ENHANCED TEXTUAL CLASSIFICATION FOR SUICIDAL CONTENT DETECTION USING RECURRENT NEURAL NETWORKS AND ATTENTION MECHANISMS

JIAYING HOU [HJIAYING@SEAS.UPENN.EDU]

ABSTRACT. This project delves into the exploration and assessment of recurrent neural network (RNN) architectures for the classification of textual content derived from Reddit's "r/SuicideWatch" and "r/teenagers" subreddits [1]. This investigation focuses on the implementation of machine learning models to discern posts expressing suicidal ideation from those unrelated to such concerns. Through meticulous data preprocessing, feature engineering, model training, and evaluation, this project scrutinizes the effectiveness of plain RNN, GRU, and GRU with Attention Mechanism in accurately identifying and categorizing posts. The report illuminates the performance metrics, including accuracy, precision, recall, and f1-score, attained by each model, thereby offering crucial insights into their respective capabilities in detecting potentially concerning content within online forums.

1. INTRODUCTION

The pervasive nature of online platforms as avenues for self-expression and support-seeking behavior has underscored the imperative need for proactive identification and intervention in instances where individuals convey distress, especially concerning suicidal thoughts or intent. The exponential growth of user-generated content on social media and community forums has made it challenging to manually sift through and detect concerning content efficiently. Addressing this challenge, this study focuses on harnessing the power of deep learning and natural language processing techniques to automate the classification of posts related to suicidal ideation within the aforementioned Reddit communities.

The fundamental goal is to employ sophisticated recurrent neural network architectures to accurately distinguish between posts that signal an individual's distress, explicitly discussing thoughts of self-harm or suicide, and those that do not exhibit such indications. The ability to automate this process could significantly aid in the timely identification of individuals in emotional distress, facilitating prompt intervention and support provision. This project aims to elucidate the efficacy of distinct RNN models – plain RNN, GRU, and GRU with Attention Mechanisms – in this critical task, offering valuable insights into their comparative performance, strengths, and limitations. By analyzing and evaluating the performance metrics of these models, this report contributes nuanced perspectives and empirical evidence crucial for the development of effective automated systems to assist mental health support efforts in online platforms.

1.1. **Contributions.**

- Implemented a preprocessing pipeline using Spacy, involving stopword removal, lemmatization, tokenization, and lowercasing, while constructing a vocabulary set for constructing numerical representations for the textual data.
- Employed plain RNN, GRU, and GRU with Attention mechanisms to classify posts from "r/SuicideWatch" and "r/teenagers," training and evaluating these models to differentiate between suicidal and non-suicidal content.
- Conducted comparison of each model using the following evaluation metrics: accuracy, precision, recall, and f1-score.

2. BACKGROUND

In 2020, the Centers for Disease Control and Prevention (CDC) WISQARS Leading Causes of Death Reports highlighted significant findings about suicide in the United States. it emerged as the twelfth leading cause of death overall, resulting in the loss of over 45,900 lives. Shockingly, it ranked as the second leading cause of death among individuals aged 10-14 and 25-34, the third among those aged 14-24, and the fourth among individuals aged 35-44. Notably, the number of suicides, totaling 45,979, almost doubled the count of homicides[2]. What's worse is the increasing trend in the suicide rate, and this has prompted considerable focus among researchers, leading to extensive studies from various angles aiming to tackle this intractable problem.

3. Related Work

Exploring mental health detection and suicidal prevention encompasses a wide array of research methodologies across various disciplines. Clinical techniques, involving patient-client interactions and health-related records such as genetic data[12] and electric health records(EHRs)[11], along with neural imaging[10], constitute vital approaches.

The prevalence of social media platforms has led to the emergence of user-generated content as a rich source of data for mental health research. Textual data, in particular, has been extensively studied for detecting mental health-related concerns. In a comprehensive meta-analysis focusing on research papers harnessing textual data and natural language processing methods in mental illness detection, 81% of a total of 399 selected studies leverage data extracted from social media platforms like Twitter, Reddit, Facebook, Weibo, etc. Among these platforms, Reddit stands out due to its categorization of posts by topics(subreddits) and its open data policy, which makes it an attractive source for research purposes[9].

