SUICIDE IDEATION DETECTION SYSTEM

Madhav Mukund Verma, Upamanyu Ghosh, Gayatri Malladi,
Mashrukh Islam, M Sai Sandeep and Dr. Tan Wee Kek
NUS School of Computing, Computing 1, 13 Computing Drive, Singapore 117417

ABSTRACT

Suicidal ideation is one of the main predictors of the risk of suicide attempt and can be described as thoughts, ideas, planning, and desire to commit suicide. Fast detection of such ideation in early stages is essential for effective treatment. Many expressions of suicidal ideation can be found in publications in social networks, especially by young people. Previous works explore the automatic detection of suicidal ideation in social networks for the English language using machine learning algorithms. In this work, we present the first exploration of machine learning algorithms for suicidal ideation detection for the Portuguese language. We compared three classifiers in Twitter data: SVM, LSTM, and BERT (multilingual and Portuguese). Results suggest that BERT is effective for suicidal ideation identification in reddit dataset, achieving 93.33%.

I. INTRODUCTION

Around a million people are reported to die by suicide every year1 and due to the stigma associated with the nature of the death, this figure is usually assumed to be an underestimate2. Prevention research is generally felt to have been partially successful in achieving some reduction in suicide rates, particularly by increasing physician training and restricting access to lethal means3. However, recent reviews of suicide-related behaviour reported little change in prevalence estimates from the 1980s to the late 2000s4 and the World Health Organization reported that around half of member countries who submitted data experienced steady or increased suicide death rates from 2000 to 20121,5.

Suicide prevention research has been limited by methodological factors such as relative rarity of the event, short observation periods and recall bias6,7,8. One way to potentially overcome these limitations is through the use of large clinical datasets from which high-risk cohorts can be assessed, and the wealth of data contained in Electronic Health Records (EHRs), or clinical databases, provide a means to achieve this4,6,9. In addition, focusing on non-fatal suicidal behaviours, such as suicide ideation and suicide attempts, also helps to overcome some methodological limitations as their prevalence is much higher than completed suicide. Suicide ideation, is represented by the presence of current plans and wishes to attempt suicide in individuals who have not made any recent overt suicide attempts, and its severity has been proposed by Beck et al. as an indicator of suicidal risk10. Suicide attempt is defined as a non-fatal, self-directed, potentially injurious behavior with an intent to die as a result of the behavior, but that may or may not result in injury11. Both of these behaviours are also often considered vital risk factors for completed suicide and hence are well placed for suicide prevention research.

Previous studies using EHRs for suicide research have used structured diagnostic and International Classification of Disease, ninth revision (ICD-9) cause-of-injury codes (also known as E-codes) to identify patients who have attempted suicide12,13,14. However, the quality and practice of EHR suicidal behaviour coding varies widely. For example, a study assessing the recording of suicidal ideation and attempts in primary care clinical records (N = 15,761) found 1,025 patients who had suicidal ideation recorded in text fields, of whom

only 3% had a structured code indicative of this. Furthermore of 86 patients identified as having made a suicide attempt, only 19% had the relevant code for this15.

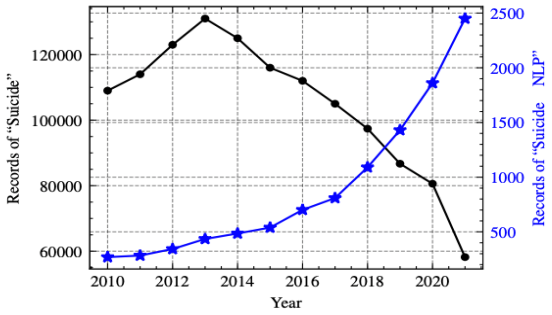

Figure 1: The number of scholarly articles published in the past twelve years. The black line with dot markers presents the records searched with query of "suicide" and the blue line with asterisk markers shows the records with query of "suicide" and "NLP".

Natural Language Processing (NLP) offers wide-ranging solutions to retrieve and classify data from EHRs. Text mining, which is part of the NLP family, is defined as the analysis of 'naturally-occurring' text driven by human specification to achieve one or a range of goals (e.g. information retrieval or artificial intelligence)16. The analysis often manifests as a set of programming rules or machine learning algorithms (i.e. algorithms generated from automated learning from manual examples), whose eventual output should represent human output as much as possible17. One of its applications can be to identify and classify instances of suicide ideation, attempts and death by suicide recorded within free-text medical notes, if the data are not readily derived from structured fields18. This use alone can significantly improve the common limitation in suicide prevention research of low case sample sizes.

