Suicide Notes: Inferring suicide diagnoses from medical notes

Dylan Warnecke
dwarnecke@utexas.edu

1 Introduction

Suicide is one of the leading causes of death worldwide. Each year, suicide kills approximately 720,000 people and is the leading cause of death in 15-29-year-olds [1]. It is also quickly worsening as a problem as between 2000 and 2022, suicide rates increased approximately 36% in the United States [2].

Despite this, suicide is also one of the most preventable causes of death. Two-thirds of victims make some form of contact with the mental healthcare system a year before their death, and many of these people leave indications of suicide attempts or ideation in their medical notes before their passing [3]. In the form of patient interviews, psychiatric exams, medical tests, and other methods, providers can document many pieces of evidence of suicidal behavior in written medical note text.

Much of this evidence is not reflected in patient ICD diagnoses however. Previous studies have shown that ICD diagnoses capture only 3% of suicidal ideation events while respective medical notes can describe up to 97% of such events [4]. Automating suicidal behavior diagnoses using medical notes and large language models could close this discrepancy and better identify patients who are at risk for suicide. This paper introduces two models to do so, a classifier and a categorizer, that can respectively classify if medical note sentences describe suicidal behavior, and categorize those suicidal sentences into groups of ICD codes.

2 Related Works

Predicting ICD codes from clinical text has been previously studied. Many models have been developed to classify appropriate ICD codes from clinical notes. However, most of these perform poorly with class imbalance adjusted F1 scores not exceeding 0.60 due to the massive scope of the projects [5]. This has focused training diagnosis models for specific contexts, including suicide.

One study developed RoBERTa-based paragraph and hospital-stay classifiers to predict if text segments contained evidence for either suicide ideation or attempts [6]. These texts were further categorized as either positive, negative, or unsure in motive with separate models. The paragraph classifier had an F1 score of 0.88 in labeling paragraphs as containing suicidal evidence, and the hospital-stay classifier had F1 scores of 0.78 and 0.60 in classifying notes as containing attempt and ideation evidence respectively. These models did not attempt to classify the notes with any specific ICD codes however, meaning that they would still not be able to automate diagnosis documentation.

Additional models have been developed outside of using clinical notes to predict suicide risk. Through a combination of decision trees and gradient boosting, one model has predicted patients at risk for suicide by using demographic features like race and religion, and clinical features like alcohol abuse and family history to achieve a 0.84 F1 score [7]. This model was only trained on a dataset of 75 patients and was not validated with an additional set of validation or test patients.

3 Methodology

3.1 Dataset

Several available datasets contain relevant training data, but a MIMIC III notes subset was used to train the available data due to the size and detail of the dataset [6]. The dataset annotates multiple paragraphs across 12,759 notes spanning 697 hospital stays and states whether the annotated text refers to an attempt or ideation. Each annotation also clarifies if the evidence is truly intentional, not intentional, or unclear in motivation. All positively intentional suicide attempt annotations are also categorized into groups of ICD codes. T36-50 indicates poisoning by medicinal substances, T51-65 indicates poisoning by non-medicinal substances, T71 indicates asphyxiation or suffocation, X71-83 indicates the use of firearms, jumping from high places, and other means, and T14.91 describes previous suicide attempts in a patient's history. Negatively intentional annotations are also grouped into an N/A category, and attempts with undeterminable intentions are grouped into an unsure category. The dataset is also split into training and validation subsets.

This dataset was further preprocessed. Annotations were first classified as either suicidal or not suicidal. Any positive suicide attempts were labeled as suicidal, and unsure attempts were also labeled as suicidal due to the severity of false negatives in the context. Any suicide ideation was also labeled as suicidal and given the respective ICD-10 suicide ideation code R45.851. All remaining sentences were labeled not suicidal. The notes were then tokenized into sentences, and the sentences were then labeled as suicidal themselves if they contained a suicidal annotation entirely within them. This produced a training and a validation set of 287,219 and 53,293 sentences labeled as either containing suicidal evidence or not, and their respective diagnosis codes if applicable. Analyzing the datasets reveals that they are highly imbalanced. In the training dataset, only 5.0% of the sentences contain evidence of attempts or ideation, and 48.6% of the given diagnosis codes belong to one category.