However, the complexity of the datasets poses challenges when using traditional machine learning techniques like Support Vector Machines(SVM)[5] and clustering methods[6], as textual data demands, feature engineering for optimal performance. This process can consume considerable time and resources.

In contrast, deep learning has garnered significant attention in recent years due to its inherent advantages over conventional models. Notably, it eliminates the need for extensive feature engineering and demonstrates the capability to achieve high accuracy. In the realm of deep learning, several architectures have been explored for textual classification in the context of mental health detection. Convolutional Neural Networks(CNN) has been used to capture local patterns and features in text data, making them suitable for tasks like sentiment analysis and mental health classification[7]; Recurrent Neural Networks(RNN), including variants like Long Short-Term Memory(LSTM) and Gated Recurrent Unit(GRU), have been employed to capture sequential dependencies in text data[8]; and Transformer based models, such as BERT has also been employed for text classification tasks[9].

4. Approach and Experiments

4.1. **Textual Data Pre-Processing.** The dataset consists of posts obtained from two subreddits: "SuicideWatch," containing labeled suicidal posts, and the "teenagers" subreddit for non-suicidal posts. Initially, the data underwent preprocessing to prepare the textual content for model training. This involved employing the spaCy library for stop word removal and lemmatization on the text data. Subsequently, the text underwent tokenization and lowercasing, and any non-meaningful special characters were removed from these tokens. The resulting cleaned and processed text data was then used for feature engineering purposes.

4.2. **Feature Engineering.** The preprocessed data serves as the foundation for creating a vocabulary set encompassing the 20,000 most frequently used words. This curated vocabulary set played a pivotal role in transforming textual tokens into numerical representations, each standardized to a length of 100. When necessary, padding was applied to ensure uniformity in text length – a crucial step in preparing the data for model input.

4.3. **Model Architecture.** Three distinct recurrent neural network (RNN) architectures—RNN, GRU, and GRU enhanced with an attention mechanism—were developed and trained for the classification task. Through rigorous experimentation with various model configurations, the finalized structures of these models are as follows:

4.3.1. *RNN.* Configured with an input size matching the constructed vocabulary set, a single recurrent layer with a hidden size of 256, and an embedding dimension of 300. The input to the network is first embedded using "nn.Embedding()", then it is passed through "nn.RNN()" followed by a linear layer and a sigmoid activation layer to generate the final output.

4.3.2. *GRU.* Mirrored the setting of the RNN model, employing similar configurations for input and output sizes, hidden size, embedding dimension, epochs, learning rate, loss function, and optimizer settings. Utilized "nn.GRU()" for the recurrent layer.

4.3.3. *GRU with Attention Mechanism.* Retained the foundational settings of the GRU model while introducing additional components. Beyond text embedding and employing "nn.GRU()" for the recurrent layer, it incorporated an attention mechanism. This involved positional encoding followed by a 3-head transformer encoder layer, aiming to enhance the model's capacity to capture significant textual patterns.

All three models underwent 50 epochs of training using a learning rate of 0.0001, with the Adam optimizer and cross-entropy loss function.

5. RESULTS

The models are trained and validated using a 70-30 train test split. While accuracy provides an overall view of model performance, metrics like precision, recall, and f1-score offer deeper insights into the model's behavior. Precision measures the accuracy of the positive predictions. This is valuable when the cost of false positives is high. Similarly, recall calculates the ratio of correctly predicted positive observations to actual positives. It is crucial when the cost of false negatives is high. In the case of this project, we want a high recall since it implies that the model can capture most of the actual positive cases, minimizing the chance of missing a potential suicidal ideation. F1-score is a metric that combines precision and recall into a single score. It is particularly useful when the problem requires a balance of both precision and recall. It provides a harmonic mean of these two metrics, giving an overall understanding of the model's performance on both false positives and false negatives. Thus along with accuracy scores, precision, recall, and f1-score are also calculated.

