature embeddings relevant to the behavioral health context.
- We created prompts/instructions sets leveraging police narrative reports to fine-tune LLMs.
- We demonstrated the effectiveness of transformer-based models, in this case, RoBERTa and LLaMa-3.2-3B, were used in our experiments for extracting domain-specific representations.
- Finally, our evaluations show that assessing a model merely on accuracy does not accurately reflect its performance in sensitive domains like behavioral health. A model may excel in detecting a specific class yet fail to capture all of the dataset's variables, resulting in a partial representation of its performance.

2. Intersection of behavioral health, machine learning and transformers

In this section, we review existing work on the intersection of behavioral health and machine learning in Subsection 2.1, and transformers in Subsection 2.2.

2.1. Behavioral/mental health and machine learning

According to SAMHSA, approximately 53 million adults in the United States suffered from a mental disorder in 2020 [9]. This staggering statistic underscores the crucial need of effective efforts to address mental health issues, which are a major source of public health concerns all over the world. During the same time period, the COVID-19 pandemic began to exacerbate a variety of disorders. According to the study [11], the pandemic has undoubtedly increased the number of cases with behavioral health issues. Behavioral health is a critical problem that requires immediate attention since, if not medically diagnosed, it can have an impact on people without being aware of it. In more recent times, machine learning has provided several techniques for detecting a broad variety of mental health issues.

Several studies have focused on detecting mental health categories such as depression, anxiety, schizophrenia, and suicide [16, 18–20, 36]. Previous research on NLP classification approaches has revealed useful frameworks for dealing with this issue. Recurrent neural networks (RNNs) are commonly used for processing sequential data in NLP tasks, however, they struggle to retain information over extended text sequences due to the vanishing gradient problem, resulting in a tendency to "forget". To address this issue, more advanced variations have been developed, which include LSTM [14], BiLSTM [30], and gated recurrent unit (GRU) [6]. LSTMs tackle the "forgetting" problem in RNNs by utilizing memory cells and gates (input, forget, and output) to maintain long-term dependence. BiLSTMs improve this by processing sequences in both directions, capturing context from previous and future tokens. GRUs simplify LSTMs by combining gates into an update gate, reducing complexity while accurately representing long-range relationships. CNNs [27], unlike RNNs, are designed for pattern recognition rather than sequential processing. However, CNNs can capture local features and patterns within text, such as n-grams, making them beneficial to RNN-based models like LSTMs, BiLSTMs, and GRUs for tasks that need both local and sequential context knowledge.

The work [23] compared neural network architectures (CNNs, GRUs, LSTMs, and Bi-LSTMs) for classification of text in unstructured medical data. The study [2] created a model utilizing self-attention techniques in order to learn contextual features from unstructured data and includes domain expert knowledge by leveraging expert-provided keywords. The integration of context and domain-specific data increases the model's understanding of social and behavioral health cues. While existing machine learning models for detecting behavioral/mental health concerns are reliable, they frequently lack the transparency needed in sensitive domains like behavioral/mental health and healthcare in general.

2.2. Transformer-based LLMs

Prior to the the introduction of transformers, deep learning models struggled with adequately capturing the context of a document. The transformer model introduced the attention mechanism to focus on the most relevant elements of a document to grasp the key idea [35]. This breakthrough has resulted in widespread success in a variety of applications. LLMs are transformer-based models with hundreds of billions of parameters that thrive at understanding natural language and accomplishing complex tasks [40]. LLMs can be fine-tuned for domain-specific tasks, as demonstrated by the success of zero-shot approaches in domain-specific machine translation [41]. Similar research like ours used online text to fine-tune an LLM for mental health prediction [37]. The work (MentaLLaMa) [38] proposed an interpretable model for mental health analysis based on ChatGPT's generating capabilities. Another interesting work [33] proposes a model with reasoning abilities comparable to human experts in the field of medicine.

Research has shown that fine-tuning LLMs can improve their performance on specific tasks. Typically, fine-tuning has focused on LLMs tailored for specific downstream tasks or settings, which often requires separate fine-tuning for each task. This approach demands extensive training resources, and has deployment and maintenance challenges. Several researchers have investigated LLMs as encoders. Another work [28], they utilized a frozen transformer block from pretrained LLMs as an encoder layer to decode visual tokens. This is an interesting approach considering that LLMs are typically used to text data, making this application in the visual domain particularly intriguing. Other researchers have used LLMs to encode clinical data and developed an automated evaluation framework to benchmark LLM performance in the clinical domain [32]. A neural semantic encoder was implemented in a different work, with a novel memory update rule and a variable-sized encoding memory that changes over time to preserve comprehension of input sequences via read, compose, and write operations [25]. Such an encoder is useful for a variety of natural language tasks, including machine translation and document sentiment analysis.

3. Transformer-enhanced lightweight framework.

In this section, we will discuss the architecture of the framework for behavioral health analysis, shown in Figure 1.

We leveraged parameter-frozen transformer-based models as an encoder to extract high-dimensional contextual embeddings for each text input. Our design has various components.

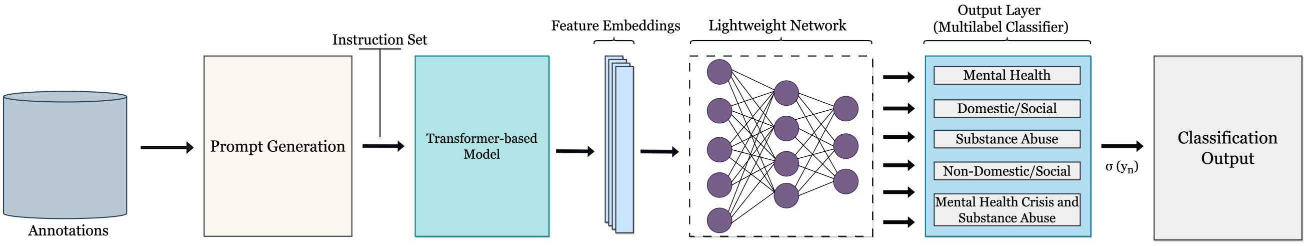


Figure 1. Transformer-enhanced lightweight framework architecture.

