EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Suicidal Ideation Detection and Influential Keyword Extraction from Twitter using Deep Learning (SID)

Xie-Yi. G.[1,*]

[1]Student, BSc (Hons) in Computer Science School of Computer Science, Asia Pacific University of Technology and Innovation, Malaysia

Abstract

INTRODUCTION: This paper focuses on building a text analytics-based solution to help the suicide prevention communities to detect suicidal signals from text data collected from online platform and take action to prevent the tragedy.
OBJECTIVES: The objective of the paper is to build a suicide ideation detection (SID) model that can classify text as suicidal or non-suicidal and a keyword extractor to extracted influential keywords that are possible suicide risk factors from the suicidal text.
METHODS: This paper proposed an attention-based Bi-LSTM model. An attention layer can assist the deep learning model to capture influential keywords of the model classifying decisions and hence reflects the important keywords from text which highly related to suicide risk factors or reason of suicide ideation that can be extracted from text.
RESULTS: Bi-LSTM with Word2Vec embedding have the highest F1-score of 0.95. Yet, attention-based Bi-LSTM with word2vec embedding that has 0.94 F1-score can produce better accuracy when dealing with new and unseen data as it has a good fit learning curve.
CONCLUSION: The absence of a systematic approach to validate and examine the keyword extracted by the attention mechanism and RAKE algorithm is a gap that needed to be resolved. The future work of this paper can focus on both systematic and standard approach for validating the accuracy of the keywords.

Keywords: attention mechanism, Bi-LSTM, deep learning, NLP, text classification.

Received on 17 January 2024, accepted on 6 May 2024, published on 13 May 2024

doi: 10.4108/eetpht.10.6042

1. Introduction

According to the World Health Organization (WHO), over 700,000 people die from suicide every year [1]. Suicide not only affects specific victims, their surviving families, and friends, it also increases the financial burden on society. The CDC noted that suicide and nonfatal self-harm cost the country close to $490 billion in 2019 [2]. Suicidal ideation (SI) is the starting point of the overwhelming tragedy. The ability to recognize SI as a signal of a potential victim and establish strategies that address suicide risk factors are the crucial breaking points in improving suicide prevention and care. Although the authorities such as the World Health Organization (WHO), National Institute of Mental Health (NIMH), and other private communities has been making effort to prevent suicide and raise suicide awareness, victims may be silent and passive in seeking help due to stereotypes and social stigmas. Moreover, traditional approaches such as self-reported ratings and current clinical risk assessments are not designed to assess rapid changes in SI [3].

With the rise of Internet and social media, younger generations who have communication gaps with their caretakers and isolated from their peers, groups that

*Corresponding author. Email: tp056669@apu.edu.my

struggles financially to meet with psychiatrist, and other middle age group that are hesitant to reach out for help due to social status and stigmas tend to turn towards social media such as Facebook, Twitter, Reddit, Instagram, WhatsApp, Weibo and more to express their emotions and seek support [4]. Public forums such as Reddit, Quora and Tumblr that support anonymous participation provide a "safe-space" for users to discuss socially stigmatized topics, including suicidal feelings and thoughts [3][5].

In relation, a trend of leaving suicide notes and suicidal ideation comments on social media [5], [6], [7], [8] has been seen throughout the years, suicidal ideation detection using social media data will be a breakthrough to understand a person's mental health status and help to prevent suicidal behaviour.

This paper proposed to build a Suicide Ideation Detection (SID) model can detect suicide ideation and extract keywords that reflect risk factors from the user-generated text collected from online forum.

