Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/55/20241512

Research on the applicability of suicide tweet detection algorithms

Shuyu Cai[1,*], Hyundo Jung[2,5,†], Jiaqi Liu[3,6,†], Weikang Liu[4,7,†]

[1]School of Physical Science, University of California Irvine, 92617, CA, The United States
[2]Korea International School Jeju Campus, Seogwipo-si, South Korea
[3]University of California Santa Barbara, 93106, CA, The United States
[4]Lakehead University, 955 Oliver Rd, Thunder Bay ON P7B5E1, Canada

*Corresponding author: SCAI10@uci.edu
[5]harold.jung0914@gmail.com
[6]ljiaqi57@gmail.com
[7]wliu29@lakeheadu.ca
[†]These authors contributed equally to this work and should be considered co-second author.

Abstract. The prevalence of social media has risen dramatically, making it a crucial platform for understanding public health issues, including the expression of suicidal behavior. This study explores the feasibility of utilizing Natural Language Processing (NLP) methods to detect suicidal tendencies through Twitter posts. We employed various advanced NLP models, such as Logistic Regression (LR) and Bidirectional Encoder Representations from Transformers (BERT), to analyze the linguistic patterns and semantic nuances inherent in tweets. Our approach also included a Majority Vote system and Term Frequency-Inverse Document Frequency (TF-IDF) techniques to enhance the detection accuracy. The objective was to develop an effective model capable of early identification of potential suicide risks, which could be crucial for timely intervention and support. This research not only contributes to the field of digital mental health monitoring but also offers insights into the potential of machine learning in addressing critical societal issues. The findings suggest that while current NLP models show promise, there are complexities and ethical considerations in applying these technologies for sensitive topics like suicide detection. The study underscores the need for continuous refinement of these models and highlights the importance of integrating human judgment in the final decision-making process.

Keywords: Suicide Detection, Social Media Analysis, Natural Language Processing, Machine Learning Algorithms, Sentiment Analysis.

1. Introduction

Suicide is one of the leading causes of human death. About 1 million people worldwide commit suicide each year [1]. As the variety of suicide factors increased, according to Hedegaard's study, suicide rate is keep increasing these years. Social media is also a cause of suicide. Shah notes that suicide rates in the population are positively correlated with the penetration of Internet users [2]. This is because social

DOI: 10.54254/2755-2721/55/20241512

media users may experience cyberbullying and online harassment and have long been the target of attacks [3]. Research shows that the number of people using social media is still increasing to this day [4]. The main social media are mainly divided into Facebook, Weibo, Twitter, etc. [4]. Of those surveyed, 40 percent prefer to post comments online and 80 percent prefer to browse or search for comments on social media. According to a 2017 survey, 2.7 billion people worldwide use social media [4]. Social media has become a medium for diverse groups of people to express themselves [5]. Many people treat social media, such as Twitter as their personal diaries [6]. Thus, social media and people are inseparable, and withing suicide are also linked. Detecting such expressions is crucial not only for the well-being of individuals involved, but also for the wider online community. Early detection of suicidal ideation can lead to timely intervention measures that may save lives. However, manually monitoring a large amount of content on platforms such as Reddit is impractical. This highlights the need for automated systems that can accurately identify posts expressing or suggesting suicidal thoughts.

The latest developments in Natural Language Processing (NLP), especially the emergence of converter-based models such as BERT, LR and its variants, have completely changed our ability to understand and process human language. In this study, we utilized the power of these models to detect suicidal ideation in Reddit posts. By fine-tuning a carefully planned dataset, we aim to develop a model that can distinguish subtle clues in text, providing a promising solution to the challenge of automated suicide ideation detection.

