Transformer vs. LSTM Approaches for Multi-Label Classification of Youth Mental Health News Articles

Juyeong Park, jpark5@stevens.edu   Shotitouch Tuangcharoentip, stuangch@stevens.edu

1 Introduction

This project compares Transformer-based models and LSTM networks for multi-label classification of youth mental health news related to YouTube and Snapchat. It aims to determine which architecture more effectively captures the complex and overlapping language patterns in mental health reporting.

The dataset is built through systematic preprocessing steps, including filtering, cleaning, and tokenization of collected news articles. To improve labeling efficiency and quality, LLM-assisted annotation is used to assign multiple mental health topic labels such as anxiety, depression, and addiction.

Two models are developed—an LSTM model and a Transformer model fine-tuned from BERT—for performance comparison. Both are evaluated using standard multi-label metrics, including subset accuracy, precision, recall, F1-score, and Hamming loss.

In terms of progress, the data preparation has been completed with the help of an LLM-based system and has already been trained on multiple models. The LSTM models have been tested and improved for a total of 3 versions, each one better than the last. In addition, the first refined version of the Transformer model was also completed and tested on datasets with some variations in its data preparation methods.

The LSTM model performed at a lower level overall compared to Transformer models as expected. So far, LSTM achieved a Hamming Loss value of 0.1720 while the Transformer model achieved the loss of 0.0866. Both models still leave significant room for improvement.

The most notable challenge encountered during the development was severe label imbalance. As will be shown later in this report, the frequency of each label varied widely from common to very rare. This problem led both LSTM and Transformer models to misclassify low-frequency labels. Weighted sampling was implemented to reduce this effect, which improved recall for rare labels but did not fully eliminate the issue.

1.1 Related Work

Multi-label text classification has been actively studied across social, medical, and behavioral health domains.

MULTIWD (Garg et al., 2024) [1] classified Reddit posts into multiple wellness dimensions related to mental and social well-being using multi-label models. The authors faced challenges such as unclear category definitions, overlapping labels, class imbalance, and short informal posts that lacked sufficient context for accurate classification.

Identification of Social Determinants of Health Using Multi-Label Classification of Clinical Notes (Stemerman et al., 2021) [4] This study applied multi-label text classification on clinical notes to identify social and mental health determinants using models such as SVM, Random Forest, and Bi-LSTM. The Bi-LSTM model achieved the highest AUC-ROC and lowest Hamming loss, demonstrating strong ability to capture contextual dependencies in unstructured medical text. The authors highlighted challenges similar to this project, including label imbalance, lexical diversity, and overlapping categories, and emphasized the importance of robust multi-label metrics such as Hamming loss and AUC-ROC.

Multi-Label News Classification Using Bi-LSTM (Goel , Samantaray, 2021) [2] This study implemented a Bi-LSTM model for multi-label news categorization, focusing on identifying multiple topics within short text segments. The model effectively captured sequential dependencies and contextual relationships in news headlines, outperforming traditional classifiers such as SVM and Naïve Bayes. The authors noted that while Bi-LSTM achieved strong accuracy, class imbalance and limited context length remained key challenges—similar to those observed in this project's LSTM experiments.

A Transformer-Driven Framework for Multi-Label Behavioral Health Classification in Police Narratives (Nweke et al., 2024) [3] This work applied a Transformer-based framework to classify police incident narratives into multiple behavioral health categories. By leveraging contextual embeddings from pre-trained Transformers, the model achieved superior performance in handling overlapping and co-occurring mental-health-related labels compared to traditional and recurrent models.

2 Problem Formulation

The task is formulated as a **multi-label text classification** problem rather than a multi-class classification task. Each news article serves as a single input instance, and the goal is to predict multiple mental-health topics that may simultaneously appear within the same article. Formally, given an article represented by a text sequence x_i, the model aims to learn a function $f : x_i \rightarrow y_i$, where $y_i \in \{0, 1\}^{16}$ is a binary vector indicating the presence (1) or absence (0) of each topic.