The following table is a summary of the results of the three models: Looking at the results of the three models, it is

TABLE 1. Results.

	Plain RNN	GRU	GRU with Attention Mechanism
Accuracy	0.76	0.94	0.89
Precision	0.78	0.94	0.89
Recall	0.76	0.94	0.89
f1-score	0.76	0.94	0.89

evident that each model performs differently across the metrics. The GRU model stands out with the highest scores across all metrics. The plain RNN model lags behind at 0.76, showcasing a significant difference in performance. For the GRU and GRU with Attention Mechanism models, precision, recall and f1-score values are identical. This suggests that these models are performing consistently across correctly identifying true positives, true negatives, false positives, and false negatives. However, the plain RNN model has lower precision, recall, and f1-score compared to the other models. This implies that the plain RNN is not as effective in correctly classifying both positive and negative instances. In summary, the GRU model exhibits superior performance across all metrics, while the plain RNN model performs notably poorer.

6. DISCUSSION

Interestingly, GRU with Attention Mechanism has lower scores than the plain GRU. Even though attention mechanisms generally empower models to selectively focus on pertinent segments within input sequences, bolstering performance in tasks where specific elements carry greater significance, it did not result in a better performance in this case. One possible reason might be that the dataset used in this project is relatively simple and lacks intricate patterns. In this case, adding more complexity using the attention mechanism might not provide significant advantages, and using a simpler structure might generalize better on the dataset.

There are several directions for further explorations if more time and data is available. First, the dataset used in this project, while informative, does not fully capture the complexity and diversity of suicidal and non-suicidal content found on social media. Expanding the dataset and including more diverse sources of text data could provide a more comprehensive understanding of the challenges associated with suicidal content detection. Additionally, exploring other deep learning architectures such as Transformers and BERT models could be valuable, as these models have demonstrated state-of-art performance in various natural language processing tasks. Investigating ensemble models that combine the strengths of multiple architectures could also lead to a more robust mode.

REFERENCES

[1] Komati, N. (n.d.). Suicide Watch. Kaggle. Retrieved from https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch

[2] "Suicide." National Institute of Mental Health, U.S. Department of Health and Human Services, www.nimh.nih.gov/health/statistics/suicidepart7688. Accessed 14 Dec. 2023.

[3] Zhang, T., Schoene, A.M., Ji, S. et al. Natural language processing applied to mental illness detection: a narrative review. npj Digit. Med. 5, 46 (2022). https://doi.org/10.1038/s41746-022-00589-7

[4] Su, C., Xu, Z., Pathak, J. et al. Deep learning in mental health outcome research: a scoping review. Transl Psychiatry 10, 116 (2020). https://doi.org/10.1038/s41398-020-0780-3

[5] Bernice Yeow Ziwei and Hui Na Chua. 2019. An Application for Classifying Depression in Tweets. In Proceedings of the 2nd International Conference on Computing and Big Data (ICCBD 2019). Association for Computing Machinery, New York, NY, USA, 37–41. https://doi.org/10.1145/3366650.3366653

[6] Park, A., Conway, M., & Chen, A. T. (2018). Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: a text mining and visualization approach. Computers in human behavior, 78, 98-112.

[7] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention. In The World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 514–525. https://doi.org/10.1145/3308558.3313698

[8] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[9] Zhang, T., Schoene, A. M., Ananiadou, S. (2021). Automatic identification of suicide notes with a transformer-based deep learning model. Internet interventions, 25, 100422.

[10] Kuang, Deping Guo, Xiaojiao An, Xiu Zhao, Yilu He, Lianghua. (2014). Discrimination of ADHD Based on fMRI Data with Deep Belief Network. 225-232. 10.1007/978-3-319-09330-727.

[11] Smoller, J. W. (2018). The use of electronic health records for psychiatric phenotyping and genomics. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 177(7), 601-612.

[12] Laksshman, S., Bhat, R. R., Viswanath, V., Li, X. (2017). DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning. Human mutation, 38(9), 1217-1224.