2. Literature Review

Proposed by Sawhney and Mannchanda [7], Daine and colleagues carried out an investigation in which they examined all articles published between 1991 and 2011 in databases such as PsycINFO, MEDLINE, EMBASE, Scopus, and CINAH that focused on the linguistic aspects of self-harm or suicide. Their results indicate that the Internet can function as an intervention tool for individuals below the age of 25. Nevertheless, it is essential to note that not every language structure that includes the term "suicide" signifies a desire for self-harm. Specific semantic structures can also be employed to predict whether sentences imply a proclivity towards self-injury. Wang and Luo[8] introduced C-Attention models and hand-crafted features to detect a person's suicidal tendencies 30 days earlier. The results suggest that social media is useful for determining whether a user is contemplating suicide. Many people discuss their suicidal intentions and methods directly online[9]. Studies have shown that Twitter can help researchers predict a user's risk and probability of suicide[3]. By sifting through tweets and keywords related to the high risk of suicide from the Twitter site, the data and bias were calculated and compared with official national suicide rates. It found that 1.8% of the tweets were suicidal. Therefore, Twitter has been shown to be a way to detect suicide risk [3]. Sawhney and Joshi et al.[10] proposed a model called STATENet. They enhanced the language model to identify suicidal intentions in English tweets so that the model could be used to screen for suicides on social media. Du and Zhang et al.[11] refer to a previous study using a keyword approach to detect suicide risk in tweets. However, this method can detect a lot of irrelevant information. They built automatic binary classifiers to refine the data further using convolutional neural networks and tagging mental stressors. They also found that CNN has a significant advantage in this field and can extract mental stressors from social media. Hassib and Hossa et al.[6] created a AraDepSu dataset which consists of more than 20000 tweets, they separate the dataset into three parts, which are “depression”, “depression with suicidal”, and “non-depression”. Then they used 11 models, including mBERT, GigaBERT, AraBERT, and XLM-RoBERTa to do the model pre-trained on languages, tweets, and Modern Standard Arabic. The results show that Arabic people actually do show their negative emotions on social media by using depressive words and symptoms. Based on their predictive model, they classify and labeled the tweets as “depressed”, “suicidal”, and “neutral”. And their predictive accuracy can reach 91.20%. A similar experiment conducted by Deshpande, Rao et al.[12] also used sentiment analysis to test the tweets and images. They separate the tweets into “neutral” and “negative” by using a vector machine and Naive-Bayes classifier. In 2017, Birjali and his team construct a word set to solve the lack of terminological resources about suicide. Then they used Weka, which is a tool for better analysis based on machine learning algorithms. Roy and Nikolitch[13]

DOI: 10.54254/2755-2721/55/20241512

selected more than 500000 tweets on Twitter, training Machine Learning (ML) techniques to predict the risk of suicidal ideation and generate metrics in using sentiment polarity and neural networks. Then they used the matrix to train the random forest model to predict the suicidal ideation.

3. Data

The paper uses the Reddit Data set "SuicidalPost" which consists of 83,342 Reddit posts regarding various subjects. Out of these, 2,709 of those data are posts indicating suicidal ideation, which are labeled “Yes,” and 80,633 data are not suicidal posts, which are labeled “No” on suicidal column. Each sample contains a text of reddit post written in English and indication if the post was suicidal or not with “Yes” and “No” respectively.

Table 1. Reddit Dataset

	Title	Suicidal
1	This is the first thing I see when I open my eyes in the morning	No
2	I just don't know anymore	Yes
3	I can't even begin to explain the place where I am at.	Yes

The dataset is highly imbalanced with approximately 97% of the data labeled as "No" in the suicidal column. This imbalance in the dataset was deliberately chosen due to following reasons. The dataset contains extensive Reddit posts covering a wide range of topics. Furthermore, this imbalance in the data accurately mirror the social media platforms, where posts regarding suicide are rare and difficult to identify. Thus, a model capable of performing sufficiently on an imbalance dataset could provide promise for applications of real-world suicide intention.

4. Methodology

To understand how machine learning models detect suicidal posts and which model is most appropriate for the task, we trained and tested following three models with the dataset: logistic regression, random forest, and Bidirectional Encoder Representations from Transformers (BERT). Transformation algorithm of the term frequency inverse document frequency transformation algorithm was applied to logistic regression and random forest. "BertTokenizer" from "PreTrainedTokenizer" with padding and truncation was utilized for transformation and to formulate a consistent input length for BERT. Three models were trained and its outputs on the test data set were evaluated. Comparing outputs of three models for each data, the performance of three models combined were calculated through majority vote. Following the evaluation of the individual models, we employed a Majority Vote mechanism to further enhance decision-making. The Majority Vote method is a widely used ensemble technique where the final prediction is based on the majority output of all the models under consideration. The flowchart can refer to **figure 1 Model Training Progress**. It should be noted that the model training for LR, RF, and BERT is conducted separately, but the output results of all three models will be simultaneously concentrated in Majority Vote for final real-world applications.