Using NLP (and text mining) is a relatively new venture in suicide prevention research14,19,20 and only recently have some studies reported using intuitive text mining approaches to identify suicidal behaviour in clinical notes21,22,23. Text mining methods used in suicide research have evolved from simplistic dictionaries and search engines18,24,25 to more NLP orientated approaches. Haerian et al., described a hybrid process26 developed for radiology databases by Friedman et al.27,28, combining use of structured E-codes and NLP to identify suicide thoughts and attempts in their clinical EHR database. In 2014, Ben-Ari et al. integrated their simple text search approach with a random forest classifier to identify suicide attempts in clinical notes23. More recently, Metzger et al. utilised seven standard machine learning techniques (Support Vector Machines, predictive association rules, decision trees, logistic

regression, Naïve Bayes, random forest and neural networks) to classify suicidal ideation from suicide attempt from a list of demographic and clinical variables22.

Contributing to the sparse literature on text mining methods used to identify suicidal behaviour for research4,29, the aim of this study is to describe the independent development of two NLP tools – one for detecting the presence of recorded suicide ideation and the other for detecting a recorded suicide attempt – and to evaluate each tool's performance against manual text annotation using precision and recall. To our knowledge, this is the first time rules-based or machine learning have been used to detect suicidality in a psychiatric database. We seek to present the strengths and weaknesses of our approaches, with a view to future development and improvement of text analysis in this topic area. We recognize that there are no standard rules outlining how to use text mining techniques to extract data from observational datasets and so, in this manuscript, we present two working approaches.

Suicide means ending one's own life intentionally. The reasons behind this act can be divided into two major parts. One can be any particular incident that triggers a person to commit suicide. This is very sudden, unplanned and most of the time we cannot prevent or take early measures to stop the initiative; this is known as impulsive suicide. The second reason behind committing suicide is a mental disorder, where a person suffers from depression, PTSD, anxiety and other mental problems and start to think ending life is a solution for these sufferings. These mental problems create a lack of mental stability which leads to the thought of killing oneself from extensive thought and detailed planning. This is called Suicidal Ideation or Suicidal Thoughts.

When it comes to NLP, handling sequential data is very important. And as transformers does not require processing sequential data in order, it performs better than other recurrent neural networks like Bi-LSTM. Since transformer models allow parallelism it made possible training on larger datasets, and that is how the pre-trained systems like BERT and its variants ALBERT, ROBERTa, XLNET were developed. In our study, we made a classification model that shows Transformer based model such as BERT, ALBERT, ROBERTa, and XLNET perform significantly better than old recurrence based neural networks like Bi-LSTM in the area of sentiment analysis, suicidal ideation detection to be precise.

In this paper, we investigate Suicidal Ideation detection from social media posts of users. We have used user posts from SuicideWatch. SuicideWatch is a subReddit where users anonymously post about their sufferings, traumatic incidents, or their fight with mental illness. These posts often have words or intentions that indicate suicidal thoughts in their head. We have taken these posts with full user privacy and used Natural Language Processing (NLP) to create a model that will detect Suicidal Ideation in these posts. Our paper mainly focuses on the possibility of using Transformers, a state of the art Deep Learning model to detect Suicidal Ideation.

II. DATASET COLLECTION

We selected the notes from the MIMIC-III (Johnson et al., 2016) dataset, which consists of the de-identified EHR data of patients admitted to the Beth Israel Deaconess Medical Centre in Boston, Massachusetts from 2001 to 2012.

Data includes notes, diagnostic codes, medical history, demographics, lab measurements among many other record types.

We chose MIMIC-III because it is publicly avail- able under a data use agreement and allows clinical studies to be easily reproduced and compared.

The diagnostic ICD codes for the patients are provided at hospital-stay level in MIMIC with admission identification numbers (HADM_ID in MIMIC database).

For each stay, multiple de-identified notes such as nursing notes, physician notes, and discharge summaries are available.

ScAN consists of 19, 690 unique evidence annotations for the suicide relevant sections of 12, 759 EHRs of 697 patient hospital-stays. There are a total of 17, 723 annotations for SA events and 1, 967 annotations for SI events. The distribution for both SA and SI events is provided in the following table.