3.1. Instruction set

A hard prompt/instruction is generated to fine-tune the LLM based on the annotations. The input is organized in this prompt generation phase, as shown in Table 1. Usually, it entails developing a set of instructions to guide the LLM to classify the input text into one of more behavioral health classes. Formally, we define the instruction as

$$\text{Prompt} = f(x, A),$$

where x is the input text (narrative), and $A = \{a_1, a_2, \ldots, a_k\}$ represents the set of annotations or behavioral health class labels provided by SMEs. The function f constructs an instruction set that organizes the input and specifies the task for the LLM, such as

$$\text{Sample Prompt} = \begin{cases} \textbf{Public Narrative Report: } x, \\ \textbf{Prompt: } \text{Classify the narrative into one or more behavioral health classes:} \\ \qquad\qquad \{A_1, A_2, \ldots, A_n\}, \\ \textbf{Output: } \{y_1, y_2, \ldots, y_m\}. \end{cases}$$

A practical instance of a prompt based on the dataset is presented in Table 1.

Table 1. An example of a public narrative report as a instruction set.

Public Narrative Report	Prompt	Output
Person6 contacted ORG1 to report that her **child** (an infant) was injured and needed medical attention. PERSON2 stated that her brother (PERSON3) had **grabbed her arm causing the child to fall to the ground**, see incident report CARDINAL...	Classify the narrative into one or more behavioral health classes: mental health, substance abuse, domestic/social, non-Domestic/social, and mental health crisis and substance abuse.	Classification: `[‘‘mental health’’]`

3.2. Feature embedding extraction

A transformer-based model is fundamental to the architecture. The model transforms the input text into high-dimensional contextual embeddings that capture the narrative's semantic meaning, this model acts as an encoder. For the transformer models, we employed RoBERTa [22], known for its lightweight design and higher performance over BERT and other models within the family. LLaMA-3.2-3B, based on hardware constraints, is capable of extracting domain-specific details with fewer parameters.

- **Input Encoding:** Upon tokenizing and embedding the input text x, a sequence of token embeddings is generated
$$\mathbf{x} = [x_1, x_2, \ldots, x_n],$$
where x_i represents the i-th token in the input sequence.
- **Contextual Embedding Generation:** For each token, we compute contextual embeddings using the transformer-based LLM, which produces
$$\mathbf{H} = \text{LLM}(\mathbf{x}) = [h_1, h_2, \ldots, h_n],$$
where h_i is the high-dimensional contextual embedding for token x_i.
- **Sentence Embedding:** Token embeddings are usually pooled to generate the final embedding h for the input narrative
$$h = \text{Pooling}(\mathbf{H}).$$

The output of the transformer-based model is a set of feature embeddings, which are high-dimensional representations of the input text. These embeddings h combine the semantic context of the narrative in a high-dimensional space, enabling the downstream components to leverage this information for classification.

3.3. Dimensionality reduction and trainable layers

The framework's initial layer performs dimensionality reduction, which optimizes computing efficiency while preserving essential features. This layer uses a linear transformation to reduce the dimensionality of the embedding to 256. The reduced embeddings are then passed to a trainable layer, allowing the model to learn task-specific patterns efficiently.

$$h_{\text{reduced}} = W_{\text{reduce}} \cdot h + b_{\text{reduce}},$$

where $W_{\text{reduce}} \in \mathbb{R}^{256\times d}$ and $b_{\text{reduce}} \in \mathbb{R}^{256}$ are the learned weights and biases of the layer.

The resulting $h_{\text{reduced}} \in \mathbb{R}^{128}$ is a lower-dimensional embedding that retains the relevant information for multi-label classification. Now, it is then passed to additional dense layers, which further refine and capture relationships between different features within the narrative report text. These layers help in capturing complex interactions between features, enhancing the model's capacity to distinguish between behavioral health classes.

3.4. Behavioral health classifier

The output layer (classifier) applies a linear transformation to the reduced and refined embedding vector obtained from the last hidden layer. Let $h_{\text{reduced}} \in \mathbb{R}^{256}$ be the output of the last hidden layer. The output layer computes the logits for each behavioral health class as follows:

$$\text{logits} = W_{\text{output}} \cdot h_{\text{reduced}} + b_{\text{output}},$$

where

- $W_{\text{output}} \in \mathbb{R}^{\text{num\_labels}\times 256}$ is the weight matrix of the output layer.

- $b_{\text{output}} \in \mathbb{R}^{\text{num\_labels}}$ is the bias vector of the output layer.
- num_labels = len(bhr_type_classes) is the number of behavioral health classes.

Remember that our goal is to classify the input into one or more behavioral health classes, resulting in a multi-label classification task.

$$\text{output} = \sigma(\text{logits}),$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function applied element-wise.

The resulting output $\in [0, 1]^{\text{num\_labels}}$ vector has probabilities showing the likelihood of the input belonging to each behavioral health class.

$$\hat{Y}_i = \begin{cases} 1, & \text{if output}_i \geq \theta, \\ 0, & \text{if output}_i < \theta, \end{cases}$$

where $\hat{Y}_i$ is the predicted label for class i, and θ is the threshold.

To optimize the framework during training, we used binary cross-entropy (BCE) loss as the loss function:

$$\mathcal{L} = -\frac{1}{\text{num\_labels}} \sum_{i=1}^{\text{num\_labels}} \left[Y_i \log(\text{output}_i) + (1 - Y_i) \log(1 - \text{output}_i) \right],$$

where $Y_i \in \{0, 1\}$ is the ground truth label for class i.

4. Experimental setup

In this section, we will provide an overview of the dataset, alongside details about its creation, implementation settings, and the evaluation metrics used to assess model performance in our experiments.

4.1. Data description and analysis

The data utilized for this work was annotated by an SME who used an annotation tool to extract ground truths from police narrative reports. These reports are classified into one of several behavioral health classes as defined by SMEs. We utilized around 2000 annotated reports for the experiments that we conducted. The classes are presented in the Table 2.

Table 2. Behavioral health classes.