2. Materials and Methods

2.1. Deep Learning Approaches for Text Classification

Recent advancements in NLP have leveraged deep learning (DL) to improve the state of the art for language modelling [9] and text classification [5], [10], [11]. In health informatics, experimental studies are implementing DL models to classify medical notes, electronic health records (EHR), progress notes, etc. [12], [13]. DL is a large artificial neural network made up of numerous processing layers to pick up relevant features using interrelating layers of information-processable neurons, mimicking human neural architecture in a knowledge acquisition process [14]. DL architectures provide significant advantages over ML approach for text classification as they perform at very high accuracy with lower-level engineering and processing, eliminating disadvantages such as sparse feature vectors, dimensional explosion, and troublesome feature extraction [15], [16]. Moreover, the highly progressive and greedy-natured algorithm of DL enables machines to take in more complex information and handle non-linear relationships between input and output data [17].

2.1.1 LSTM

A standard LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The 3 gates take control of the flow of information into and out of the cell while the cell collects values over arbitrary time intervals [10], [18]. The capability of LSTM to capture long-term dependencies facilitates the architecture to attain contextual understanding behind the input [19] has made significant improvements in the performance of NLP text classification models. However, the researchers [20] have pointed out that LSTM has limits in contextualise information from the future tokens and is not sufficient to extract the local contextual information despite being great at handling variable-length sequences[20]. Furthermore, LSTM does not have the ability to differentiate the relevance between each part of the document [12].

2.1.2 Bi-LSTM

Bi-LSTM (Bidirectional LSTM) is an enhancement of LSTMs. The fact that each training sequence is presented forwards and backwards to two independent recurrent nets, connected to the same output layer in Bidirectional Recurrent Neural Networks (BRNN) reflects its ability to comprehend sequential knowledge about all points before and after each point in each sequence [20], [21].

Many researchers have made efforts to improve LSTM based model performance by using Bi-LSTM model and their result has proven the fact that by adding one more LSTM layer, which reverses the direction of information flow allows the model to produce more accurate result [13], [15], [20]. Haque et al. [15] suggest that it is resulted by the ability of Bi-LSTM to efficiently access relevant information across lengthy tweets through the forward–backward dependencies from feature sequences aids in resolving gradient disappearance and helps in long-term dependence. The ability of Bi-LSTM to obtain both historical information and future information through the bidirectional propagation mechanism helps to achieve better performance in NLP tasks [22].

2.2. Attention Mechanism as Influential Keywords Extractor

The concept of the attention mechanism is to allow the decoder to utilize the most relevant parts of the input sequence in a flexible manner, which is reflected through the weighted combination of all encoded input vectors, with the highest weights assigned to the most relevant vectors [23], [24]. Besides offering a performance gain [25], the attention mechanism is implemented as an approach to interpret the behaviour of neural architectures [12], [26], [27]. For instance, the weights generated by attention could give insight into which relevant information or irrelevant features have been discarded or input by the neural network. In fact, the knowledge stored in neural networks is in form of numeric value that is meaningless without interpretations. Therefore, visualizing highlights of attention weights could be instrumental in analysing the outcome of neural networks [12], [27], [28]. In the NLP context, the visualization of attention weights can help researchers to evaluate the model's performance. It can also be used as a tool to identify influential keywords that affect the classifier in making its decision. Inspired by the

work of Li et al. [25] and Tang et al. [13] on using attention mechanism to interpret the "blackbox" of neural networks, the SID model applies the similar concept to extract the influential keywords of the deep learning model that may reflects suicide risk factors. As shown in Fig. 1 below, adding an attention layer after the information processing layer and before the classification layer to attain the attention weights that influence the classifier's decision. With adequate mapping and visualization technique, the numeric output could be transformed into meaningful words to be further analysed.

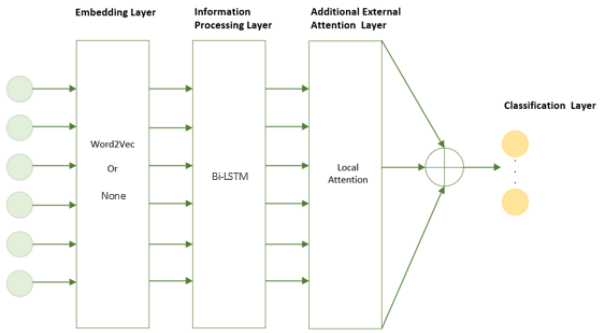

Figure 1. Attention-based deep learning model architecture for SID and Influential Keyword Extraction inspired by Tang et al.[13]

To compare the keywords that are extracted by the attention mechanism, RAKE algorithm is implemented as another approach to extract the important keywords from each sample that falls under suicide class. The result of both methods to extract possible suicide risk factors (keywords) will be displayed as word cloud in section below.

