Deep Learning Suicide Ideation Detection Model

KIPKEBUT ANDREW[1], EMMANUEL CHESIRE[2]
[1, 2] *Computer science and I.T, Kabarak, Nakuru, Kenya*

Abstract- ***Suicidal Ideation in society today has become very common, this is due to stressful societal issues. Recently social platforms have gained special attention regarding this phenomenon. Mental health issues like depression, frustration, hopelessness, and bullying among others directly or indirectly influence suicidal thoughts. Early detection of suicidal intent can help people to diagnose and get proper treatment before it is too late. Deep Learning has played an important role in NLP-related predictions and detection. In this study, a novel detection approach that uses a deep learning approach is proposed. Essentially the study analyses raw natural language data from different sources such as social networks among others, and classifying the indication of suicidal ideation. This study focuses on Deep learning techniques as a base for suicidal ideation. The experiment shows that the BERT model can achieve an optimal classification result.***

Indexed Terms- ***Deep Learning, Suicide, Ideation, Natural Language Processing.***

I. INTRODUCTION

After the Covid 19 pandemic, mental health has become more important than ever before. Anxiety and depression, are becoming increasingly concerned in modern society, as they turn out to be more severe especially in developed countries and emerging economies. Severe mental disorders without effective treatment can turn to suicidal ideation or even suicide attempts [1].

Globally 800,000 people die from suicide every year, this is twice the number from homicide. Suicide is one of the leading causes of death in young people. 1.4% of global deaths in 2017 were from suicide. In some countries this rate goes up to 5%. Take, for example, Kenya, where World Bank data shows the suicide rate is 6.1 per 100,000 people, with men as the highest category. Suicide ideation is viewed as a tendency to end personal life that is caused by depression, hopelessness, and frustration [2].

According to some studies, most of the individuals with suicide ideation do not attempt suicide. A study by Klonsky et al. [3] believes that most of the risk factors (depression, hopelessness, frustration) connected with suicide are predictors of suicide ideation, not the progression from the ideation to the attempt. According to WHO, early detection of suicide ideation should be developed and implemented as a national suicide prevention strategy at the global level to reduce suicide rates by 10% in the future [1]. Social media has become a powerful tool for the mental health and well-being of its users, mostly young individuals are active on platforms such as Facebook, Twitter, and Instagram. In these social platforms written suicidal signs can be viewed as a worrying sign, and such individuals should be interrogated on the existence of suicidal thoughts. According to Choudhury et al. [4], social media text, such as blog posts, forum messages, tweets, and other online notes, is important in mining the ideation. Social platforms moreover give a valuable research platform for the development of new technological approaches and improvements which can bring novelty to suicide detection and its prevention [5]. The primary objective of this study is to explore the knowledge of suicide ideation using the various data sets from various platforms in Kenya using an effective deep learning technique. The main task is to explore the potential BERT algorithm for suicide ideation and detection.

II. RELATED WORK

In the latest research done by scholars, a considerable number of experiments have been developed to emphasize the influencing muscle of social media on suicide ideation. Choudhury et al. [4] developed a statistical method based on standardized comparison scores to obtain multiple variables for suicidal ideation to determine changes in mental health. According to the authors, this transition can be accompanied by

three unique psychological stages: thinking, ambivalence, and decision-making.

First, it includes thoughts of anxiety, hopelessness, and distress. The second is related to lowered self-esteem and reduced social cohesion, and the third is related to the violence and suicide plans that the subjects carry out. Additionally, Coppersmith et al. [6] examined the behavioural shifts of the users who identified a significant growth of tweets with feelings of sadness expressed in the weeks before a suicide attempt. Many studies reciprocated the role of social platforms in suicide ideation detection.