Class	Description
Mental health crisis & substance abuse	Diagnosed mental disorder induced by substance abuse, like schizophrenia.
Mental health	Diagnosed mental disorder, such as schizophrenia or suicidal ideations.
Domestic/social	Involves multiple individuals in a home setting, such as domestic disputes.
Non-domestic/social	Involves multiple individuals outside of a home setting, such as crimes against unrelated persons.
Substance abuse	Persistent drug/alcohol abuse problems.

Table 3 provides the statistics of the annotated report done by SME. These statistics allow for a more in-depth analysis of the dataset's representation.

Table 3. Annotated report statistics.

Description	Value
Number of sentences	67,090
Number of tokens	1,273,954
Average number of sentences per document	32.36
Average number of tokens per document	614.55
Average sentence length	18.9
Count of mental health crisis and substance abuse cases	97
Count of mental health cases	331
Count of domestic/social cases	1394
Count of non-domestic/social cases	169
Count of substance abuse cases	561
% of train samples	80
% of test samples	20

4.2. Instruction dataset construction

We constructed the instruction sets to adhere to the structure of instruction, input, and output. Where the instruction specifies an action to be performed, the text is the public narrative report, and the output is the class label(s).

To construct a dataset suitable for classification learning in behavioral health analysis, we formatted it in an instruction-like format as shown in Table 1, which allows the model to classify inputs. We achieved this format using a customized function, ensuring consistency and adaptability across a wide range of data.

4.3. Evaluation

To evaluate the effectiveness of our study, we used evaluation criteria such as precision, recall, f1-score, and accuracy. These metrics have been selected as they provide useful insights into the model's ability to effectively identify and classify behavioral health classes within reports.

4.3.1. Accuracy

The model's ability to correctly identify the class. It can be expressed as a ratio of precise predictions to total predictions. Mathematically, it is calculated as the ratio of the sum of true positive (TP) and true negative (TN) to the sum of TP, TN, false positive (FP), and false negative (FN).

4.3.2. Precision

It measures how well the model can identify positive values.

4.3.3. Recall

It measures the model's ability to properly identify the TP value.

4.3.4. F1-score

This metric provides a balance between the recall and precision score. It is also known as the harmonic mean of precision and recall.

4.4. Implementation details

We implemented the framework with PyTorch, a versatile deep learning library that supports dynamic computational graphs and powerful graphics processing unit (GPU), making it ideal for dealing with complex neural network architectures. We fine-tuned the LLM on a single high-performance NVIDIA A4500 GPU with CUDA. To improve efficiency and reduce memory usage without sacrificing performance, we fine-tuned LLaMA using low-rank adaptation (LoRA) [15], leveraging LLM's feature embeddings as learnable weights for the framework. First, we reduced the dimensionality of embeddings from 1024 to 256. In addition, we bound the length of the narrative report (text) to n characters due to computational constraints.

To extensively assess model performance, we split the dataset into an 80:20 training:testing split. The training set was used to fine-tune the model's parameters and the test set to evaluate the model's generalization to unseen data.

5. Results

This section presents the findings of our experiments, focusing on key research questions. These questions emphasize on the model's performance and applicability.

- **RQ1:** How effectively does our proposed framework perform in comparison to state-of-the-art (SOTA) models?
- **RQ2:** Can the proposed framework classify cases with complex and overlapping classes?
- **RQ3:** How does the use of the transformer-based models to generate feature embeddings impact the semantic understanding of narrative reports?
- **RQ4.** What are the common errors made by the framework?

5.1. SOTA models (RQ1)

We compared SOTA models to our fine-tuned LLM encoder for multi-label classification of behavioral health cases. We compared performance using deep learning models, including LSTM, BiLSTM, and CNN+LSTM. We additionally evaluated the results of the classifier enhanced with transformer-based models. Each deep learning model was evaluated across four different configurations, as summarized in Table 4. Our experiments utilized widely adopted word embeddings: Word2Vec [24], GloVe [29], and fastText [4]. Word embedding is an essential technique in NLP that represents text and vocabulary as numerical vectors in a continuous vector space. This representation captures semantic relationships, with comparable words having closer vector representations which help with contextual interpretation. Word2Vec uses predictive models to learn word associations, GloVe uses co-occurrence statistics to capture global relationships, and fastText improves conventional embeddings by taking into account subword information, which efficiently handles uncommon and out-of-vocabulary words.

Table 4. Evaluation results for multi-label classification over 20 epochs.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Deep Learning Models				
LSTM [without pre-trained embeddings]	78.0%	82.8%	56.2%	66.9%
LSTM [Word2Vec embeddings]	80.7%	83.5%	65.0%	73.1%
LSTM [GloVe embeddings]	80.9%	84.3%	67.8%	75.1%
LSTM [fastText embeddings]	81.4%	81.3%	63.7%	71.4%
BiLSTM [without pre-trained embeddings]	75.9%	88.6%	63.1%	73.7%
BiLSTM [Word2Vec embeddings]	81.2%	82.7%	79.6%	81.2%
BiLSTM [GloVe embeddings]	80.9%	84.0%	75.9%	79.8%
BiLSTM [fastText embeddings]	80.9%	82.3%	80.0%	81.2%
CNN+LSTM [without pre-trained embeddings]	78.8%	82.4%	58.9%	68.7%
CNN+LSTM [Word2Vec embeddings]	77.5%	82.9%	60.5%	70.0%
CNN+LSTM [GloVe embeddings]	79.3%	86.4%	53.6%	66.2%
CNN+LSTM [fastText embeddings]	79.0%	81.6%	59.6%	68.9%
Transformer-based models				
Base LLM (RoBERTa) + multi-label classifier	53%	76%	69%	72%
Fine-tuned LLaMA + multi-label classifier	49%	69%	57%	62%

Table 4 shows the evaluation results of various models for identifying behavioral health classes in police narrative reports. We categorized the models into two categories: deep learning models and transformer-based models, with each demonstrating its efficacy.