DOI: 10.54254/2755-2721/55/20241512

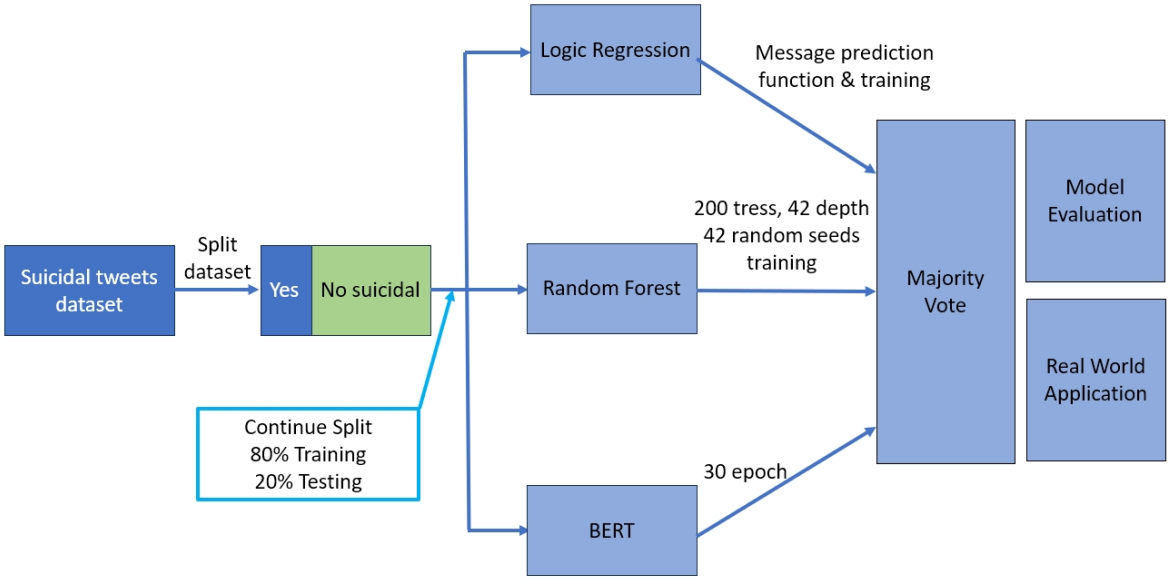

Figure 1. Model Training Progress

4.1. Data Loading and Preprocessing

Our main dataset comes from excel files tailored for this study, which contain social media-based messages and related tags indicating potential suicidal tendencies. Our dataset has nearly 90,000 samples, each was labeled with whether there is a suicidal tendency. To prevent the occurrence of sample blanks, we first removed all samples that were empty under "message" or "conclusion". We ensured effective data manipulation by loading the **Pandas** library. In order to partition the dataset, we allocated 80% of the data for training using the '**train_test_split**' function. The remaining 20% was divided equally for validation and testing, allowing for iterative model refinement and performance assessment. We allocated tweets with and without suicidal tendencies based on a relatively realistic ratio. The proportion of tweets with and without suicidal tendencies was almost the same in all sub datasets, to ensure that different methods did not have significant training bias after using the dataset.

4.2. Feature Extraction

We employed the term frequency inverse document frequency transformation algorithm (tf-idf) to convert text into numerical vectors, setting a limit of 50,000 terms to balance computational efficiency with capturing textual nuances. The method learns the vocabulary from the training data and returns the tf-idf weighted term-document matrix for the training data.

The tf-idf transformation algorithm is a numerical statistic that reflects how important a word is to a document in a collection or corpus. It's often used in information retrieval and text mining.

$$tf = \frac{number\ of\ times\ a\ term\ appears\ in\ document}{total\ number\ of\ terms\ in\ document}$$

$$idf = log\frac{total\ number\ of\ documents\ in\ corpus}{number\ of\ documents\ with\ terms\ in\ it}$$

$$tf\ idf = \frac{number\ of\ times\ a\ term\ appears\ in\ document}{total\ number\ of\ terms\ in\ document} \times log\frac{total\ number\ of\ documents\ in\ corpus}{number\ of\ documents\ with\ terms\ in\ it} = tf \times idf$$

DOI: 10.54254/2755-2721/55/20241512

4.3. Model Selection and Training

Three models, logistic regression (LR), random forest (RF), and Bidirectional Encoder Representations from Transformers (BERT), were trained on the dataset. The trained models were tested upon same datasets to compare which model performed most accurately on predicting suicidal posts.