Jashinsky et al. [7] demonstrated the spatial relationship between suicide and the nature of dangerous situations in tweets. Colombo et al. [8] resonated that the tweets containing suicide ideation based on the users' behaviour in social network interactions result in a high degree of reciprocal correlation. Another interesting observation according to the Werther effect [9]. His work indicates a notable increase in users' posting frequency and the shifts in their linguistics in social media, the shift was observed in a direction toward more negative and self-focused posts with lower social integration. Similarly, Ueda et al. [10] delved into 1 million Twitter posts after the suicide of 26 celebrities in Japan between 2010 and 2014. Identification of regular language patterns in social media texts leads to more effective recognition of suicidal tendencies. It is often supplemented using various machine learning techniques for different NLP techniques. Desmet et al. [11] built a suicide note analysis method to detect suicide ideation using Support Vector Machine (SVM) classifiers. Huang et al. [12] developed a psychological dictionary based on the Chinese mental dictionary (Hownet). He used the SVM method for classification recognition to develop a real-time suicide prevention strategy to be implemented on the Chinese Weibo (equivalent to Twitter). Braithwaite et al. [13] demonstrated that machine learning algorithms are efficient in differentiating people from those who are and who are not at suicidal risk. Sueki et al. [14] studied the suicidal intent of Japanese Twitter users in their 20s, and they stated that language framing is important for identifying suicidal markers in the text. Additionally, O'Dea et al. [15] showed that the use of human code and automatic learning machines (LR, SVM) based on TF-IDF features can distinguish colors of suicidal messages. Wood et al. [16] identified 125 Twitter users followed their tweets preceding the data available before their suicide attempt. Using simple and linear classifiers, they found 70% of the users with a suicide attempt and identified their gender with 91.9% accuracy. Sawhney et al. [17] improved the performance of the Random Forest (RF) classifier for the identification of suicide ideation in tweets. Logistic regression classification algorithms applied by Aladag et al. [18] showed promising results in detecting suicidal content with an 80–92% accuracy rate.

With recent advances in deep learning in language processing, the use of deeper learning techniques has brought new benefits for investigating suicidal ideation, making machine learning better. Recurrent Neural Networks (RNNs) are specially designed for modelling sequences [19].Sony et al. This study demonstrates the advantages and potential of C-LSTM-based models for suicide detection compared to other deep learning and machine learning methods. Ji et al. [20] compared the LSTM classifier with five other machine learning models and demonstrated the feasibility and effectiveness of these models.

Recently, CNN neural networks with convolutional, nonlinear, and pooling layers have been successfully applied to a wide range of NLP tasks and have proven to gain better performance than traditional NLP methods [19]. Matsumoto et al. [21] proposed an efficient hybrid model which combines a fast deep-learning model with an initial information retrieval model to effectively and efficiently detect suicide ideation. According to all the studies done by different scholars, deep learning remains an optimal technique in suicide ideation yet it remains unexplored. In this study, a novel deep learning framework is proposed that effectively and accurately detects suicide ideation in modern society.

2.1 DATA SET

The dataset contains both primary and secondary data collected from different sources in Kenya that include the Department of Criminal Investigation-DCI and local websites including www.bonga.or.ke/share, https://www.fuzu.com and social platforms such as Twitter and Facebook for a period nine months. To

preserve the users' privacy and anonymity, their personal information is replaced with a unique ID. The dataset for this test comprised training, and testing data. In the training and Testing data, each of them has 8 columns: TID(Text_id), text_data, Gender, Month, AgeRange, SuicideAttempts, Region and class (suicidal, non–suicidal),
Table 1 shows examples of the data set.

TID	Text-data	class	AgeRange	Gender	Month2022	Region	Attempts
001	2 pm on Friday my life will be over I'll finally be free.	Suicide	15-24	Male	January	Coast	3
002	I will vote for my candidate come August 9th	Non - suicidal	25-34	Female	July	Central	1
003	Ndugu yangu Nimechoka na hii dunia	Suicide	35-44	Male	June	Rift-Valley	4
004	I will kill myself one day	Suicide	45-54	Male	September	Nyanza	6