Table 1: Distribution of unique annotations at the patient, hospital-stay and notes level

General Statistics	**Patients** **669**	**Hospital-stays** **697**	**Notes** **12, 759**
Suicide Attempt	Positive 14, 815	Negative 170	Unsure 2, 738
Suicide Ideation	Positive 1,167	Negative 800	

III. PROPOSED METHODOLOGY

In this study, we intend to introduce a new natural learning processing (NLP) model to improve the performance of text classification to detect suicidal ideation in a social media post. Figure 1 gives us a general overview of our proposed model. We have divided our proposed model into three parts data layer, Embedding layer, and Classification layer according to their work.
In the era of pretraining, many pretrained large language models have been applied to fine-tune suicide text classifiers. Those pretrained models achieved superior classification performance. One straightforward question in this position paper is whether the pretrained models can "understand" the latent intention to some extent. Taking the sentence "I am going to buy a knife" as an example, it can be recognized as purchasing intention in a daily context or suicide attempt (i.e., to commit suicide with a knife). Starting with our question, we conduct the fill-mask language modeling task with the sentence "I am going to buy a knife and [MASK]", using several masked language models, including BERT and RoBERTa , and their domain-specific variants MentalBERT and MentalRoBERTa (Ji et al., 2022b). Figure 3 shows the output of word probabilities. The suicidal ideation does not appear in the prediction of BERT. RoBERTa and MentalBERT predict the suicidal intention ("die") in the fifth place. MentalRoBERTa recognizes the suicidal intention as the first place. Then, we use another example sentence "This life is not worth living. I am going to buy a knife and [MASK]." that provides a bit more contextual information. The results in Figure 4 show that RoBERTa and MentalBERT tend to predict suicidal intention ("die") with higher probabilities. MentalRoBERTa predicts "die" with a significantly high probability. These two examples showcase the abilities of masked language models to predict intention as a fill-mask task. Domainadaptive continued pretraining helps with suicide keyword prediction, although we do not consider the memorization issue of BERT here. While MentalRoBERTa outputs "die" with a high probability, which can be interpreted as suicidal intention, we also see the outputs of RoBERTa ("shoot") can be interpreted as potential anti-social and criminal actions (e.g., shoot at others, although it does not make sense with a knife.).

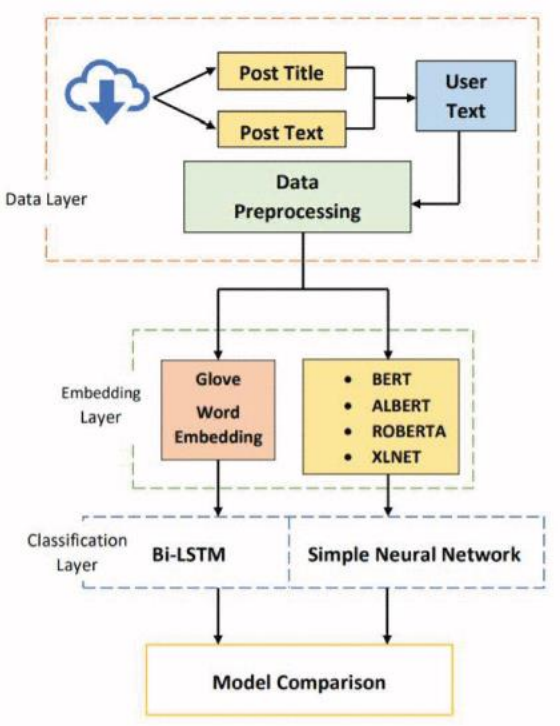

Figure 1: Data flow in the proposed model

III.1. Data Layer

The data layer part consists of preprocessing the suicidal suggestive texts. For the importance of both “Post Title” and “Post Text” parts, we have concatenated both user text and post of each user. The resulting user text needs to be passed into filters to improve the accuracy of our proposed model. As raw user posts have abbreviated words and it will be difficult for any automated machine to get the main context of the post. For this, we have expanded the abbreviated words to its correct form. We also removed any URL addresses and redundant whitespaces with a single white space. Again, We removed any emoticon which could give us an inconsistent result. This preprocessing will make the user raw texts to a format that can be understood by the word embedding.

III.2. Embedding Layer

The embedding layer consists of two different types of word embedding. Word embedding is a natural language processing where each word has been represented as a vector in a pre-trained vector space with the help of distributional semantics, which states that similar meaning words occur in the same context. Generally, this word embeddings are pre-trained by the help of an

alternative objective with a large unlabeled dataset, such as predicting terms based on their context. By doing so the vector can learn the words semantical details. Glorot et al. [11] used embeddings with autoencoders in every layer for domain adaptation in the classification of review sentiments. It shows an improved state of the art results. The wide use of word embedding in the recent literature can state that in deep learning text-based classification, word embedding plays a crucial role.