4.3.1. Logistic Regression

We chose the LogisticRegression model as a baseline model for its simplicity. Logistic Regression (LR) is a widely used statistical method for binary classification. At its core, LR models the probability that a given input point belongs to a particular category. Logistic regression predicts the probability of an outcome in a binary scenario.

4.3.2. BERT

As a basis for feasibility comparison, we used a commonly used Natural Language Processing (NLP) training model: Bidirectional Encoder Representations from Transformers (BERT). BERT stands for "Bidirectional Encoder Representations from Transformers" and was introduced by Google in 2018 for NLP tasks. Unlike traditional models, BERT is deeply bidirectional, considering both left and right context in all layers. It's based on the Transformer architecture, which uses self-attention mechanisms for flexible representations. BERT is pre-trained using two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). After pre-training, BERT can be fine-tuned on specific tasks with smaller datasets. There are various BERT sizes, and models like 'RoBERT' have built upon BERT's foundation. BERT has been applied to numerous NLP tasks, from question answering to sentiment analysis. One criticism of BERT is its high computational requirement, especially when training from scratch. BERT's success highlights the power of transfer learning and large-scale pre-trained models in NLP.

4.3.3. Random Forest

For further comparison, Random Forest (RF), a commonly used machine learning algorithm which computes a single output based on the output of multiple decision trees, was selected. RF are composed of multiple decision trees. Decision trees are supervised learning algorithms that are often used for classification and regression operations. Decision trees subdivide and classify the dataset into multiple subsets based on outcomes of decision nodes. Being composed of multiple Decision trees, RF algorithms produce more accurate results and are less prone to bias and overfitting than a single decision tree.

4.3.4. Training Process

We initialize logistic regression classifiers and random forest t to train them on the "tf-idf" representations of the training data and their corresponding labels. We set 200 trees and each trees have 42 depths when we use random forest to train our model, and the random seeds is 42. BERT judge suicidal tendencies by predicting missing text, and we set up 30 layers to complete it. After training, we used the trained logistic regression model to predict the labels for the validation and test data. The "classification_report" function from Scikit-learn provides a detailed report on the classification performance, including metrics like precision, recall, and F1-score for each class.

4.3.5. Majority Vote

We use Majority Vote to produce the final output of the results of the three models, using the basic algorithm of Majority Vote: by collecting the results of each model on the same data point, and then classifying them, the classification with the highest results will be used as the final output. It does not perform any NLP related result processing on its own, but simply summarizes the results of the previous three models and selects the most likely result, which is the result with the highest number of votes, in order to increase the reliability and accuracy of the model.

DOI: 10.54254/2755-2721/55/20241512

5. Result

5.1. Model Performance

Table 2. Model Training Report

	Precision	Recall	F1-score	Support
tf - idf LR				
positive	**0.93**	**0.78**	**0.72**	**658**
negative	**0.97**	**0.99**	**0.98**	**8069**
tf-idf RF				
positive	**0.94**	**0.236**	**0.38**	**1341**
negative	**0.94**	**0.999**	**0.97**	**16113**
BERT				
positive	**0.79**	**0.60**	**0.68**	**84**
negative	**0.97**	**0.99**	**0.98**	**1116**
Majority Vote				
positive	**0.97**	**0.59**	**0.73**	**147**
negative	**0.97**	**1.00**	**0.98**	**1853**

The accuracy of LR model of classifying not potential suicidal post is 97%, but the accuracy drops to 93% for potential suicidal post. The macro average f1-score, which gives equal weight to the f1-scores of both classes, is 85%, The weighted average f1-score, which weights the f1-scores by the number of true instances for each label, is 96%. In summary, the model trained report using this method exhibits excellent performance in filtering out 'No Potential Suicide posts'. However, there is still a lot of improvement should make in determining whether a statement contains suicidal tendencies.

RF model portrays a similar pattern. The model accurately predicts non-suicidal post and potential suicidal post by 94%. Although the model presents high recall and F1-score on non-suicidal posts, the model exhibits significant low recall and F1-score, which only have 0.23 recall and 0.38 F1-Score. The result of random forest model with tf-idf transformation algorithm indicates the model have been overfitted to classifying non suicidal post and does not perform well on potentially suicidal post.