Figure 1 A look on the dataset

III. METHODOLOGY

The purpose of the study is to implement a BERT deep learning classifier to improve the performance of language modelling and text classification for detecting suicide ideation based on the data set provided. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained deep learning model that functions on the sequence-to-sequence learning of text. BERT models are performing well in understanding textual data on suicide identification Martinez-Castano et [22]. The types of BERT models differ based on the number of transformer layers, self-attention layers, number of parameters, types of fine-tuning, masking, and word embedding. In the experiment, the study presents a technical description of approaches using various approaches in the BERT technique.. The Training phase consists of the following steps: Pre-Processing, preparing the input sentence for the BERT encoder, and the words are converted to tokens with input ids and tags using a standard tokenizer. The labels are also encoded and assigned weights based on the input data distributed. BERT Encoding: The next step is to apply the pre-trained model to the current input data. This step tries to map the vector to words in the context with high precision. The BERT layer is fine-tuned by adjusting the learning rate and optimization. Dropout: The dropout function tends to adjust the weights assigned in each layer to normalize the weights among the words. This layer is significant as it differentiates the words related to depression vs the rest of the words. This step also distributes the weight to avoid overfitting. Classifier: The final step is to map the previous layer with the words to the two classes ('Suicide', and 'non-suicide') which forms the labels for this task.

3.1 Pre-Processing

The dataset provided has a mixture of raw text obtained directly from social media platforms and online repositories. This data needs cleaning so that it can be used by the model effectively to improve the

prediction rate. Therefore, a series of refining works were done to the dataset before actual training and testing which includes the removal of duplicates among others

3.1.1 Removal of stop words

where the length of texts on the dataset is quite long. To reduce the model size and improve accuracy, we remove the stop words. The list of stop words is obtained from the Python Natural Language Toolkit and stop words packages. Shorter sentences are post-padded and the longer ones are post-truncated.

3.1.2 Tokenization

The words are tokenized using the TensorFlow tokenizer with a vocabulary size of 1024 words. Since the small BERT model is used for model building, tokenization is also performed by the pre-trained model. In addition, transfer learning also includes a preprocessing layer that takes care of tokenization. Tokenization is a crucial step for BERT-based models as they prepare the data to be processed by segmenting them between the [CLS] and [SEP] tags.

3.1.3 One-hot encoding labels

The label binarizes encoder function is used to encode our labels as one-hot vectors to perform multiclass classification. In the case of our BERT model, the tf.one_hot function is used to encode the labels into tensors with a depth of 3[22].

3.2 Fine-tuned BERT Model

The pre-trained small_Bert model for classification is retrieved from TF Hub. However, to adapt the model to the given dataset, the researcher added layers to the model i.e., the dropout and the dense layer. The model is trained with 10 epochs with a learning rate of 2e-5. The number of epochs and learning rate may be scaled up however this will add overhead to the performance. Figure 1 below shows the parameters used for BERT.

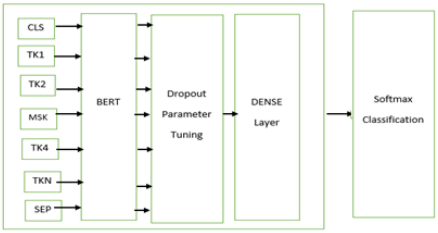
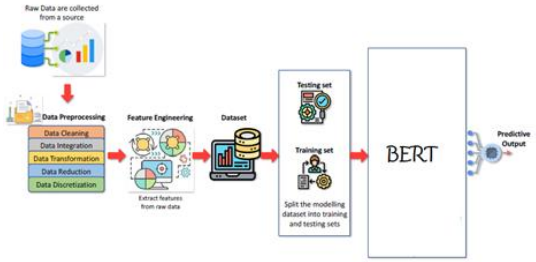


Figure 2 BERT parameters

Figure 2 shows a general overview of the proposed framework. It consists of data pre-processing techniques and features extraction with NLP techniques (TF-IDF, BOW, Statistical Features, and word embedding) employed to encode the words to be further processed by the BERT model.



3.3 Experimental parameter setting

Different hyperparameters may have different effects on the experimental results. Although parameter tuning itself is not the main research content of this paper, for the sake of fairness, this paper considers the overall effect of the experiment and finally determines the hyperparameters of BERT. Some hyperparameters of the BERT model are shown in Table.