2.3. Workflow of SID Model

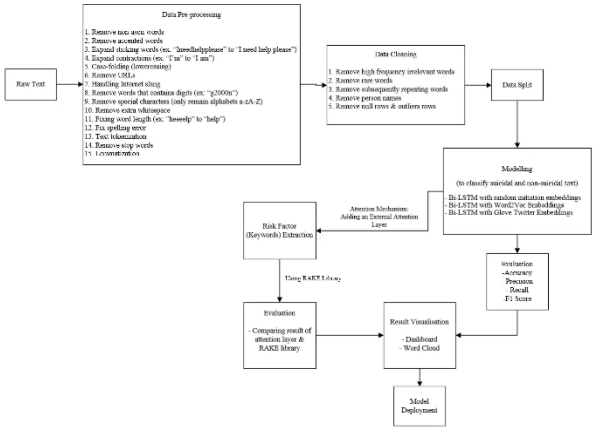

Figure 2. Visualization of Workflow for SID model

Figure 2 shows the workflow initiated by preprocessing the raw text, including removing non-ascii words, special symbols, URLs, hashtags, expanding contractions, handling Internet slangs, spelling corrections, tokenization, lemmatization and stop word removal. Data cleaning cleans out left-over noise in dataset, including removing irrelevant words, drop out rows that contains only symbols or rows that contain no meaningful string via several iterations of checking. Outlier samples that have extra-long text are also discarded during this process for optimized model training [29].

Bi-LSTM model acts as the suicide ideation detector with the functionality to classify text input into suicidal class or non-suicidal class. Several Bi-LSTM model variants are created using different word embeddings, namely Word2Vec embeddings and GloVe embeddings. Regularization techniques, including dropout layer, L2 regularizer, optimizer, adjusting learning rate and weight decay is implemented to prevent overfitting or underfitting [30]. The model with highest model performance for each model variant is then added on with the attention mechanism that acts as the suicide risk extractor.

(a) Bi-LSTM model variants:
- Model 1: Random Initialization
- Model 2: Custom Word2Vec (256-dimensions)
- Model 3: Pre-trained GloVe (200-dimensions)

The top 10 attention weights of each text sample that falls under suicidal class will be captured and transformed into meaningful keywords. The steps to interpret and visualize attention weights output are shown below.

(b) Interpreting & visualizing top attention weights:
- (i) Get the attention weights for each token.
- (ii) Get word index mapping from tokenizer.
- (iii) Map the top 10 attention weights (index) to corresponding word.

3. Results and Discussion

This section discusses on the experimental result of applying different regularization techniques on the Bi-LSTM model variants, selecting out the best Bi-LSTM model of each model variants to be compared with the LSTM model performance from Jiayi et al.'s work[31], which acts as the base model. model performance before and after applying the attention mechanism is also examined.

Table 1. Model performance of Bi-LSTM model variants with different optimizer & regularization

Bi-LSTM Variants	Optimizer	Accuracy	Precision	Recall	F1-Score
Random Initialise	Adam, L2	0.93	0.94	0.91	0.93
	AdamW	**0.94**	**0.93**	**0.94**	**0.94**
Word2Vec embedding	Adam	**0.95**	**0.95**	**0.95**	**0.95**
	Adam, L2	0.93	0.94	0.92	0.93
	AdamW	0.94	0.94	0.95	0.94
GloVe Twitter embedding	Adam	**0.95**	**0.95**	**0.95**	**0.95**
	Adam, L2	0.93	0.92	0.94	0.93
	AdamW	0.94	0.95	0.93	0.94