BERT accurately predicted 97% of non-suicidal post, yet 79% of potential suicidal post. The contrast between performance on potential and non-suicidal post expands on recall and F1-score. The contrast implies BERT is definitely overfitted with non-suicidal post and fails to accurately classify potential suicidal post with high precision.

The Majority Vote model presents a distinctive performance trend when analyzing its metrics. The model correctly identifies non-suicidal posts with remarkable precision and recall values, both standing at 0.97. For the positive class, which represents the non-suicidal posts, the model achieves a commendable precision of 0.97, coupled with a recall of 0.59, leading to an F1-score of 0.73. This is particularly impressive given the support count of 147 for the positive class. On the contrary, for the negative class, which pertains to potentially suicidal posts, the precision remains consistent at 0.97. However, the recall peaks at 1.00, resulting in a near-perfect F1-score of 0.98, supported by a massive count of 1853 instances. This suggests that the Majority Vote model excels at categorizing non-suicidal content but may be inclined to misclassify some of the potentially suicidal posts. Similar to the RF model, even though we have applied drop-out function during the training progress, the Majority Vote demonstrates a potential overfitting issue, particularly towards the non-suicidal posts, indicating areas that warrant further investigation and refinement. We can also found that in the final report, the performance of RF was very close to that of Majority Vote.

DOI: 10.54254/2755-2721/55/20241512

5.2. Model Comparison

When selecting the most appropriate model, its performance on positive conclusion, potential suicidal post, was most emphasized. In particular, recall of the model, which penalizes misclassifying an instance as negative, were heavily weighted since the model's performance decides matters of life and death. Based on following measure, we concluded that the logistic regression model best performed on detection of suicidal posts among the three models. Logistic regression model not only had highest precision on classifying potential suicidal post but also the highest recall.

6. Real World Application

We developed a function to facilitate real-time evaluations of messages. This utility processes a raw message and employs our trained model to predict its label, making it instrumental for immediate evaluations of user-generated content.

- message = "Test message here"
- predicted_label = predict_message(message, vectorizer, clf)
- print(predicted_label) #Result output

Sample Prediction: To demonstrate the model's applicability, we tested it on a sample message, showcasing how it can be used in real-world scenarios to assess content and provide immediate feedback. We collected over 100 messages from social media such as Twitter and Facebook in different partitions and created them into an information dataset.

Table 3. Manually collected dataset

	Message	**Conclusion (M)**
1	I can't help or stop how I feel.	Potential Suicide post
2	My own personal war.	Potential Suicide post
3	Loving the sun but is upset she can't make the picnic	No Potential Suicide post

This is a part of our manually collected dataset, which displays three typical types of tweets: significant suicidal tendencies, potential negative emotions, and descriptions of events.

Data input: This is the information we manually input for testing, with a total of 100 items, evenly distributed among them are descriptions of events, descriptions of direct suicidal tendencies, abstract descriptions of potential negative emotions, and descriptions of positive emotions.

Conclusion (M): This is our manual, expert labeled, judgment result used to compare the accuracy of model results. This result is not input into the model for proofreading before the model completes the label but is used as a comparison standard with the model output results.

After all training reports come out, we will compare all the comparable values to check which training method is more suitable in the application of suicidal tweets detection.

Table 4. Model output dataset(example)

	Message	**Conclusion**
1	I can't help or stop how I feel.	Potential Suicide post
2	My own personal war.	No Potential Suicide post
3	Loving the sun but is upset she can't make the picnic	No Potential Suicide post

We first manually jotted down whether each piece of information contained a label indicating suicidal tendencies. Then we applied this dataset to the logistic regression model we trained, and the result showed approximately 88% accuracy. Among them, tweets that cannot be accurately judged include: some obscure negative emotions may be misjudged, and descriptions of negative events may be ignored. For example, in **Table 2 Manually collected dataset**, the information section in the second row expresses a sense of loneliness and difficulty, and we can feel a sense of despair and helplessness when we hear such words. So manually judging this sentence carries a certain tendency towards suicide.

DOI: 10.54254/2755-2721/55/20241512

However, in the **Table 3 Model output dataset**, the model does not have the message in the second row as potential suicidal post. This is a typical problem that occurs in the output of NLP training models. The model may not be able to identify negative events or suicidal tendencies covered in abstract expressions.