The deep learning models, with and without pretrained embeddings, were evaluated. Across all configurations, models with pretrained embeddings outperformed those without, demonstrating the benefit of pretrained embeddings. LSTM with GloVe embeddings outperformed other LSTM models in terms of F1-score (75.1%) and recall (67.8%), while LSTM without embeddings performed poor (F1-score of 66.9%). The BiLSTM model using Word2Vec had the highest F1-score (81.2%), followed by fastText. CNN+LSTM models had reduced performance, compared to BiLSTM, suggesting that the added CNN layer did not provide significant improvements in this task.

The RoBERTa model exhibited strong performance across every metric for transformer-enabled models, achieving 53% accuracy, 76% precision, 69% recall, and 72% F1-score, while utilizing shorter sequence lengths and a two-layer network. In comparison, the fine-tuned behavioral health LLM performed similarly to RoBERTa, which excels in encoding contextual embeddings for classification rather than text generation.

5.2. Class-wise performance (RQ2)

Each narrative report in the dataset is labeled with one or more behavioral health classes, enabling multi-label classification. Figure 2 shows the distribution of behavioral health in the dataset. "Domestic/social" class is the most common class, with 1,394 cases. "Substance abuse" class has 561 cases, while “mental health” class has 331 cases. There are 169 cases of “nondomestic/social” class. “Mental health crisis and substance abuse” class includes 97 cases, which are a composite of two (2) classes. The disparity between classes highlights the necessity for techniques to ensure balanced model performance.

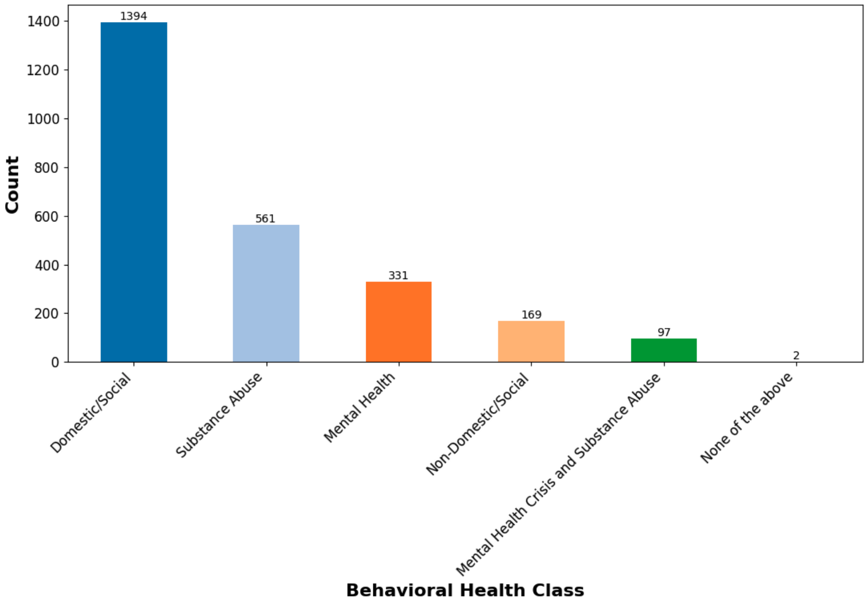


Figure 2. Distribution of behavioral health classes.

As shown in the Figure 3, the LSTM is much more better at identifying the “domestic/social” and “substance abuse” classes compared to other classes, whereas the BiLSTM model performs comparatively good across the classes other than the “nondomestic/social” class, shown in Figure 4. This could be due to the low availability of class samples in the test set. Figure 5 shows that the CNN+LSTM model performed poorly in identifying classes other than “domestic/social” and “substance abuse”.

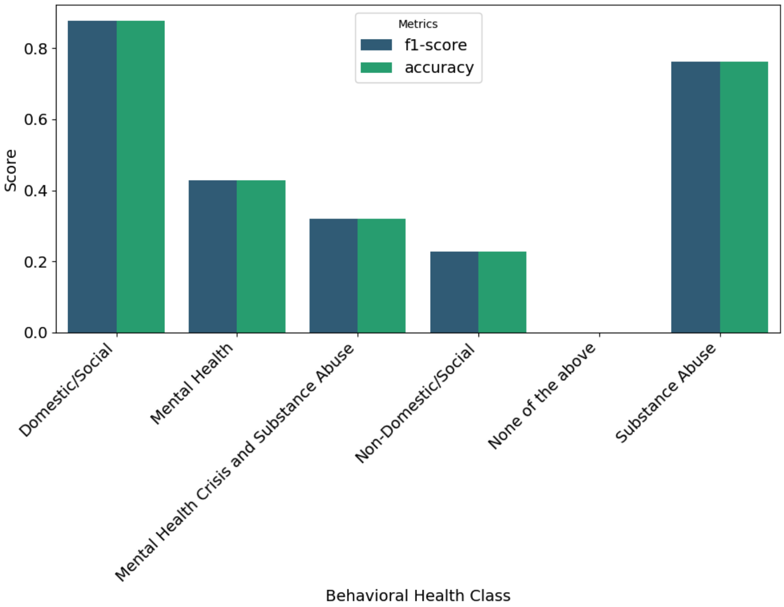


Figure 3. Class-wise performance for LSTM.

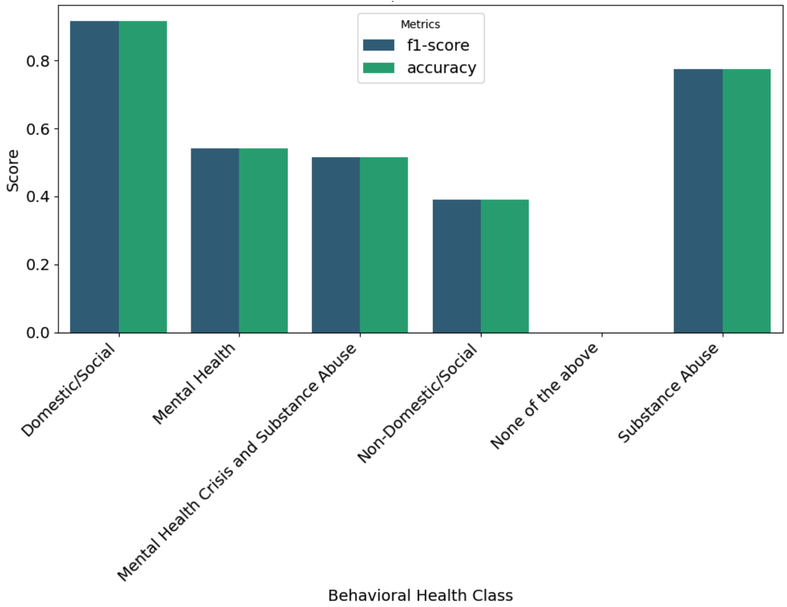


Figure 4. Class-wise performance for BiLSTM.