Table 1 shows two different optimizers- Adam and AdamW, and kernel regularizer (L2) are implemented as regularization. Kernel regularization adds penalties to the kernel layers to reduce weights of the neural network [32]. Adding penalty factors to the weights assists the neural network to expedite its update process, accelerating model convergence with proper weights for the next update. The exclusion of updating the bias is beneficial for obtaining lighter models to prevent the overfitting of complex neural networks [33]. The combination of Adam optimizer with L2 regularizer is often seen to be implemented in deep learning models to help reduce overfitting when training model [34], [35].

On the other hand, AdamW is an improve version of Adam proposed by authors Loshchilov and Hutter [36]. It improves model generalization by decoupling the weight decay from the gradient-based update. The authors [34], [36] argues that better training loss can be obtained compared to Adam as the weight decay is performed only after controlling the parameter-wise step size in AdamW, ensuring the regularization term does not end up in the moving averages but only the proportional weight itself. In this case, the combination of Adam optimizer with L2 regularizer works most effectively to help the model obtain a good fit. It can be observed that adding kernel regularizer will decrease the F1-score but will help the model to obtain a good fit learning curve. Each model variant with the highest F1-score is compared with the base models.

Table 2. Model comparison table of base models and Bi-LSTM models

Model Variants	Accuracy	Precision	Recall	F1-Score
LSTM Model 1	0.87	0.86	0.8	0.83
LSTM Model 2	0.93	0.94	0.86	0.90
LSTM Model 3	0.88	0.92	0.76	0.83
Bi-LSTM Model 1	0.94	0.93	0.94	0.94
Bi-LSTM Model 2	**0.95**	**0.95**	**0.95**	**0.95**
Bi-LSTM Model 3	**0.95**	**0.95**	**0.95**	**0.95**

Table 2 shows that each Bi-LSTM model variant has higher model performance when compared to the LSTM models [31]. As suggested by previous studies [12], [13], [15], [20], [22], [26], Bi-LSTM which provide forward and backward propagation allows the model to capture more information and addresses non-linear relationship in text data, which prompts the model to obtain highest F1-score of 0.95. However, these models that have the highest F1-scores show overfitting in their learning curve. Hence, it is to be debated whether a higher F1-score will produce a better performance or a good fit learning curve with lower F1-score can achieve better classification performance when dealing with new and unseen data.

Table 3. Model performance of attention based BiLSTM model variants

Attention based Bi-LSTM Model Variants	Accuracy	Precision	Recall	F1-Score
Random Initialisation	0.93	0.93	0.93	0.93
Word2Vec embeddings	**0.94**	**0.95**	**0.93**	**0.94**
GloVe embeddings	**0.94**	**0.94**	**0.94**	**0.94**

Table 3 shows the model performance of the attention-based Bi-LSTM model variants. All the attention-based models apply L2 regularizer. The presence of kernel regularizer in model training results in a good fit model. Although the F1-score dropped 0.1 when being compared with model trained without kernel regularizer (0.95), if being compared with the models that applies kernel regularizer in model training as shown in Table 1, the model performance increases by 0.1, which aligns with the research [13], [20], [26], [27], suggesting that attention mechanism could help the model to focus on more crucial part of the text to produce good classification result.

Several rounds of testing with new unseen data are carried out on the model to examine the model's performance when dealing with new and unseen data. The tested model is listed below.

Tested Model:

- Word2Vec Bi-LSTM [accuracy: 0.95, F1-score: 0.95]
- Attention-based Word2Vec Bi-LSTM [accuracy: 0.94, F1-score: 0.94]
- Attention-based GloVe Bi-LSTM [accuracy: 0.94, F1-score: 0.94]

The testing examples include sentences that contain topic sensitive words such as “depressed”, “died”, “depression” but is not suicidal as well as sentences that does not have obvious words that relates to suicide ideation but is categorized under suicidal class.