3.2 Models

Sentence granularity at the input for these models was chosen because of the importance of model explainability in healthcare today. Clinicians making diagnosis decisions need evidence for why algorithms like the proposed models make predictions, and the ability to cite a specific sentence as evidence is highly beneficial in this regard. The task of identifying appropriate ICD-10 codes was split into two tasks with two distinct models. One model would be trained to identify if a sentence contained any evidence of suicidal behavior, and the other would be trained to identify a diagnosis code group for those sentences containing suicidal behavior evidence. This split was made to improve training due to the dataset imbalance and prevent the increased computational complexity of a single model.

Both of the models were made using the same design except for the final dimension output. Figure 1 and Figure 2 outline the design of the models below. The models clean and tokenize the sentences before converting them into embeddings and passing them through transformer and linear layers to classify the models. Due to computing resource limits, other pre-trained large language models were not used as base models. Instead, the transformer layers were trained from random initializations and the embedding vectors were sourced from the 400,000 vocab GloVe word vectors [8]. The embedding vectors are compressed into a lower dimension using a linear layer without activation or bias to further reduce the computational complexity of the model. Transformer modules with pre-layer normalization and residuals are used for sequence transformation. Adaptive max pooling is used to eliminate the sequence length dimension. Linear layers with ReLU activation are used to transform the features until final classification logits are calculated.

Models were trained using cross entropy loss and dropout was applied after the embedding layer and within the transformers to prevent overfitting. Upsampling was performed on the minority class data to address the class imbalance.

4 Results

Classifiers fine-tuned and trained in this design achieved a 0.65 F1 score with a 0.66 precision and a 0.65 recall on the validation dataset. This does not exceed the baseline 0.88 F1 score achieved by other suicidal evidence classifiers [6]. This also indicates that the model does not achieve the performance that it potentially could, and more work could be done to improve the model in the future. Still, the classifier performs better than a uniformly random guess that achieves an approximate 0.09 F1 score, showcasing the use of the classifier. Categorizers fine-tuned and trained in this design achieved a weighted F1 score of 0.78 on the validation dataset. This exceeds the baseline 0.60 score achieved by other text-to-ICD classifiers [5]. The difference between these scores indicates the usefulness of the new categorizer, though it retains the limitation that it can only be used with sentences that are predicted to be suicidal.

Both of these models could be improved primarily through larger, more contextual embedding vocabularies. The small GloVe embedding vocabulary does not contain the medical and abbreviated terms often seen in medical text like “Seroquel” or “pt” that carry significant meaning, and are reduced to unknown token vectors in the current model. Furthermore, the GloVe vector embeddings do not provide healthcare-specific context in the embeddings. This makes it harder to differentiate medical terms by conflating words like “dose” and “overdose” with similar embeddings and misrepresent words like “depression” that have multiple meanings outside of a medical context.

Training with more data could also improve validation performance. There are only approximately 580 hospital admissions included in the training set, and the sentences in each of these admissions may be correlated due to them describing the same patient’s story, thus having similar words and similar embeddings. This can result in large-scale overfitting across many training samples in the dataset. Adding more notes to the training data could help generalize the model across more unseen words and stories, and further improve performance.

5 Conclusions

This paper introduces two large language models: a classifier that can indicate if a sentence contains evidence of patient suicidal behavior, and a categorizer that can label those suicidal sentences with an appropriate ICD-10 code. The F1 scores for these models were 0.65 and 0.78 respectively. These are mixed results compared to the baseline results of other models in similar tasks, but still show predictive power when compared to results in random guessing. Further improvements could be made to these models by incorporating larger and more contextual embedding vocabularies as well as adding additional training data. Searching online reveals that these may be the best algorithms in predicting specific ICD suicide codes currently, meaning that these models could help enable clinicians to properly document and diagnose suicidal behavior, and potentially save lives.

References

[1] World Health Organization. Suicide, August 2024. URL https://www.who.int/news-room/fact-sheets/detail/suicide.

[2] Center for Disease Control. Facts about suicide | suicide prevention, July 2024. URL https://www.cdc.gov/suicide/facts/index.html.