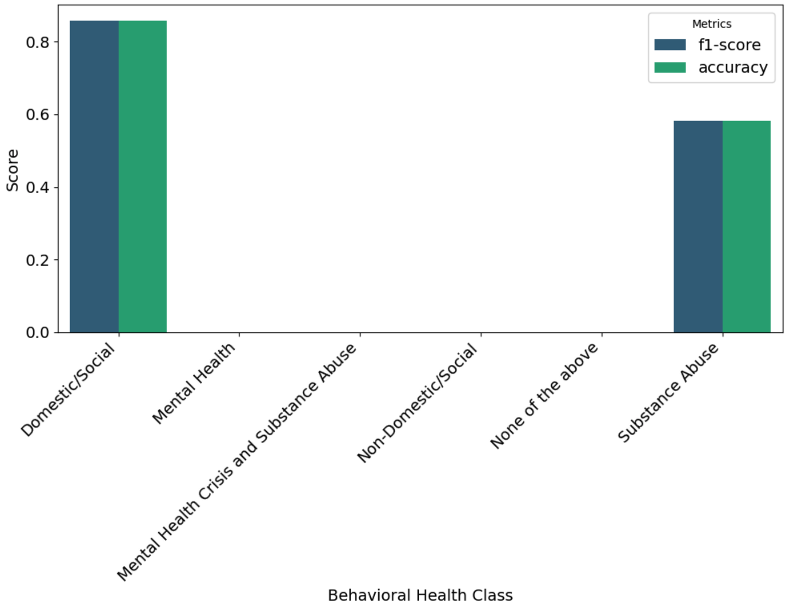


Figure 5. Class-wise performance for CNN+LSTM.

Figure 6 shows that RoBERTa accurately captured the contextual representation of the embeddings, as demonstrated by its ability to distinguish between behavioral health classes. The lightweight framework had high F1-scores and accuracy in well-represented classes including “domestic/social” and “substance abuse”, showing excellent results in these classes. However, the results for classes with low representation, such as “mental health” and “mental health crisis and substance abuse”, show significantly lower F1-scores compared to accuracy, indicating difficulties in dealing with class imbalances.

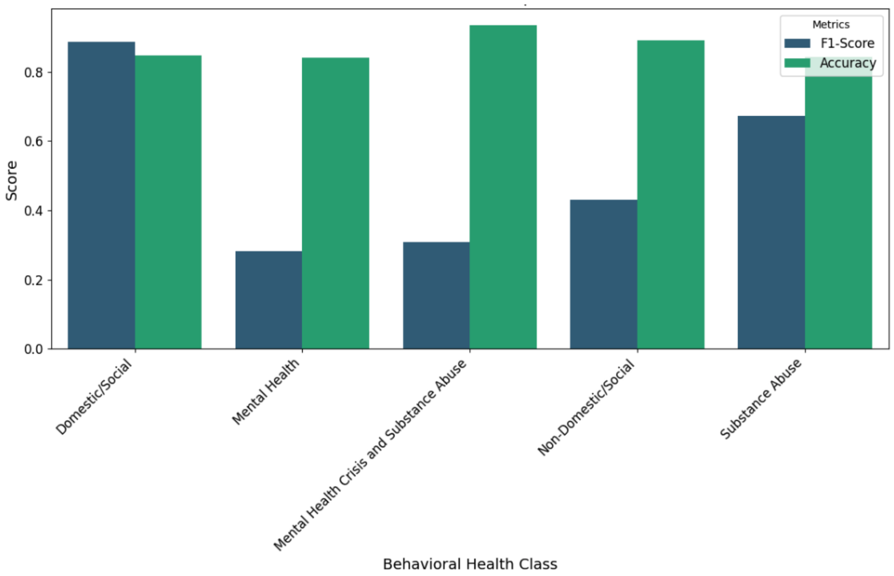


Figure 6. Class-wise performance for RoBERTa + multi-label classifier.

The lightweight framework, enhanced by feature embeddings from the fine-tuned LLM, performed well in distinguishing the "domestic/social" class. However, as shown in Figure 7, it performed fairly in the "substance abuse" class and struggled with other classes. This low performance in less prevalent classes can be attributed to the nature of LLaMA, which is designed primarily for text generation tasks rather than encoding fine-grained contextual embeddings for classification. In contrast, RoBERTa, which was specifically built for encoding tasks, is better suited to capture nuanced distinctions across classes, making it more effective for balanced multi-label classification. The findings in Figures 3–7 and Table 4 suggest that, while fine-tuned LLaMA embeddings benefit dominant classes, they are less effective at handling the peculiarities of imbalanced datasets.