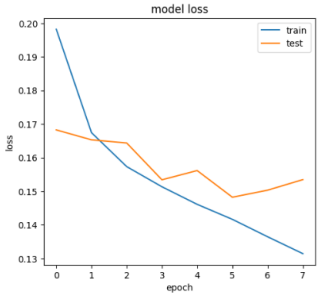

Figure 3. Bi-LSTM Word2Vec loss graph (higher F1-score, 0.95 but overfitting)

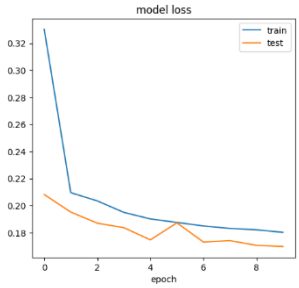

Figure 4. Attention-based Bi-LSTM Word2Vec model loss graph (lower F1-score, 0.94 with good fit)

The observation gain is the attention-based model with lower F1-score and accuracy but with a good fit in its learning curve can perform better when dealing with the new unseen sample data. The fact that overfitting models tend to be overly sensitive [30] as it picked up noises which affect the model performance when dealing with new data. It can be summarized that a model with good fit as shown in Figure 4 produces more accurate classification results when being compared to a model with higher F1-score but has an overfitting learning curve as shown in Figure 3.

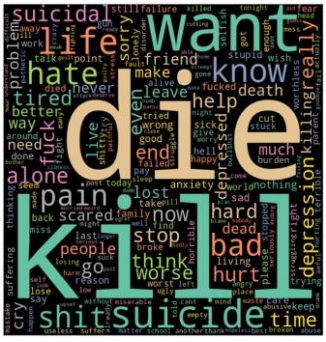

Figure 5. Word cloud of influential keywords extracted by attention mechanism

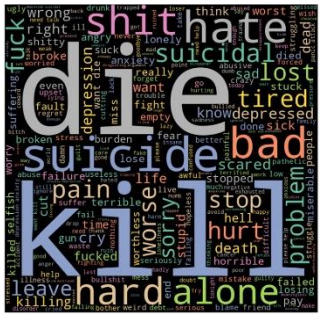

Figure 6. Word cloud of keywords extracted by RAKE algorithm

From Figure 5 and Figure 6, it can be observed that many similar keywords that reflects suicide risk factors using both methods, reflecting attention mechanism’s potential as a keyword extractor and its ability to help reserachers understand the model’s decision. The keywords extracted via RAKE algorithm contains greater amount of adjectives and verbs which carries crucial information and sentiment. However, both of the methods often pick up words that is less important or not reflecting the cause of suicide ideation. Moreover, the absence of a systematic approach to validate and examine the keyword extracted by the attention mechanism and RAKE algorithm is a gap that needed to be resolved.

3. Conclusion

The SID model is a text analytic-based solution that can detect suicidality in user input text and extract influential keywords which is the possible suicide risk factor. This can help the suicide prevention communities to detect suicidal signals and take action to prevent the tragedy. The most crucial concern when developing a DL solution is the quality of the dataset as DL works on a garbage-in-garbage-out principle. The volume of the data as well as the variation of data will highly impact on the model’s performance. While accuracy and F1-score are important metrices used to evaluate the model performance, the learning curve also gives many insights of the model performance. A high accuracy but overfitted model performs not as good as model that has a good fit when dealing with new data as it cannot generalize well. Limited understanding on the internal operation of the attention mechanism and its approach to initialize and distribute the weights for each input affects the model effectiveness to extract the accurate keywords from the text. Hence, techniques to exploit the full potential of attention mechanism for keyword extraction are to be further explored. Transformers approaches such as BERT, XLNet, and RoBERT that implement based on attention is a great topic to approach. Furthermore, there is still a gap in having a standard and systematic approach to validate the accuracy and correctness of extracted keywords in terms of

the relevance of the selected keywords with the cause of the specific suicide ideation text. Hence, future work can focus on the mentioned areas.

