Topic Extraction and Depression Risk Prediction in Online Mental Health Posts

Group9: Guli Zhu, Joey Qiu, Kaixiang Jiao, Yilin Yan

2024-12-17

GitHub Repository: https://github.com/Cameron-zgl/Biostat_625_umich_final_project

---

1 Abstract

This project aims to analyze public mental health-related data to extract relevant discussion topics and predict depression risk in user posts. Using a data set of Reddit posts from mental health subreddits, the analysis was conducted in two phases:

1. Topic Extraction: Unsupervised clustering techniques were applied to identify recurring themes in posts grouped by time intervals.

2. Depression Prediction: Annotated data was used to train supervised models, including LSTM with BERT embedding, to classify posts as indicating depression risk or not.

The project incorporates real-world data, computational challenges such as text pre-processing, feature extraction, and model building, and employs advanced machine learning techniques for analysis.

2 Introduction

2.1 Background and Motivation

Mental health is a growing concern worldwide, and online platforms like Reddit provide an opportunity to study how individuals discuss their struggles. Analyzing such data can help identify key topics of discussion and provide tools to assess depression risk, which could assist in early interventions.

This study aims to develop a comprehensive two-step framework that integrates thematic analysis and predictive modeling to address the complexities of analyzing unstructured textual data. By leveraging NLP techniques and advanced machine learning algorithms, the research seeks to uncover latent themes, predict mental health conditions such as depressive tendencies, and construct robust, interpretative systems for enhanced mental health assessment.

2.2 Objective

- **Objective 1:** Identify major discussion topics within mental health-related posts.
- **Objective 2:** Develop a predictive model to classify posts based on depression risk.

2.3 Data Source

The dataset used is sourced from Zenodo:

- Link: https://zenodo.org/api/records/3941387/files-archive
- The dataset consists of Reddit posts from mental health-related subreddits, including depression and other general posts.

The dataset was preprocessed and split into training and testing sets for analysis.

3 Methods

The following flowchart (Figure 1) provides a clear and systematic overview of our approach, ensuring that all steps are logically connected and easy to follow.

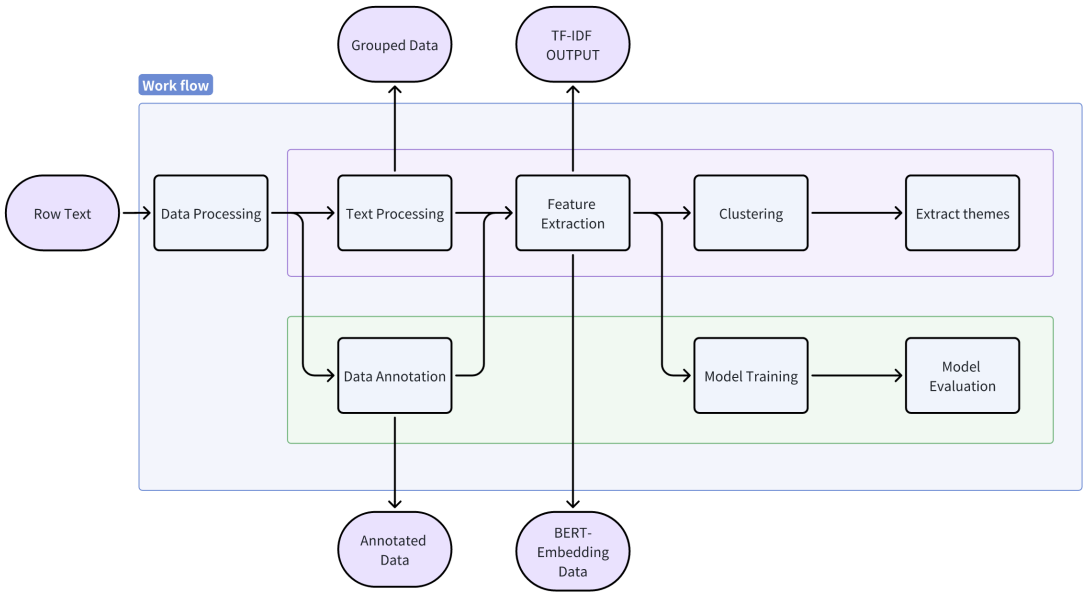

Figure 1: Flowchart of the Methods

3.1 Step1: Topic Extraction

In Step 1, topic extraction was conducted to identify prominent themes within mental health-related Reddit posts. The process commenced with text preprocessing, which involved cleaning the raw textual data through the removal of stop words, sentence tokenization, and conversion of all text to lowercase to achieve consistency and uniformity.