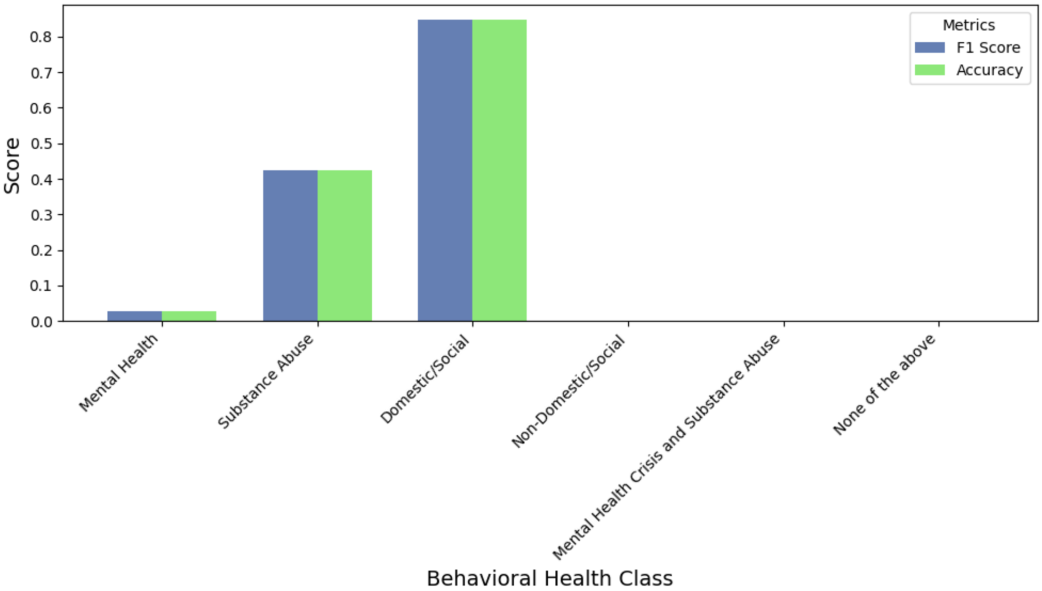


Figure 7. Class-wise performance for fine-tuned LLM + multi-label classifier.

5.3. t-SNE plot analysis of the classes (RQ3)

First, we tried to address the class imbalance. We applied oversampling to the extracted features and labels, ensuring balanced representation. Figure 8 depicts the reduced embeddings for test data across classes, giving insight on the distinction of class features. Each point in Figure 8 represents a sample and the colors indicate certain classes, such as "domestic/social" (blue) and "substance abuse" (brown). Some classes, including "domestic/social" and "mental health crisis and substance abuse" (green), form distinct clusters, demonstrating the model's capacity to distinguish attributes. However, overlaps between classes such as "mental health" (orange) and "substance abuse" (brown) indicate feature similarity or ambiguity, which affects classification. While oversampling corrected class imbalance during training, more improvements in feature distinctiveness are necessary for better separation and performance.

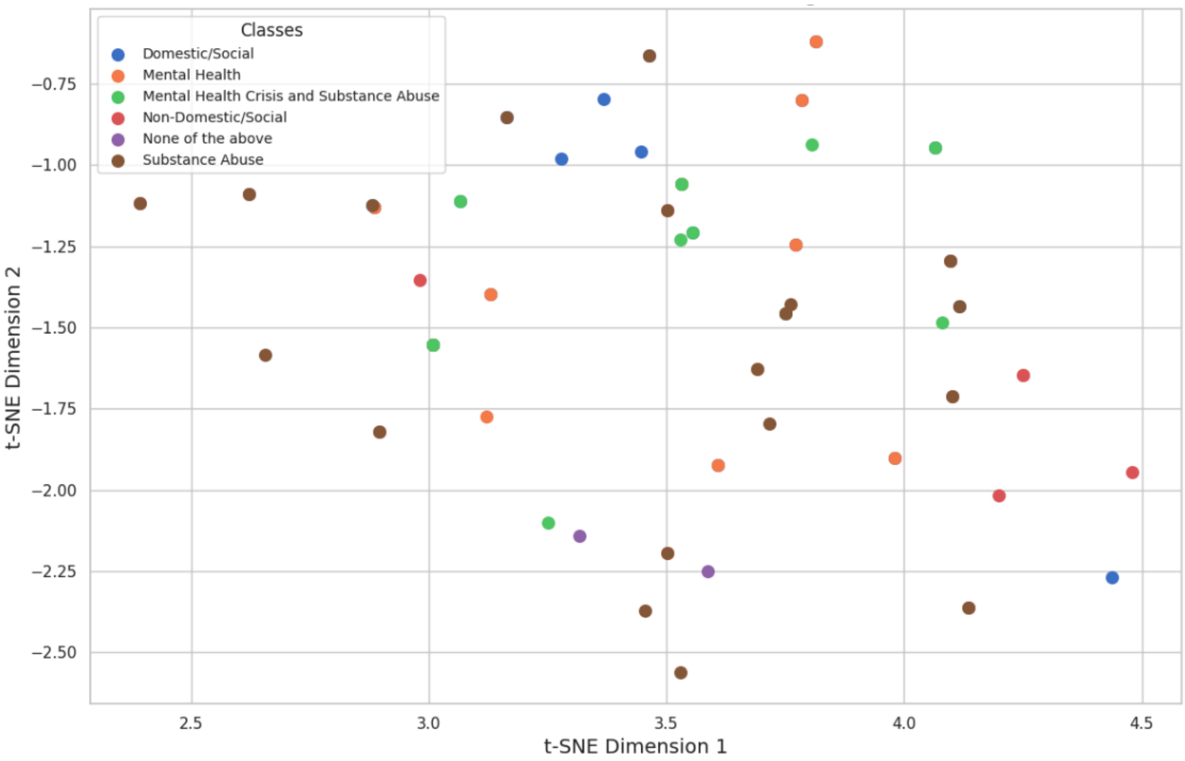
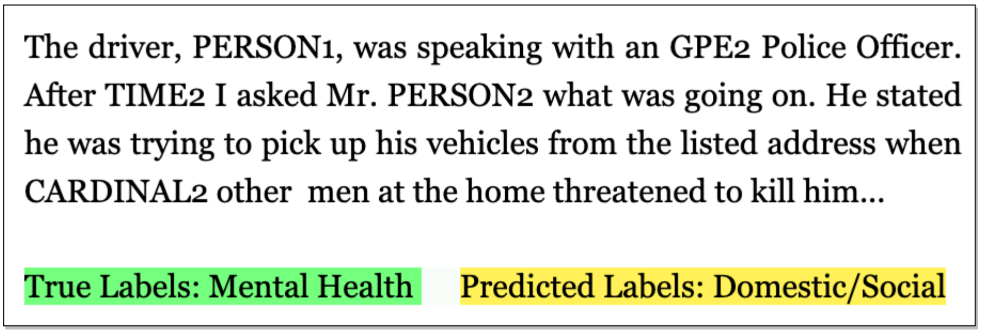


Figure 8. Visualizing the relationship among the classes.