The dataset defines sixteen possible mental-health topics: *access_to_care, anxiety, apps_telehealth, bullying_cyberbullying, depression, education_stress, family_peers, pandemic_stress, school_policy, self_esteem_bodyimage, stigma_awareness, substance_use, suicide_prevention, therapy_CBT, violence_safety, and youth_wellbeing.*

Each label vector $\mathbf{y}_i$ is represented as a **multi-hot encoding** of dimension 16×1, where each position corresponds to one topic category:

$$\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{i16}], \quad y_{ij} \in \{0, 1\}.$$

A value of $y_{ij} = 1$ indicates that the j^{th} topic is discussed in article i, while $y_{ij} = 0$ denotes its absence. The model outputs predicted probabilities $\hat{y}_{ij}$ for each topic using a sigmoid activation function, and a threshold τ (typically 0.5) is applied to convert probabilities into binary label predictions.

This formulation allows the model to assign multiple mental-health topics to each article, reflecting the overlapping and co-occurring nature of issues discussed in youth mental health reporting.

3 Methods

3.1 LSTM Model

3.1.1 Method and Implementation

The Long Short-Term Memory (LSTM) network was chosen as a sequence-based baseline to model the temporal and syntactic structure of the textual data. The architecture consisted of an **Embedding layer** (128-dimensional), followed by an **LSTM layer** with 64 hidden units, and a **Dense output layer** with sigmoid activation for multi-label prediction. In the final version, an additional **Global Average Pooling 1D layer** was applied to the embedding output, and its result was concatenated with the LSTM output to combine local sequential features and global contextual information.

The models were trained using the Adam optimizer with a learning rate of 0.001 and the Binary Cross-Entropy loss function. Three versions were developed:

- **LSTM 1:** Baseline model trained without any weighting.
- **LSTM 2:** Added sample weighting inversely proportional to label frequency to handle imbalance.
- **LSTM 3:** Introduced Global Average Pooling and a softened weighting scheme proportional to $\sqrt{1/frequency}$.

The networks were implemented in TensorFlow/Keras and trained on padded integer sequences representing tokenized text, with batch normalization applied implicitly through standardization.

3.1.2 Rationale for Method Selection

The LSTM model was selected because it effectively captures sequential dependencies and contextual order within text data, which are essential in understanding nuanced mental health news articles. Unlike bag-of-words or feed-forward models, LSTMs retain contextual memory through gating mechanisms, allowing the network to model how topic cues evolve within an article. Adding the Global Average Pooling layer enabled the model to retain global semantic information that complements the LSTM's localized pattern learning. This design is suitable for the dataset because the text often contains long phrases and co-occurring emotional or social indicators that require sequential understanding.

3.1.3 Challenges and Solutions

A major challenge in the LSTM model was **class imbalance**, where common labels such as *youth_wellbeing* appeared in over 70% of articles, while rare labels such as *education_stress* occurred in less than 3%. This imbalance caused the baseline LSTM to predict frequent labels more reliably while neglecting rare ones. Weighted sampling was introduced to compensate, but strong weighting destabilized learning and reduced overall accuracy. A moderated weighting scheme (square-root adjustment) was then adopted, yielding a better trade-off between recall and precision. Another challenge was **long text length**; articles contained thousands of tokens, requiring truncation to maintain feasible computation. Padding and truncation ensured consistent input shape, while Global Pooling reduced the dependency on sequence length and improved generalization.

3.2 Transformer Model

3.2.1 Method and Implementation

The Transformer-based model used a pre-trained **BERT Base Uncased** architecture, which consists of 12 encoder layers, each with 768 hidden units and 12 self-attention heads. This model was fine-tuned using the `BertForSequenceClassification` implementation from the Hugging Face Transformers library. The architecture included:

- The BERT encoder to generate contextual token embeddings.
- A dropout layer to prevent overfitting.
- A classification head applied to the [CLS] token, mapping the 768-dimensional representation to 16 sigmoid-activated outputs.

The model was fine-tuned for 3 epochs with a batch size of 8, using the AdamW optimizer (learning rate = 2×10^{-5}) and Binary Cross-Entropy with Logits Loss. The input text was tokenized with a maximum sequence length of 128, padded or truncated as needed, and formatted as PyTorch tensors including `input_ids`, `attention_mask`, and `labels`.

