Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/95/2024CH0061

# Improvement for deep learning for suicide and depression identification with unsupervised label correction

**Jiaqi Liu[1,4,†], Yincheng Zhang[2,5,*,†], Weiyu Qian[3,6,†]**

[1]School of software, Shandong University, Shandong,250100, China
[2]Freeman School of Business, Tulane University, New Orleans,70118, The United States
[3]Department of Statistics, University of Wisconsin–Madison, Madison,53706, The United States

[4]202100300203@mail.sdu.edu.cn
[5]yzhang120@tulane.edu
[6]wqian26@wisc.edu
*corresponding author
[†]co-first authors

**Abstract.** Our study try to address the challenge of accurately identifying depression and suicidal ideation on social media platforms by introducing an enhanced novel methodology for unsupervised feature selection and label correction. Utilizing advanced word embedding models like BERT and the Universal Sentence Encoder, we transform textual content into dense numerical vectors that capture the nuanced emotional context of online discussions. Our approach enhances these embeddings with a deep neural network (DNN) to extract distinctive features, reducing the dimensionality of the data through Principal Component Analysis (PCA). For label correction, we employ clustering techniques including OPTICS, K-medoids, and hierarchical clustering, which are robust against noisy data points. We then train classifiers using CNN, DNN, logistic regression, and random forest algorithms, evaluated with metrics such as accuracy, precision, recall, F1 score, and AUC. This methodology improves the accuracy of classifying depressive and suicidal sentiments to some extent, assisting to utilize the vast data available on social media to advance mental health diagnostics and interventions.

**Keywords:** Mental Health Diagnostics, Deep Learning, Suicide/Depression Detection, Feature Selection, Unsupervised Label Correction.

## 1. Introduction

One of the most prevalent mental health disorders, depression, affects hundreds of millions of individuals from a global perspective. The World Health Organization reports that over 264 million people suffer from depression, and if not addressed on time or adequately, many cases progress to suicide [1]. Consequently, early detection and intervention are pivotal in providing timely support and potentially saving lives. However, traditional diagnostic methods, which rely heavily on Electronic Health Records (EHRs), clinical interviews, and self-reported questionnaires, face significant data availability and reliability limitations. This probably impede the process of detection and mitigation in this flexible and open community[2, 3].

Fortunately, in recent years, the rapid development of Internet community and social media platforms has created golden opportunities for mental health management. Platforms, such as Reddit, offer a precious environment where users can share their feelings anonymously. This kind of circumstance could facilitate a more honest and open disclosure about users' true conditions or concerns. Some researches have proved that this anonymity encourages users to share personal feelings that might be withheld in traditional settings. For instance, it was found that social media could offer "critical insights" into the mental health states of users regarding their posts and interactions [4]. Similarly, other experts viewed that online communities are crucial environments for those searching for mental health support [5]. This means that social media could offer a valuable data pool that may tackle with poor availability from conventional approaches.

However, utilization of online data for diagnostics of mental health condition could lead to noisy labels simultaneously. The introduced challenges of tackling noises have the potential risk of degrading the performance of our models. Traditional methods may struggle with these noisy datasets particularly [6]. Therefore, our study proposed a modified approach through a novel combination of unsupervised label correction and feature selection techniques. Unlike previous methods, our approach does not include prior knowledge of the noise distribution in the data as a prerequisite, which may improve the accuracy as well as increase its adaptability to various datasets.

Feature selection is crucial in initializing classification models in the prior stage. According to Li & Yang and Fan and Qin conventional techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF), have been extensively used in Natural Language Processing (NLP)[7, 8]. Nevertheless, these methods could fail to capture the complex context of human language, especially in daily emotional settings, which is the basis of comments on social media platforms here. By comparison, BERT (Bidirectional Encoder Representations from Transformers), representing the recent developments in embedding models, may provide more sophisticated feature extraction by offering context-aware representations of text data [9, 10].

Our work integrated embedding models with feature selections for label correction and clustering algorithms. Specifically, we utilized dimensionality reduction Principal Component Analysis (PCA) and clustering KMEANS to refine noisy labels from the web-scraped datasets.

In summary, this paper frameworks a relatively comprehensive strategy for enhancement in the detection of depression and suicidal ideation amid the circumstance of noisy social media platforms. Our approach of unsupervised label correction, combined with the embedding models and deep learning technologies, may represent an advance in accessible and daily mental health diagnostics.