Hyperparameter	characterization	Value
Ƞ	Learning rate	2e-5
Epoch	Number of iterations	10
Hidden	Number of hidden units	200
Batch	Batch size	16
D	Word vector size	40

Table 1 Parameters used for BERT model3.3.1 Experimental environment setting

The hardware configuration used in this paper to experiment has the following specifications; CPU Intel Core i7-4600U and clock speed of 2.70GHz with 8GB of installed RAM; it runs on Windows 10 operating system; while the development environment used is Google Colab running on a GPU device; libraries used include PyTorch 2.0, Keras and transformers.

3.4 Performance Evaluation parameter

The majority of state-of-the-art sentiment analysis makes use of accuracy, F1 score, and precision. Sentiment analysis using deep learning architectures:

a review utilizes recall and accuracy as performance metrics. These metrics are as follows;

Precision - Precision is defined as the ratio of correctly classified positive samples to the total number of samples predicted as positive. This metric can be used to indicate the strength of the prediction.

Recall - Recall is also known as sensitivity. It is defined as the ratio of actual positive instances out of the total number of positive instances present in the classification. It measures the misclassifications done by the model.

F1 score - F1 score is the harmonic mean of Recall and Precision. It is the most used metric after Accuracy. It is used when we are unable to choose between Precision or Recall

IV. RESULTS AND DISCUSSION

Since we have a limited dataset with a few numbers of training samples, many iterations may lead to overfitting problems, and also few iterations may be insufficient for the model to learn all the features. Therefore, the dataset was split into 10% for testing and 90% for training and validation.

Precision, Recall and F1- score precision measure techniques were selected to be used as evaluation indicators. The experiment results are in the table below

Model	Precision	Recall	F1
BERT	0.91	0.87	0.90
MultinomialNB	0.68	0.68	0.63
BernoulliNB	0.67	0.67	0.74
ComplementNB	0.77	0.50	0.55

Table 2 Comparison of Results from BERT and Naïve Bayes models

Tasks like text classification, machine translation and language modelling rely greatly on the use of sequential modelling. Even though RNN and LSTM were perfect for these operations, the computation time taken due to the processing of single input at a time led to the popularity of transformer models. Thus, the use of BERT is justified in many classification problems due to their efficiency of being pre-trained in large datasets and being deeply bidirectional.

1. BERT's transformer architecture and model size helped it learn the features better. One instance where it was able to predict the label correctly is given below

"., I always make the wrong choices and I can't get myself out of the depressed state of mind and I feel like my life is over ..."

The above sentence was tagged as moderate as the context in which the words 'depressed' and 'failure' are used is also studied. The other two models had labelled it as severe.

2. The BERT model works well for context with opposite terms as it works on parallel processing of layers. Since contextual words form the key feature in this learning model, the absence of words directly related to depression in the context had an impact on the performance of the model. For instance, the below sentence though looks like a serious case of self-harm, was tagged moderate due to the lack of words directly related to depression.

"... I'll finally get to take a breather. Today I think I'll die."

3. BERT models work well on diverse and representative training data. Pretraining the BERT model on a large dataset enables it to learn general patterns in the language. The sentence below which was tagged moderate was possibly due to the absence of explicit words related to depression. This indicates that the model's prediction can be influenced by the availability and relevance of specific training examples.

"I just can't move the talking stages to the next step"

4. On the average evaluation metrics, BERT outperforms the Naïve Bayes models in terms of precision, recall and F1 score. MultinomialNB and BernoulliNB exhibit moderate performance, while ComplementNB shows high precision but lower recall and F1 score. This shows that the BERT model can work effectively in classifying suicide ideation texts compared with different Naïve Bayes models.

5. On average there was an equal number of suicide attempts across all regions with significant differences between males and females. But in the Coast region, there were 6 male suicide attempts and 1 female suicide attempt. Also, according to the dataset male suicide attempts were significantly higher than female suicide attempts. The figure below shows the summary of the dataset.