For our model we have used two different word-embedding techniques:

GloVe Word Embedding: The glove is an adjusted weight least-square word-embedding model that gives us a vector space of word vectors. It was created by Pennington et al. [12] The word matrix is normalized by counts and log smoothing operation. The word is also factorized to get low dimensionality by using "reconstruction loss". The output is then fed into our Bi-LSTM model.

We used the pre-trained word embedding model that was trained on web data using a common crawler which has 840 billion tokens with a vocabulary of 2.2 million. The dimension of the embedding is 300.

Contextual Word Embeddings: The vector representing the traditional word embeddings such as word2vec [13] and glove only express the words by creating a global vector representation for all the sentences. This creates a problem for the context of the word based on the surrounding words. This context-based problem was solved by contextualized word embedding that fine-tunes the model architecture of transformer encoders [14]. In our study, we have compared BERT, ALBERT, ROBERTA, and XLNET as they give the result in the NLP domain.

III.3. Text Representation

The BERT-based pre-trained models play a vital role in the field of natural language understanding (NLU) tasks such as machine translation, question and answering, automated reasoning, and newsgathering. In this paper, the word and sentence features or embedding vectors are extracted from the text using BERT. The BERT dynamically produces the context information. Moreover, it accepts input in a specific token format [CLS] TEXT [SEP]. The [CLS] and [SEP] indicate the beginning of the text and the separation of the text into one or two sentences, respectively. First, the texts are tokenized into vocabulary list and indexes.

Second, the text tokens are defined with Token IDs, Mask IDs, Positional Embeddings, and Segment IDs. Later, these data are converted into tensors and fed into the BERT-based transformer models.

III.4. Transformers-Based Classifiers BERT

BERT: Bidirectional Encoder Representations from Transformers (BERT) uses the "masked language model" (MLM) that helps representing the left and the right context in the model. This gives the ability to train a bidirectional transformer model. The masked language model works by masking token from the text and to predict the token based on the context. The BERT can be denoted by the number of layers. One of them is the BERTLarge which has 24 layers and BERTBase which has 12 layers. We have worked with the BERTBase as it takes less computational time, and it gives almost the same results in sentiment analysis as BERTLarge.The BERT is pre-trained over the different pretraining task with unlabeled data. For classification, the pre-trained BERT takes the user text token as input with the [CLS] tag. We fine-tune the BERT end to end hyperparameters as necessary for our task.

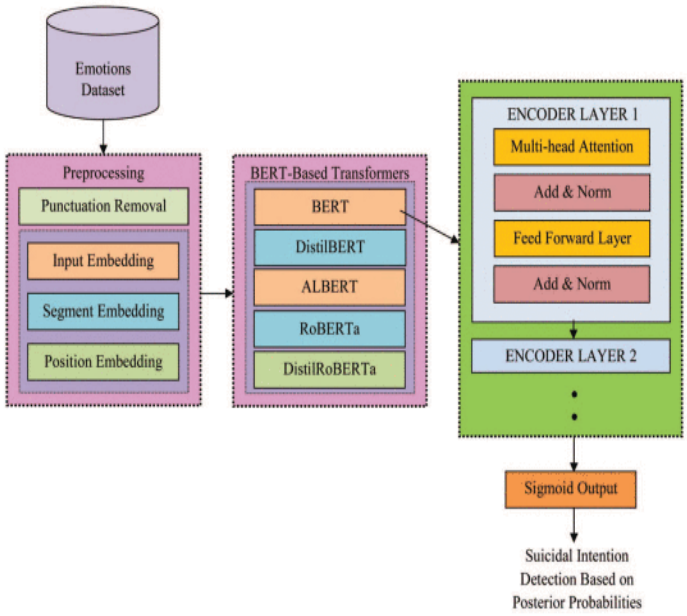

Figure 2: Suicide Ideation Detector (SID) – Architecture

BERT is an encoder representation of the transformer model that is designed to train bidirectional context representations from both left and right direction of a text or comment. It

is pre-trained for two objectives, namely, masked word prediction and next sentence prediction. The first one mask 15% of tokens in the input sequence. Then, this sequence is given as input to the BERT model for predicting the masked tokens. The latter one takes a pair of input sequences like A and B. Then, the BERT model learns to identify whether the sequence A follows B. In particular, RNN models learn context representation, either from left to right or right to left direction. Conversely, the BERT model processes the entire input sequence at the same time in parallel using the self-attention mechanism. The encoder layers of the BERT model is shown in Figure 1. In the pre-trained model, a Sigmoid function is added to its top for fine-tuning purpose. The BERT model is designed with two variants, namely, BERT-Base and BERT-Large. The first one is built with 110M parameters based on 12 encoders, 12 attention heads, and 768 hidden states. Similarly, the latter one is built with 340M parameters based on 24 encoders, 16 attention heads, and 1024 hidden states. Specifically, the BERT-base model is used for the task of suicidal intention detection.