5.4. Error analysis (RQ4)

We will look into the errors in the multi-label classifier enhanced by fine-tuned LLM for behavioral health analysis. There are some test cases incorrectly identified from our experiment.

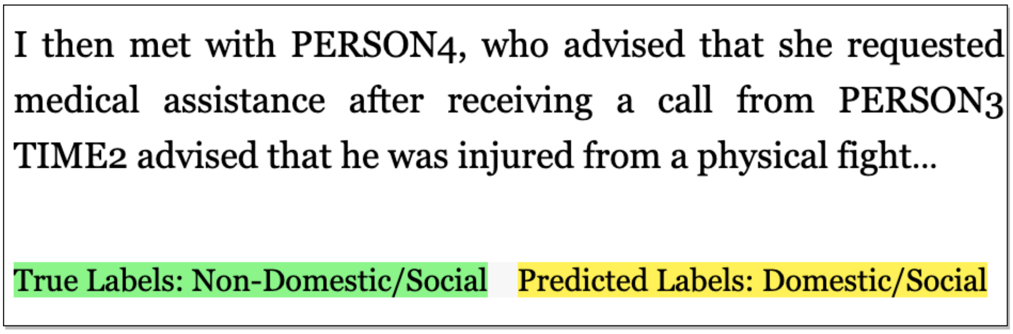
Figures 9 and 10 demonstrate that the LLM + classifier could not identify the correct class. For the first sample, it misclassified the narrative report as "domestic/social" rather than the correct label: "mental health". In addition, it wrongly identified a report from the "nondomestic/social" class as "domestic/social. It is possible that training the classifier for additional epochs may improve its ability to distinguish across classes.



The driver, PERSON1, was speaking with an GPE2 Police Officer. After TIME2 I asked Mr. PERSON2 what was going on. He stated he was trying to pick up his vehicles from the listed address when CARDINAL2 other men at the home threatened to kill him...

True Labels: Mental Health Predicted Labels: Domestic/Social

Figure 9. Incorrectly classified as “domestic/social”.



I then met with PERSON4, who advised that she requested medical assistance after receiving a call from PERSON3 TIME2 advised that he was injured from a physical fight...

True Labels: Non-Domestic/Social Predicted Labels: Domestic/Social

Figure 10. Incorrectly classified as “domestic/social”.

6. Conclusions

In this work, we provide an approach that employs transformer-based models as an encoder to extract feature embeddings for training fully connected layers for multi-label classification. Our results demonstrate that a lightweight network enhanced with parameter-frozen transformer-based LLMs, while having fewer layers and shorter sequence lengths, can match or rival traditional and hybrid models such as LSTM, BiLSTM, CNN+LSTM in terms of contextual embedding quality and classification performance. In this case, the RoBERTa + classifier achieved high F1-scores and accuracy in a number of behavioral health classes, demonstrating its usefulness and robustness despite its streamlined architecture. Furthermore, our findings show that accuracy alone is insufficient for evaluating model performance in sensitive areas such as behavioral health, where a model may perform well in some classes but fail to provide a complete representation of the dataset.

6.1. Discussion and limitations

We evaluated several models alongside a trainable framework enhanced with transformer-based models, providing crucial insights into the efficacy of our proposed approach. Our study was constrained by a number of variables. First, class imbalance was a significant concern, which could affect the accuracy of classifications and stability of the transformer-based LLMs, as shown in Section 5.4. To address this, we utilized an oversampling technique in which positive samples from each class were reproduced in proportion to the number of negative samples, resulting in a balanced representation across the dataset. For better computation performance, we applied this to extracted embeddings rather than individual sets (training and testing). This eliminated the need for repeated tokenization. In addition, we refined the instruction sets to improve clarity and granularity, which reduced ambiguity and increased model output stability to some degree. These improvements increased the ability of the model to learn nuanced patterns and improve robustness and performance by fairly representing all classes across the dataset. Also due to hardware constraints, we lowered some parameters to their minimal values to reduce computing demands and training time. With better hardware, we would like to explore more powerful transformer-based models and fully use their capabilities to generate richer feature embeddings.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

Md Abdullah Al Hafiz Khan is an editorial board member for Applied Computing and Intelligence and was not involved in the editorial review and the decision to publish this article.

References

1. Open AI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, et al., GPT-4 technical report, arXiv: 2303.08774. https://doi.org/10.48550/arXiv.2303.08774

2. A. Azmee, M. Brown, M. Khan, D. Thomas, Y. Pei, M. Nandan, Domain-enhanced attention enabled deep network for behavioral health identification from 911 narratives, *Proceedings of IEEE International Conference on Big Data*, 2023, 5723–5732. https://doi.org/10.1109/BigData59044.2023.10386126

3. A. Azmee, M. Murikipudi, M. Al Hafiz Khan, Yong Pei, Sentence level analysis for detecting mental health causes using social media posts, *Proceedings of IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2023, 1388–1393. https://doi.org/10.1109/COMPSAC57700.2023.00211

4. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, **5** (2017), 135–146. https://doi.org/10.1162/tacl_a_00051

5. M. Brown, M. Al Hafiz Khan, D. Thomas, Y. Pei, M. Nandan, Detection of behavioral health cases from sensitive police officer narratives, *Proceedings of IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2023, 1398–1403. https://doi.org/10.1109/COMPSAC57700.2023.00213

6. K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et al., Learning phrase representations using RNN encoder-decoder for statistical machine translation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1724–1734. https://doi.org/10.3115/v1/D14-1179

7. J. Chung, J. Teo, Mental health prediction using machine learning: taxonomy, applications, and challenges, *Appl. Comput. Intell. S.*, **2022** (2022), 970363. https://doi.org/10.1155/2022/9970363

8. J. Devlin, M. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

9. B. Figg, Substance abuse and mental health services administration <www.samhsa.gov>, *J. Cons. Health Internet*, **22** (2018), 253–262. https://doi.org/10.1080/15398285.2018.1513760

10. *Froedtert and the Medical College of Wisconsin, Get the facts about behavioral health*, Froedtert & the Medical College of Wisconsin health network, 2024. Available from: `https://www.froedtert.com/behavioral-health/understanding`.