3.2.2 Rationale for Method Selection

The Transformer model was chosen for its ability to capture long-range dependencies and contextual relationships across the entire text through the self-attention mechanism. Unlike LSTMs, which process text sequentially, BERT's bidirectional attention enables it to model relationships between all words simultaneously, making it particularly effective for multi-label classification where overlapping themes often occur. Since the dataset contains nuanced and co-occurring mental-health topics, BERT provides richer semantic understanding than traditional sequential architectures. Additionally, fine-tuning a pre-trained language model reduces the need for large domain-specific data and improves performance on imbalanced and small datasets.

3.2.3 Challenges and Solutions

One key challenge was the limitation on maximum sequence length (128 tokens), which may truncate important information in longer articles. Future experiments will address this by extending the sequence limit to 256 or 512. Another issue was **class imbalance**; unlike the LSTM, no sample weighting was applied in BERT training, yet the model still performed better due to its robust contextual representations. Computation cost was also a concern—fine-tuning Transformers requires more memory and training time. This was managed by using smaller batch sizes and efficient data loading through PyTorch's `DataLoader`. Overall, the Transformer model proved more stable and effective, outperforming the LSTM models even without explicit balancing techniques.

4 Datasets and Experiments

The dataset used in this project, `news_combined_final_with_id_bool.csv`, consists of 2,002 news articles collected from **YouTube** and **Snapchat** in 2024. Each article discusses topics related to youth mental health and is annotated with one or more topical labels such as *anxiety*, *depression*, and *access_to_care*. These annotations were created through semi-automated large language model (LLM)-assisted labeling to ensure consistency across the dataset.

After data cleaning and label refinement, the final dataset includes **16 active labels**, excluding *social_media* and *workplace_stress* due to their extreme imbalance. The label frequencies vary substantially—from only 57 instances of *education_stress* to 1,421 instances of *youth_wellbeing*—reflecting the strong real-world skew often observed in social and mental-health reporting. This imbalance presents a major challenge for machine learning models, as it biases predictions toward high-frequency topics.

4.1 Label Imbalance Analysis

The frequency distribution of the labels is shown in Fig. 1 as well as in Table 1. It is evident that certain categories such as *youth_wellbeing*, *access_to_care*, and *family_peers* dominate the dataset, while others like *education_stress* and *school_policy* are severely underrepresented.

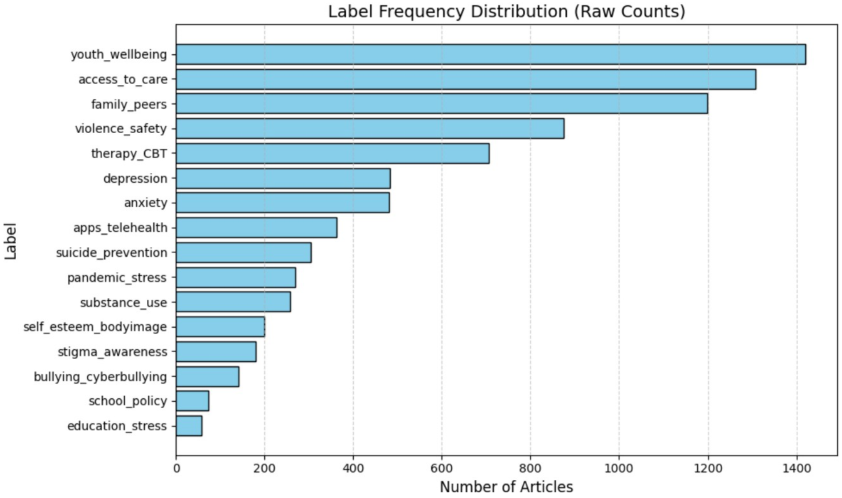

Figure 1: Bar chart of label frequencies across all 16 mental health topics.