Acknowledgements.
Major thanks to Mr. Raheem Mafas (Asia Pacific University of Technology and Innovation) for the supervision.

References

[1] "World Health Organization (WHO)," "Home/Newsroom/Fact sheets/Detail/Suicide," WHO, 2021. https://www.who.int/news-room/fact-sheets/detail/suicide (accessed Jun. 22, 2023).
[2] "Centers for Disease Control and Prevention (CDC)," "Facts About Suicide," CDC, 2022. https://www.cdc.gov/suicide/facts/index.html (accessed Apr. 23, 2023).
[3] E. D. Ballard, J. R. Gilbert, C. Wusinich, and C. A. Zarate, "New Methods for Assessing Rapid Changes in Suicide Risk," Front Psychiatry, vol. 12, Jan. 2021, doi: 10.3389/fpsyt.2021.598434.
[4] M. Chatterjee, P. Kumar, P. Samanta, and D. Sarkar, "Suicide ideation detection from online social media: A multi-modal feature based technique," International Journal of Information Management Data Insights, vol. 2, no. 2, p. 100103, Nov. 2022, doi: 10.1016/j.jjimei.2022.100103.
[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, Dec. 2019, doi: 10.3390/a13010007.
[6] T. M. DeJong, J. C. Overholser, and C. A. Stockmeier, "Apples to oranges?: A direct comparison between suicide attempters and suicide completers," J Affect Disord, vol. 124, no. 1–2, pp. 90–97, Jul. 2010, doi: 10.1016/j.jad.2009.10.020.
[7] B. Desmet and V. Hoste, "Emotion detection in suicide notes," Expert Syst Appl, vol. 40, no. 16, pp. 6351–6358, Nov. 2013, doi: 10.1016/j.eswa.2013.05.050.
[8] T. Zhang, A. M. Schoene, and S. Ananiadou, "Automatic identification of suicide notes with a transformer-based deep learning model," Internet Interv, vol. 25, p. 100422, Sep. 2021, doi: 10.1016/j.invent.2021.100422.
[9] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing," Aug. 2017.
[10] D. Raihan, "Deep learning techniques for text classification," Nanyang Technological University, 2021.
[11] F. Wei, H. Qin, S. Ye, and H. Zhao, "Empirical Study of Deep Learning for Text Classification in Legal Document Review," Apr. 2019, doi: 10.1109/BigData.2018.8622157.
[12] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," BMC Med Res Methodol, vol. 22, no. 1, p. 181, Dec. 2022, doi: 10.1186/s12874-022-01665-y.
[13] M. Tang, P. Gandhi, M. A. Kabir, C. Zou, J. Blakey, and X. Luo, "Progress Notes Classification and Keyword Extraction using Attention-based Deep Learning Models with BERT," Oct. 2019.
[14] J. Brownle, "TensorFlow 2 Tutorial: Get Started in Deep Learning with tf.keras," Machine Learning Mastery, Aug. 02, 2022.
[15] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, "A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning," Technologies (Basel), vol. 10, no. 3, p. 57, Apr. 2022, doi: 10.3390/technologies10030057.
[16] H. Wang and F. Li, "A text classification method based on LSTM and graph attention network," Conn Sci, vol. 34, no. 1, pp. 2466–2480, Dec. 2022, doi: 10.1080/09540091.2022.2128047.
[17] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," Electronic Markets, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.
[18] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," Artif Intell Rev, vol. 53, no. 8, pp. 5929–5955, Dec. 2020, doi: 10.1007/s10462-020-09838-1.
[19] N. Nuzulul Khairu, "Text Messages Classification using LSTM, Bi-LSTM, and GRU," MLearning.ai. Medium, Aug. 21, 2022. https://medium.com/mlearning-ai/the-classification-of-text-messages-using-lstm-bi-lstm-and-gru-f79b207f90ad (accessed May 10, 2023).
[20] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," Neurocomputing, vol. 337, pp. 325–338, Apr. 2019, doi: 10.1016/j.neucom.2019.01.078.
[21] E. Zvornicanin, "Differences Between Bidirectional and Unidirectional LSTM," Baeldung, Jun. 08, 2022. https://www.baeldung.com/cs/bidirectional-vs-unidirectional-lstm (accessed May 18, 2023).
[22] A. Agarwal, "Sentiment Analysis using Bi-Directional LSTM," LinkedIn Article, May 20, 2020. https://www.linkedin.com/pulse/sentiment-analysis-using-bi-directional-lstm-ankit-agarwal (accessed May 18, 2023).
[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," Sep. 2014.
[24] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," Apr. 2020.
[25] A. Chadha and B. Kaushik, "A Hybrid Deep Learning Model Using Grid Search and Cross-Validation for Effective Classification and Prediction of Suicidal Ideation from Social Network Data," New Gener Comput, vol. 40, no. 4, pp. 889–914, Dec. 2022, doi: 10.1007/s00354-022-00191-1.
[26] B. Jang, M. Kim, G. Harerimana, S. Kang, and J. W. Kim, "Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism," Applied Sciences, vol. 10, no. 17, p. 5841, Aug. 2020, doi: 10.3390/app10175841.
[27] A. Galassi, M. Lippi, and P. Torroni, "Attention in Natural Language Processing," IEEE Trans Neural Netw Learn Syst, vol. 32, no. 10, pp. 4291–4308, Oct. 2021, doi: 10.1109/TNNLS.2020.3019893.
[28] J. Li, S. Zhang, Y. Zhang, H. Lin, and J. Wang, "Multifeature Fusion Attention Network for Suicide Risk Assessment Based on Social Media: Algorithm Development and Validation," JMIR Med Inform, vol. 9, no. 7, p. e28227, Jul. 2021, doi: 10.2196/28227.
[29] E. Snorrason, "Understanding Outliers in Text Data with Transformers, cleanlab, and Topic Modeling," Towards Data Science, Oct. 07, 2022.