7. Conclusion

In recent decades, natural language processing and artificial intelligence have developed to an extent to which machines can communicate and interact with humans. Its sophistication is reaching to a point where artificially formulated texts are indistinguishable or even better than what humans have written. Furthermore, it can not only generate messages and texts but also revise and evaluate them to significant extent. These developments must be noted with caution. On the other hand, its progressive development can be used to protect and save people who are overwhelmed and suffering in modern society. Considering the content and message of social media that increases exponentially every second, it is extremely overbearing and inefficient for humans to classify every social media post. Thus, applying natural language processing for detecting suicidal post and saving individuals on danger of suicide could provide innovation to the field of mental health and timely support.

References

[1] Orsolini, L., Latini, R., Pompili, M., Serafini, G., Volpe, U., Vellante, F., Fornaro, M., Valchera, A., Tomasetti, C., Fraticelli, S., Alessandrini, M., La Rovere, R., Trotta, S., Martinotti, G., Di Giannantonio, M., & De Berardis, D. (2020). Understanding the Complex of Suicide in Depression: From Research to Clinics. Psychiatry Investigation, 17(3), 207–221. https://doi.org/10.30773/pi.2019.0171

[2] Dhita Atrasina Ghaisani, 14312213. (2018). EFISIENSI KINERJA KEUANGAN PERUSAHAAN ASURANSI KONVENSIONAL DAN ASURANSI SYARIAH DENGAN PENDEKATAN DEA (Data Envelopment Analysis) Tahun 2014 dan 2015. https://dspace.uii.ac.id/handle/123456789/6770

[3] Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2014). Tracking Suicide Risk Factors Through Twitter in the US. Crisis, 35(1), 51–59. https://doi.org/10.1027/0227-5910/a000234

[4] Ghaisani, A. P., Handayani, P. W., & Munajat, Q. (2017). Users' Motivation in Sharing Information on Social Media. Procedia Computer Science, 124, 530–535. https://doi.org/10.1016/j.procs.2017.12.186

[5] Swain, D., Khandelwal, A., Joshi, C., Gawas, A., Roy, P., & Zad, V. (2021). A Suicide Prediction System Based on Twitter Tweets Using Sentiment Analysis and Machine Learning. In D. Swain, P. K. Pattnaik, & T. Athawale (Eds.), Machine Learning and Information Processing (pp. 45–58). Springer. https://doi.org/10.1007/978-981-33-4859-2_5

[6] Hassib, M., Hossam, N., Sameh, J., & Torki, M. (2022). AraDepSu: Detecting Depression and Suicidal Ideation in Arabic Tweets Using Transformers. Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), 302–311. https://doi.org/10.18653/v1/2022.wanlp-1.28

[7] Sawhney, R., Manchanda, P., Singh, R., & Aggarwal, S. (2018). A Computational Approach to Feature Extraction for Identification of Suicidal Ideation in Tweets. Proceedings of ACL 2018, Student Research Workshop, 91–98. https://doi.org/10.18653/v1/P18-3013

[8] Wang, N., Luo, F., Shivtare, Y., Badal, V. D., Subbalakshmi, K. P., Chandramouli, R., & Lee, E. (2021). Learning Models for Suicide Prediction from Social Media Posts (arXiv:2105.03315). arXiv. http://arxiv.org/abs/2105.03315

[9] Mbarek, A., Jamoussi, S., Charfi, A., & Ben Hamadou, A. (2019). Suicidal Profiles Detection in Twitter (296). https://doi.org/10.5220/0008167600002366

[10] Sawhney, R., Joshi, H., Gandhi, S., & Shah, R. R. (2020). A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. Proceedings of the 2020 Conference

DOI: 10.54254/2755-2721/55/20241512

on Empirical Methods in Natural Language Processing (EMNLP), 7685–7697. https://doi.org/10.18653/v1/2020.emnlp-main.619

[11] Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., & Xu, H. (2018). Extracting psychiatric stressors for suicide from social media using deep learning. BMC Medical Informatics and Decision Making, 18(2), 43. https://doi.org/10.1186/s12911-018-0632-8

[12] Deshpande, M., & Rao, V. (2017). Depression detection using emotion artificial intelligence. 2017 International Conference on Intelligent Sustainable Systems (ICISS), 858–862. https://doi.org/10.1109/ISS1.2017.8389299

[13] Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., & Kaminsky, Z. A. (2020). A machine learning approach predicts future risk to suicidal ideation from social media data. Npj Digital Medicine, 3(1), Article 1. https://doi.org/10.1038/s41746-020-0287-6