IV. SOCIAL IMPACT

Early detection of suicide ideation provides a solution to early intervention so that social workers can help people living with mental health issues through proactive conversations. However, no significant evidence shows that suicidal risk assessment can guide decision-making in clinical practice (Large et al., 2017). We suggest that people experiencing a mental health condition seek professional help from psychiatric services. Research on suicidal intention understanding and risk assessment does not aim to replace psychiatrists. It can empower social workers to prioritize social resources for people with mental conditions. The sensitive nature of suicide-related data requires our research to protect privacy. This study uses social media posts from anonymous users that are manifestly available on the website. Furthermore, these collected posts are stored on passwordprotected servers. We do not attempt to identify or contact social users.

V. CONCLUSION

In this study, we explored the idea of using transfer models in the field of suicidal ideation detection. We showed a complete and comprehensive approach of using pre-trained transfer models that cover all the variety of BERT and concluded that in terms of performance this

comparatively newer technology surpasses conventional Deep Learning models. We tried to make an impact in the automatic suicidal threat detection field. Our work is just paving the path for using more modern and advanced technology in the mental health domain. We faced difficulties in our study with the dataset as the volume of data is small and has annotation bias. For further research, we need more annotated data that are labeled with a set of rules defined by mental health specialists. As other mental health issues like depression, anxiety are closely related to suicidal ideation, a set of rules must be set for further studies in this field.

VI. FUTURE ENHANCEMENT

We hope our research will help to create an efficient automated system in social networks which will have the ability to detect early suicidal threats in an individual and can help them accordingly. Furthermore, in future neural models should focus more on linguistic features as linguistics are an integral part of emotion. Along with this a benchmark dataset will help to create a stronger and robust model which will detect suicidal ideation more precisely as it is a tricky problem to classify.

VII. *REFERENCES*

[1] Guntuku, S. C., Buffone, A., Jaidka, K., Eichstaedt, J. C., & Ungar, L. H. (2019, July). Understanding and measuring psychological stress using social media. In Proceedings of the international AAAI conference on web and social media (Vol. 13, pp. 214-225).

[2] Shensa, A., Sidani, J. E., Dew, M. A., Escobar-Viera, C. G., & Primack, B. A. (2018). Social media use and depression and anxiety symptoms: A cluster analysis. American journal of health behavior, 42(2), 116-128.

[3] World Health Organization, 2018. National Suicide Prevention Strate- gies: Progress, Examples and Indicators. World Health Organization, Geneva. https://apps.who. int/iris/handle/10665/279765.

[4] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 con- ference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre- training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[8] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art naturallanguage processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (pp. 38- 45).

[9] De Beurs, D., Fried, E. I., Wetherall, K., Cleare, S., O'Connor, D. B., Ferguson, E., ... & O'Connor, R. C. (2019). Exploring the psychology of suicidal ideation: A theory driven network analysis. Behaviour research and therapy, 120, 103419.

[10] Cao, L., Zhang, H., Feng, L., Wei, Z., Wang, X., Li, N., & He, X. (2019, November). Latent Suicide Risk Detection on Microblog via Suicide- Oriented Word Embeddings and Layered Attention. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Pro- cessing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 1718-1728).

[11] Wang, N., Fan, L., Shivtare, Y., Badal, V., Subbalakshmi, K., Chan- dramouli, R., & Lee, E. (2021, June). Learning Models for Suicide Prediction from Social Media Posts. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Im- proving Access (pp. 87-92).

[12] Iob, Eleonora, Andrew Steptoe, & Daisy Fancourt. Abuse, self-harm and suicidal ideation in the UK during the COVID-19 pandemic. The British Journal of Psychiatry 217.4 (2020): 543-546.

[13] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of suicide ideation in social media forums using deep learning. Algorithms, 13(1), 7.

[14] Mila Kingsbury, Bjørn-Atle Reme, Jens Christoffer Skogen, Børge Sivertsen, Simon overland, Nathan Cantor, Mari Hysing, Keith Petrie, & Ian Colman (2021). Differential associations between types of social media use and university students' non-suicidal self-injury and suicidal behavior. Computers in Human Behavior, Volume 1