This analysis confirms that the dataset is highly imbalanced, motivating the use of techniques such as weighted loss functions and class-balancing strategies in the LSTM experiments, while exploring the inherent robustness of Transformer architectures in handling skewed label distributions.

4.2 4.2 LSTM Model

4.2.1 Data Preprocessing

A complete preprocessing pipeline was developed before model training:

- **Column filtering:** Removed irrelevant metadata fields (`id`, `title`, `provider`, `publication_date`).

Table 1: Frequency of each label in the dataset.

Label	Count	Percentage (%)
access_to_care	1307	65.3
anxiety	481	24.0
apps_telehealth	363	18.1
bullying_cyberbullying	141	7.0
depression	482	24.1
education_stress	57	2.8
family_peers	1198	59.8
pandemic_stress	269	13.4
school_policy	74	3.7
self_esteem_bodyimage	199	9.9
stigma_awareness	179	8.9
substance_use	258	12.9
suicide_prevention	304	15.2
therapy_CBT	705	35.2
violence_safety	875	43.7
youth_wellbeing	1421	71.0

- **Text cleaning:** Removed HTML tags, punctuation, and numbers; normalized case and whitespace; eliminated stop words; and applied lemmatization using WordNet.
- **Label encoding:** Converted the topic tags into multi-hot vectors of 16 binary elements.
- **Train/validation/test split:** Randomly divided into 80% training, 10% validation, and 10% testing.
- **Tokenization and padding:** Built a vocabulary (10,000 tokens) using Keras' `Tokenizer`, converted text to integer sequences, and padded them to a uniform sequence length (3,000–4,000 tokens).
- **Sample weighting:** Calculated inverse-frequency weights to balance rare vs. frequent labels during training.

4.2.2 Experiments Conducted

Three LSTM-based models were trained with identical hyperparameters (embedding = 128, LSTM units = 64, batch = 32, epochs = 10–15, optimizer = Adam), but with different architectural or weighting configurations.

Table 2: LSTM Model Configurations

Model	Description	Key Features
LSTM 1	Baseline	Embedding → LSTM → Dense (sigmoid); no weighting
LSTM 2	Weighted sampling	Added class weighting (1/freq); same architecture
LSTM 3	Enhanced model	Global Average Pooling + $\sqrt{1/freq}$ weighting

All models were evaluated using subset accuracy, micro/macro/weighted precision, recall, F1-score, and Hamming loss.

4.2.3 Results and Analysis

The third version (LSTM 3) achieved the best balance between accuracy and recall, showing clear improvement over the baseline. The addition of a Global Average Pooling layer improved the macro F1 by approximately 0.1, indicating that pooling helps capture document-level meaning. Using softened weights ($\sqrt{1/freq}$) stabilized learning and reduced Hamming loss, suggesting that mild balancing was more effective than strong weighting.

4.2.4 Future Experiments

Future work will focus on:

Table 3: Performance of LSTM Variants

Metric	LSTM 1	LSTM 2	LSTM 3
Subset Accuracy	0.1452	0.0161	0.1935
Precision (Micro)	0.7240	0.6519	0.7338
Recall (Micro)	0.4747	0.5699	0.6251
F1 (Micro)	0.5735	0.6082	0.6751
F1 (Macro)	0.1572	0.1951	0.2855
F1 (Weighted)	0.3999	0.4576	0.5632
Hamming Loss	0.2019	0.2100	0.1720

- Replacing binary cross-entropy with **Focal Loss** to dynamically down-weight easy samples.
- Performing **threshold optimization** per label to better calibrate sigmoid outputs.
- Testing **hybrid pooling** (Global Max + Average Pooling) for richer text representations.

4.3 4.3 Transformer Model

4.3.1 Data Preparation

The text data used for the Transformer model was taken from the same cleaned dataset as the LSTM experiments. Each article was tokenized using the BERT tokenizer (`bert-base-uncased`) with a maximum sequence length of 128 to balance coverage and efficiency. Shorter sequences were padded and longer ones truncated to maintain uniform input length. The tokenizer produced input IDs and attention masks, which were then formatted into PyTorch tensors suitable for BERT-based training. A custom `NewsDataset` class was developed to handle the tokenized text, attention masks, and label tensors. The dataset was split into 80% training, 10% validation, and 10% testing, following the same proportions used for the LSTM model. Finally, the data was loaded into PyTorch `DataLoader` objects with a batch size of 8 to enable efficient GPU training.