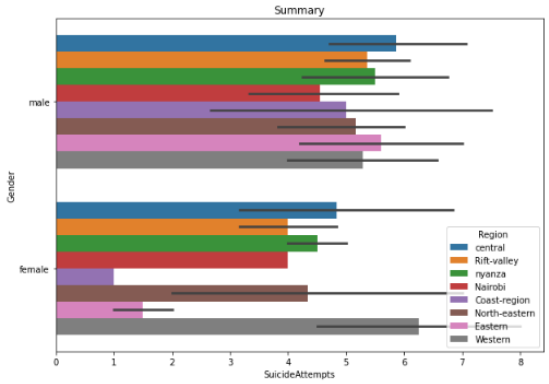


Figure 3 Summary of the dataset used

6. There was a higher number of suicide attempts among people aged 25 15 and 34 years which is a group made of the youth. This made more the half of the number of suicide attempts. Also, for most age groups the number of male suicide attempts was higher compared to the number of female suicide attempts. The figure below shows the number of suicide attempts per age group.

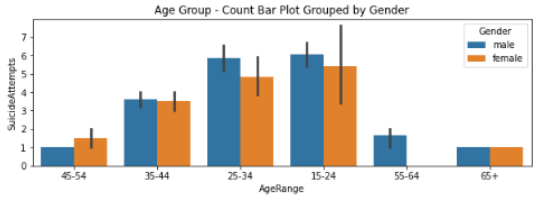


Figure 4 Age group counts

7. The central and Western regions were the leading regions in number of suicide attempts. With other regions having nearly the same number of suicide attempts. The figure below shows the number of suicide attempts per region.

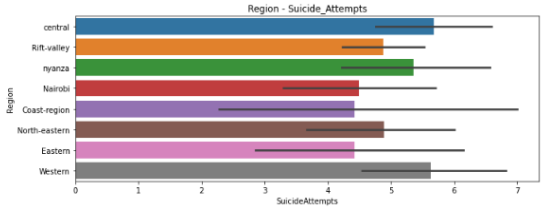


Figure 5 Number of Suicides per region

8. The period between September and December had more suicide attempts compared to other months in 2022. The period between June and August recorded the lowest number with June 2022 having the lowest number of suicide attempts.

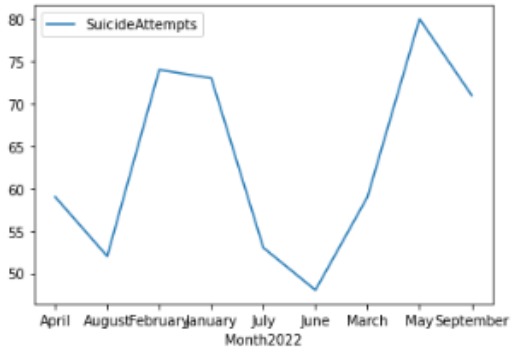


Figure 6 Time series of suicide attempts from April to September 2022

9. A word cloud analysis on texts from people who had attempted suicide show that some of the most common words in these text messages are *feel, life, people, want, hate, and social.* This shows what people talked about before a suicide attempt. The figure below shows a word cloud analysis on the dataset with attempted suicide.

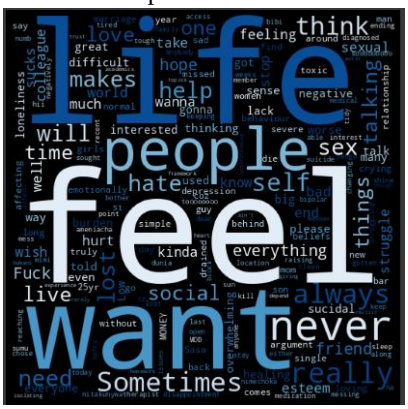


Figure 7 Most common words in texts which were classified as positive

CONCLUSION AND FUTURE WORK

Detecting the signs of depression from just a collection of words is a huge accomplishment in the field of Artificial Intelligence. This is because we have come to a point where even machine identifies the emotions of a person from the words he or she speaks. This work produces one such model that is capable of detecting depression by exploiting the efficiency of the BERT transformer model. Further, for a model pre-trained in a completely different context, the fine-tuned BERT model performed reasonably well when compared to other Deep learning models such as LSTM and Embedded models. The model could be enhanced further to address superficial and unclear words by understanding the context better and by redistributing the weights among words in the encoder layer.