[3] Alex Luedtke Alan M. Zaslavsky Ronald C. Kessler, Robert M. Bossarte and Jose R. Zubizarreta. Suicide prediction models: a critical review of recent research with recommendations for the way forward. *Molecular Psychiatry*, 25(1):168–179, 2024. doi: https://doi.org/10.1038/s41380-019-0531-0.

[4] Elias Brandt Rodney D. Nielsen Richard R. Allen Anne M. Libby David R. West Heather D. Anderson, Wilson D. Pace and Robert J. Valuck. Monitoring suicidal patients in primary care using electronic health records. *Journal of the American Board of Family Medicine*, 28(1):65–71, 2015. doi: https://doi.org/10.3122/jabfm.2015.01.140181.

[5] Dat Quoc Thanh Vu and Anthony Nguyen. A label attention model for icd coding from clinical text. In *IJCAI, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (Main track)*, pages 3335–3341, 2020. doi: https://doi.org/10.24963/ijcai.2020/461.

[6] Hong Yu Bhanu Pratap Signh Rawat, Samuel Kovaly and Wilfred Pigeon. Scan: Suicide attempt and ideation events dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022. doi: https://doi.org/10.18653/v1/2022.naacl-main.75.

[7] Mohd Halim Mohd Noar Noratikah Nordin, Zurinahni Zainol and Lai Fong Chan. An explainable predictive model for suicide attempt risk using an ensemble learning and shapley additive explanations (shap) approach. *Asian Journal of Psychiatry*, 79:103316, 2023. doi: https://doi.org/10.1016/j.ajp.2022.103316.

[8] Richard Socher Jeffrey Pennington and Christopher D. Manning. Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. doi: https://doi.org/10.3115/v1/D14-1162.

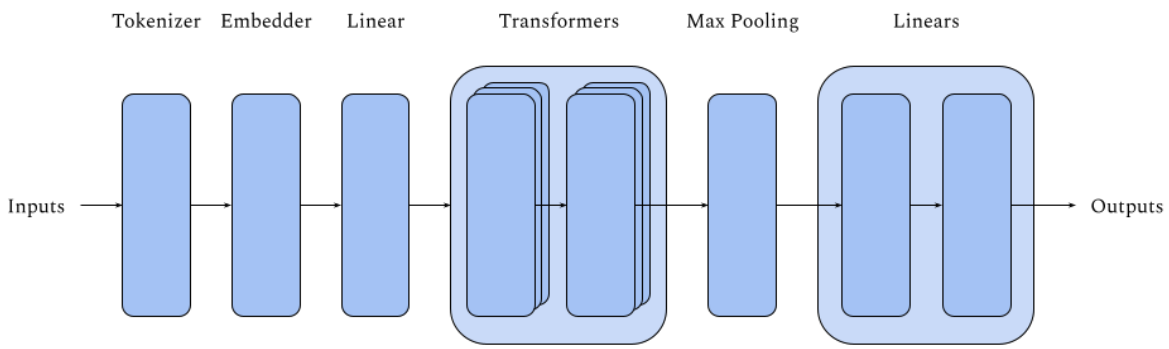

Figure 1: The design of the classifier and categorizer models

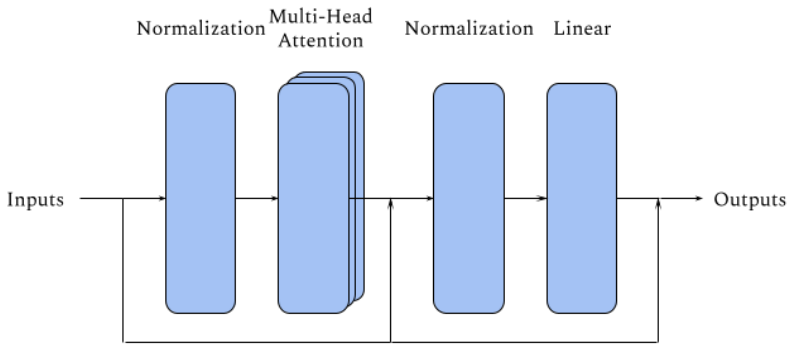

Figure 2: The design of the transformer modules