Subsequently, feature extraction was performed utilizing TF-IDF (Term Frequency-Inverse Document Frequency), a statistical method that transformed the preprocessed text into numerical feature representations. This approach assigned greater weights to words that appeared frequently within individual posts while maintaining lower importance for words common across the entire dataset. Following feature extraction, an unsupervised clustering algorithm, such as K-means clustering, was applied to the TF-IDF features to systematically group semantically similar posts into distinct clusters.

To facilitate cluster interpretation, representative keywords were extracted from each group, enabling meaningful labeling and thematic analysis. Topic keywords for each cluster were saved in *merged_clusters_output.csv*. This structured methodology successfully identified recurrent discussion topics, providing a robust foundation for subsequent depression risk prediction analysis.

3.2 Step2: Depression Prediction

Step2 mainly focused on depression prediction. In this step, annotated data was used to train supervised models, including Logistic Regression model and LSTM model with BERT embedding, to classify posts as indicating depression risks.

We first used the combination of VADER Model and Transformers Model (from Hugging Face) to analyze the sentiment of the text and label them, where the threshold could be adjusted according to specific needs. Among the text data, 48159 items were annotated as positive examples and 186503 items were marked as negative examples. Then we converted text data into numerical TF-IDF features for subsequent machine learning modeling.

Logistic Regression model was used to analyze baseline performance and prediction. Then four indicators (Precision, Recall, F1-score, Support) were calculated to evaluate the effectiveness of the model, where:

- Precision: The proportion of positive classes that are truly positive in the prediction.
- Recall: The proportion of all positive classes that have been successfully predicted as positive.
- F1 Score: The harmonic average of precision and recall.
- Support: The actual number of samples for each category in the test set.

For advanced modeling, We developed a BertLSTMClassifier model that combines the contextual embedding capabilities of a pre-trained BERT model with the sequence modeling strengths of an LSTM layer to perform a text classification task. The architecture was designed to effectively capture both the rich semantic features of the input text and the long-term dependencies within sequences.

1.BERT Component: We employed the BERT model as the backbone of our architecture to extract high-dimensional contextualized representations from the input text. Specifically:

- **BertEmbeddings:** Token embeddings and positional embeddings were processed and normalized to provide the initial input to the transformer layers.
- **BertEncoder:** The embeddings were passed through multiple transformer layers, producing a sequence of hidden states with a dimensionality of 768.
- **BertPooler:** The pooled output representation from BERT was further refined for downstream tasks.

2.LSTM Layer: To model sequential dependencies and reduce the dimensionality of the BERT output, we added an LSTM layer. The LSTM takes the 768-dimensional hidden states from BERT as input and transforms them into a 128-dimensional output, capturing sequential relationships in the data.

3.Linear Output Layer: The final LSTM hidden state is passed through a fully connected linear layer, which outputs a 2-dimensional vector for binary classification.

The resulting model contains a total of 109.9 million parameters, with most parameters coming from the BERT component. The forward and backward passes consume approximately 83.67 MB of memory, and the model's overall estimated size is 523.44 MB.

By integrating BERT with an LSTM layer, we achieved a model architecture that combines robust contextual feature extraction with sequential pattern recognition. This design allows the model to effectively perform text classification tasks, providing a strong foundation for capturing both semantic meaning and structural dependencies within the text. During the modeling process, we also observed that the combined model of TF-IDF and LSTM yielded a prediction accuracy of only 40-50%, which is insufficient for practical use.

Eventually, we could predict the potential depression risk of text data by using these two models.

4 Results

4.1 Step1: Topic Extraction

Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) was employed to reduce the high-dimensional feature space to two dimensions. A scatter plot was generated to visualize

the clustering results, providing an intuitive representation of the group structure within the data. Figure 2 presented a example of cluster for 2020-1.

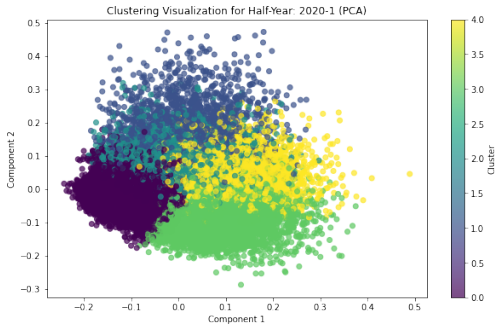

Figure 2: Example of Clustering

The sample output was presented in the GitHub Repository, and here is an example for clustering of Jan.2020.