4.3.2 Model Training

The model used for this experiment was **BERT Base Uncased**, which contains 12 Transformer encoder layers, each with 768 hidden units and 12 self-attention heads. This pre-trained architecture was fine-tuned for the multi-label classification task by adding a classification head on top of the base encoder. The model structure is summarized as follows:

- **BERT Base Encoder:** Provides contextualized token embeddings.
- **Dropout Layer:** Prevents overfitting during fine-tuning.
- **Classification Head:** A linear layer applied to the `[CLS]` token output to generate 16 label predictions corresponding to the mental health topics.

The model was implemented using the `BertForSequenceClassification` class from the Hugging Face Transformers library. Training used the `AdamW` optimizer with a learning rate of 2×10^{-5} and the `BCEWithLogitsLoss` function, which is appropriate for multi-label classification tasks. The model was trained for 3 epochs, evaluating on the validation set after each epoch. Both training and validation losses consistently decreased, indicating successful fine-tuning and generalization to unseen data.

4.3.3 Results and Analysis

The Transformer achieved substantially higher scores across all metrics compared to the LSTM models. Despite using no class weighting, it surpassed even the most optimized LSTM configuration (which used weighting and pooling), achieving a micro F1 of 0.8450 and a Hamming loss of 0.0866 versus 0.6751 and 0.1720, respectively. This demonstrates the Transformer's superior ability to learn rich semantic relationships and manage label imbalance inherently through its self-attention mechanism, rather than relying on external weighting strategies.

Table 4: Transformer (BERT) Evaluation Results

Metric	Value
Subset Accuracy	0.4920
Precision (Micro)	0.8738
Recall (Micro)	0.8181
F1-score (Micro)	0.8450
Precision (Macro)	0.8249
Recall (Macro)	0.6912
F1-score (Macro)	0.7416
Hamming Loss	0.0866

4.3.4 Future Experiments

Future work will focus on improving the Transformer model's performance and interpretability. Key directions include:

- Extending input sequences (256–512 tokens) to capture the full article context.
- Testing Transformer variants such as RoBERTa, DistilBERT, and DeBERTa for performance–efficiency comparison.
- Using Focal Loss or class-balanced loss to improve rare-label recall.
- Hyperparameter tuning of learning rate, dropout, and batch size.
- Per-label threshold optimization instead of a fixed 0.5 cutoff.

5 Project Management

5.1 Team Members and Responsibilities

Shotitouch Tuangcharoentip: Responsible for developing, training, and evaluating all LSTM-based models, including architecture tuning, sample weighting experiments, and ablation analysis.

Juyeong Park: Responsible for implementing and fine-tuning Transformer-based models, including BERT architecture setup, tokenization pipeline, and model evaluation.

Both team members collaboratively handled data preprocessing, LLM-assisted annotation of topic labels, and comparative result analysis between the two model families.

References

[1] Muskan Garg, Xingyi Liu, MSVPJ Sathvik, Shaina Raza, and Sunghwan Sohn. Multiwd: Multi-label wellness dimensions in social media posts. *Journal of biomedical informatics*, 150:104586, 2024.

[2] Y Kumar Goel and D Samantaray. Multi-label news classification using bi-lstm. *International Journal of Creative Research Thoughts (IJCRT)*, 9, 2021.

[3] Francis Nweke, Abm Adnan Azmee, Md Abdullah Al Hafiz Khan, Yong Pei, Dominic Thomas, and Monica Nandan. A transformer-driven framework for multi-label behavioral health classification in police narratives. *Applied Computing and Intelligence*, 4(2):234–252, 2024.

[4] Rachel Stemerman, Jaime Arguello, Jane Brice, Ashok Krishnamurthy, Mary Houston, and Rebecca Kitzmiller. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA open*, 4(3), 2021.