11. *GAO, Behavioral health: available workforce information and federal actions to help recruit and retain providers*, US Government Accountability Office, 2022. Available from: `https://www.gao.gov/products/gao-23-105250`.

12. T. Hashmi, D. Thomas, M. Nandan, First responders, mental health, dispatch coding, COVID-19: crisis within a crisis, *Journal of Emergency Management*, **21** (2023), 233–240. https://doi.org/10.5055/jem.0664

13. P. He, X. Liu, J. Gao, W. Chen, Deberta: decoding-enhanced bert with disentangled attention, arXiv: 2006.03654. https://doi.org/10.48550/arXiv.2006.03654

14. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

15. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al., Lora: low-rank adaptation of large language models, arXiv: 2106.09685. https://doi.org/10.48550/arXiv.2106.09685

16. P. Jain, K. Srinivas, A. Vichare, Depression and suicide analysis using machine learning and NLP, *J. Phys.: Conf. Ser.*, **2161** (2022), 012034. https://doi.org/10.1088/1742-6596/2161/1/012034

17. A. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. Chaplot, D. de las Casas, et al., Mistral 7B, arXiv: 2310.06825. https://doi.org/10.48550/arXiv.2310.06825

18. G. Karystianis, R. Cabral, S. Han, J. Poon, T. Butler, Utilizing text mining, data linkage and deep learning in police and health records to predict future offenses in family and domestic violence, *Front. Digit. Health*, **3** (2021), 602683. https://doi.org/10.3389/fdgth.2021.602683

19. G. Karystianis, A. Adily, P. Schofield, H. Wand, W. Lukmanjaya, I. Buchan, et al., Surveillance of domestic violence using text mining outputs from Australian police records, *Front. Psychiatry*, **12** (2022), 787792. https://doi.org/10.3389/fpsyt.2021.787792

20. J. Kim, J. Lee, E. Park, J. Han, A deep learning model for detecting mental illness from user content on social media, *Sci. Rep.*, **10** (2020), 11846. https://doi.org/10.1038/s41598-020-68764-y

21. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: a lite BERT for self-supervised learning of language representations, arXiv: 1909.11942. https://doi.org/10.48550/arXiv.1909.11942

22. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., RoBERTa: a robustly optimized BERT pretraining approach, arXiv: 1907.11692. https://doi.org/10.48550/arXiv.1907.11692

23. H. Lu, L. Ehwerhemuepha, C. Rakovski, A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance, *BMC Med. Res. Methodol.*, **22** (2022), 181. https://doi.org/10.1186/s12874-022-01665-y

24. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv: 1301.3781. https://doi.org/10.48550/arXiv.1301.3781

25. T. Munkhdalai, H. Yu, Neural semantic encoders, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, 397–407.

26. *R. Neusteter, M. O'Toole, M. Khogali, A. Rad, F. Wunschel, S. Scaffidi, et al., Understanding police enforcement*, Vera Institute of Justice, 2020. Available from: `https://www.vera.org/publications/understanding-police-enforcement-911-analysis`.

27. K. O'Shea, R. Nash, An introduction to convolutional neural networks, arXiv: 1511.08458. https://doi.org/10.48550/arXiv.1511.08458

28. Z. Pang, Z. Xie, Y. Man, Y. Wang, Frozen transformers in language models are effective visual encoder layers, arXiv: 2310.12973. https://doi.org/10.48550/arXiv.2310.12973

29. J. Pennington, R. Socher, C. Manning, GloVe: global vectors for word representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1532–1543. https://doi.org/10.3115/v1/D14-1162

30. M. Schuster, K. Paliwal, Bidirectional recurrent neural networks, *IEEE T. Signal Proces.*, **45** (1997), 2673–2681. https://doi.org/10.1109/78.650093

31. A. Shestov, R. Levichev, R. Mussabayev, E. Maslov, A. Cheshkov, P. Zadorozhny, Finetuning large language models for vulnerability detection, arXiv: 2401.17010. https://doi.org/10.48550/arXiv.2401.17010

32. K. Singhal, S. Azizi, T. Tu, S. Sara Mahdavi, J. Wei, H. Chung, et al., Large language models encode clinical knowledge, *Nature*, **620** (2023), 172–180. https://doi.org/10.1038/s41586-023-06291-2

33. K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, et al., Towards expert-level medical question answering with large language models, arXiv: 2305.09617. https://doi.org/10.48550/arXiv.2305.09617

34. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, et al., Llama: open and efficient foundation language models, arXiv: 2302.13971. https://doi.org/10.48550/arXiv.2302.13971

35. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, et al., Attention is all you need, arXiv: 1706.03762. https://doi.org/10.48550/arXiv.1706.03762

36. B. Victor, B. Perron, R. Sokol, L. Fedina, J. Ryan, Automated identification of domestic violence in written child welfare records: leveraging text mining and machine learning to enhance social work research and evaluation, *J. Soc. Soc. Work Res.*, **12** (2021), 631–655. https://doi.org/10.1086/712734

37. X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, et al., Mental-llm: leveraging large language models for mental health prediction via online text data, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, **8** (2024), 31. https://doi.org/10.1145/3643540

38. K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, S. Ananiadou, MentaLLaMA: interpretable mental health analysis on social media with large language models, *Proceedings of the ACM Web Conference*, 2024, 4489–4500. https://doi.org/10.1145/3589334.3648137

39. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. Le, XLNet: generalized autoregressive pretraining for language understanding, arXiv: 1906.08237. https://doi.org/10.48550/arXiv.1906.08237

40. W. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, et al., A survey of large language models, arXiv: 2303.18223. https://doi.org/10.48550/arXiv.2303.18223

41. J. Zheng, H. Hong, X. Wang, J. Su, Y. Liang, S. Wu, Fine-tuning large language models for domain-specific machine translation, arXiv: 2402.15061. https://doi.org/10.48550/arXiv.2402.15061