According to this research, there were more males with suicide attempts than women, most of the intents age range from fifteen years to 35 years of age and commonly this happens in central Kenya, it will be interesting to know why this is rampant in this region, for more crystal-clear research more data on intents will, be very useful for future research

REFERENCES

[1] D. Gunnell et al., "Suicide risk and prevention during the COVID-19 pandemic," The Lancet Psychiatry, vol. 7, no. 6, pp. 468–471, 2020,

[2] N. J. Miros, Depression, anger, and coping skills as predictors of suicidal ideation in young adults: Examination of the diathesis-stress-hopelessness theory. Hofstra University, 2000

[3] E. D. Klonsky, "The functions of deliberate self-injury: A review of the evidence," Clinical psychology review, vol. 27, no. 2, pp. 226–239, 2007

[4] M. De Choudhury and E. Kiciman, "The language of social support in social media and its effect on suicidal ideation risk," 2017

[5] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using deep learning," Algorithms, vol. 13, no. 1, p. 7, 2019

[6] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural language processing of social media as screening for suicide risk," Biomedical informatics insights, vol. 10, p. 1178222618792860, 2018

[7] J. Jashinsky et al., "Tracking suicide risk factors through Twitter in the US," Crisis, 2014,

[8] J. Scourfield et al., "The response in Twitter to an assisted suicide in a television soap opera.," Crisis: The Journal of Crisis Intervention and Suicide Prevention, vol. 37, no. 5, p. 392, 2016

[9] I. M. Wasserman, "Imitation and suicide: A reexamination of the Werther effect," American sociological review, pp. 427–436, 1984

[10] R. A. Fahey, T. Matsubayashi, and M. Ueda, "Tracking the Werther Effect on social media: Emotional responses to prominent suicide deaths on twitter and subsequent increases in suicide," Social Science & Medicine, vol. 219, pp. 19–29, 2018

[11] B. Desmet and V. Hoste, "Emotion detection in suicide notes," Expert Systems with Applications, vol. 40, no. 16, pp. 6351–6358, 2013

[12] X. Huang, L. Zhang, D. Chiu, T. Liu, X. Li, and T. Zhu, "Detecting suicidal ideation in Chinese microblogs with psychological lexicons," 2014

[13] K. H. Gordon et al., "The reinforcing properties of repeated deliberate self-harm," Archives of Suicide Research, vol. 14, no. 4, pp. 329–341, 2010

[14] H. Sueki, "The association of suicide-related Twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan," Journal of affective disorders, vol. 170, pp. 155–160, 2015

[15] B. O'dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on Twitter," Internet Interventions, vol. 2, no. 2, pp. 183–188, 2015

[16] Z. Wood-Doughty, N. Andrews, R. Marvin, and M. Dredze, "Predicting Twitter user demographics from names alone," 2018.

[17] R. Sawhney, P. Manchanda, R. Singh, and S. Aggarwal, "A computational approach to feature

extraction for identification of suicidal ideation in tweets," 2018

[18] A. E. Aladağ, S. Muderrisoglu, N. B. Akbas, O. Zahmacioglu, and H. O. Bingol, "Detecting suicidal ideation on forums: proof-of-concept study," Journal of medical Internet research, vol. 20, no. 6, p. e9840, 2018

[19] R. Sawhney, A. Malik, S. Sharma, and V. Narayan, "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease," Decision Analytics Journal, vol. 6, p. 100169, 2023

[20] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," IEEE Transactions on Computational Social Systems, vol. 8, no. 1, pp. 214–226, 2020

[21] K. Kato et al., "Clinical features of suicide attempts in adults with autism spectrum disorders," General hospital psychiatry, vol. 35, no. 1, pp. 50–53, 2013

[22] R. Martínez-Castaño, A. Htait, L. Azzopardi, and Y. Moshfeghi, "BERT-based transformers for early detection of mental health illnesses," 2021

[23] S. Adarsh and B. Antony, "SSN@ LT-EDI-ACL2022: Transfer Learning using BERT for Detecting Signs of Depression from Social Media Texts," 2022