Subreddit	Cluster	Top_10_Unigrams	Top_10_Bigrams
depression	0	depression, help, feel, get, like, know, day, life, time, anyone	feel like, anyone else, mental health, wan na, gon na, side effect, amp x200b, need help, depression anxiety, get better
	1	im, dont, cant, feel, like, ive, want, know, life, really	dont know, dont want, feel like, like im, im tired, im going, wan na, know im, think im, gon na
	2	want, fucking, hate, life, die, know, tired, ca, anymore, people	want die, wan na, feel like, want live, want kill, gon na, fucking hate, really want, hate life, get better
	3	year, friend, like, time, know, get, life, feel, really, even	feel like, high school, year ago, year old, last year, every day, best friend, mental health, get better, even though
	4	feel, like, know, want, feeling, people, even, really, life, make	feel like, make feel, feeling like, feel way, anyone else, feel alone, feel empty, feel better, know feel, want feel

Table 1: Cluster for Jan.2020

4.2 Step2: Depression Prediction

The accuracy and F1 score are presented in the following table. It can be observed that the LSTM+BERT model achieves higher accuracy and F1 score, indicating its superior performance.

Model	Accuracy	F1-Score
Logistic Regression	85.40%	84.70%
LSTM + BERT	91.20%	90.50%

Table 2: Model Performance

The loss curves for the training and validation phases of the LSTM model are shown below. It can be

observed that both curves exhibit a downward trend and converge approximately at epoch 9, leading to the final model.

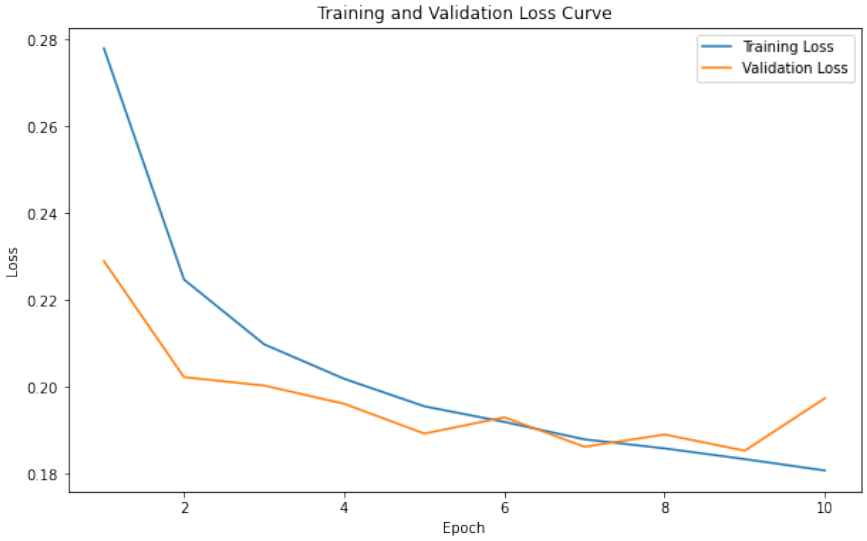

Figure 3: Loss Curve for Training and Validation Phases

5 Conclusion

This project successfully achieved its two objectives:

1. Topic Extraction: Key discussion themes were identified from Reddit posts using clustering techniques and keyword extraction.

2. Depression Prediction: The LSTM model with BERT embedding achieved strong performance in predicting posts with depression risk.

5.1 Computational Challenges

- High-dimensional text data required pre-processing and dimensionality reduction.
- Implementing BERT-based embedding for improved prediction accuracy required GPU computation.

5.2 Future Work

- Expand the dataset to include posts from broader online forums.
- Use fine-tuned BERT models for sentiment and depression prediction to further improve accuracy.

6 References

1. Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M., & De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 79–88.Available at: https://dl.acm.org/doi/abs/10.1145/3287560.3287587?casa_token=WQYwHzKDvukAAAAA:bf1RXdTX1ePljY5nDe7LiA5WRQftghq-buNbskmwbP_-6qu3lOUudWQPjSfp5GhCKO6BROjUI7FwOA (Accessed: 22 November 2024).