https://towardsdatascience.com/understanding-outliers-in-text-data-with-transformers-cleanlab-and-topic-modeling-db3585415a19 (accessed May 20, 2023).

[30] M. Alam, "Avoid Overfitting with Regularization," Towards Data Science, Dec. 29, 2020. https://towardsdatascience.com/avoid-overfitting-with-regularization-6d459c13a61f (accessed May 28, 2023).

[31] A. S. Shih Win, G. Jia Yi, L. Zi Hui, L. Xiao, and Q. Yi Zhen, "Suicidal Text Detection in Social Media Post," 2021. Accessed: May 28, 2023. [Online]. Available: https://docs.google.com/viewer?url=https://raw.githubusercontent.com/gohjiayi/suicidal-text-detection/master/docs/Suicidal-Text-Detection_Report.pdf

[32] C. Pathak, "Kernel, Bias and Activity Regularizer : what, when and why," LinkedIn Articles, Mar. 16, 2021. https://www.linkedin.com/pulse/kernel-bias-activity-regularizer-what-when-why-chiranjit-pathak (accessed May 29, 2023).

[33] M. Darshan, "How do Kernel Regularizers work with neural networks?," Mystery Vault, Jun. 25, 2022. https://analyticsindiamag.com/kernel-regularizers-with-neural-networks/ (accessed May 29, 2023).

[34] S. Gugger and J. Howard, "AdamW and Super-convergence is now the fastest way to train neural nets," fast.ia, Jul. 02, 2018. https://www.fast.ai/posts/2018-07-02-adam-weight-decay.html (accessed May 29, 2023).

[35] Knowledge Transfer, "Adam optimizer with learning rate weight decay using AdamW in keras," Knowledge Transfer, Jan. 13, 2023. https://androidkt.com/adam-optimizer-with-learning-rate-weight-decay-using-adamw-in-keras/ (accessed May 29, 2023).

[36] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Nov. 2017.