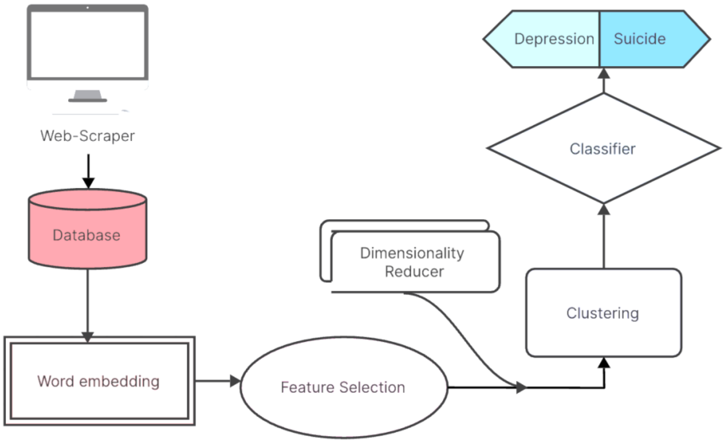

**Figure 1.** A flowchart for classifying suicide and depression by noise label correction through unsupervised learning

## 2. Methods

### *2.1. Words Embedding*

Word embedding techniques enhance the precision of our analysis at the first step. We mainly focused on models like BERT (Bidirectional et al. from Transformers) and the Universal Sentence Encoder (USE) to transform textual content into numerical vectors. This quantifying transformation step is essential, as it may capture the nuances amid complex language and emotions in discussions about mental health, which may be often missed by the commonly straightforward textual analysis tools.

BERT is known for its deep learning architecture that processes words with all the other words in a sentence rather than one-by-one. This comparatively intact contextual ability allows us to understand context more effectively. As the background is critical to understand meanings in mental health discussions, it is useful in detecting subtleties in language that indicate mental distress or suicidal signals. On the other hand, the Universal Sentence Encoder generates embeddings that are trained to naturally capture the meaning of entire sentences, making it highly effective for tasks in understanding sentence-level nuances in mental health topic.

In addition, we created a feature set that our models can analyze by converting the scraped Reddit posts into these embeddings. These features represent not only the semantic and syntactic nuances of the language used but also the emotional context. This indicates the features' essential role in accurately identifying signs of depression and suicidal thoughts respectively. This embedding method as a primer may improve our classifications' accuracy and potentially increase the effectiveness of interventions based on social media data in the further steps.

### *2.2. Feature Selection*

Based on our previous attempts, we found that directly clustering vectorized text information does not produce effective results. We aim to improve this by performing feature extraction on the raw text vectors. Initially, we vectorized the text using BERT to obtain a 756-dimensional feature vector for each text. However, this vector is highly generalized and does not effectively highlight the unique features of the text.

To do this, we introduce a new feature extraction layer. We create a simple deep neural network (DNN) model to train it, and use its feature extractor to extract features from input data. The feature extractor converts a 756-dimensional feature vector to a more manageable 128-dimensional vector. This refined feature vector can be used for subsequent clustering to enhance the clustering results by capturing unique features of the text.

### *2.3. Label Correction*

We still need to reduce the dimensionality after using trained feature selection algorithms to extract features. The team decided to run Principal Component Analysis (PCA) to reduce the high-dimensional features to 2-dimensional features because it is easier to visualize and cluster. 2-Dimensional features can also reduce the noise and allow the following algorithms to focus on the main components.

In the original paper, the authors used K-means and other clustering ways to analyze the features. We used Ordering Points to identify the Clustering Structure (OPTICS), K-medoids, and hierarchical clustering for individual analysis. We have several reasons for choosing these three methods. First, OPTICS and K-medoids can deal with noisy data points effectively, unlike the K-means method, which noise can affect. K-medoids can also be more robusted by choosing actual data points as centers. Second, OPTICS can select the number of clusters automatically, increasing accuracy. Third, K-medoids are less sensitive to initial selection than K-means. Then, we calculated the clustering performance metrics corresponding to Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI). After getting the results, we transferred the data into charts and compared them with the original paper's results. Finally, we added the clustering results to the original data and saved it as a new dataset for the following classification.

### *2.4. Classification*

In the previous section, we refined our text labels through clustering. Next, we will use the corrected text and labels as training data to train a classification model. The goal is to analyze the features of the text vectors to determine the tendency of the text, explicitly identifying whether the text contains suicidal tendencies. Ultimately, we aim to assess whether the user who made the statement might have suicidal tendencies.

We employed four algorithms—CNN, DNN, logistic regression, and random forest—to train four distinct classifiers. We then used test data to evaluate and compare the performance of these classifiers.

### *2.5. Datasets*

#### *2.5.1. Web-Scraped Depressed vs Suicidal Dataset*

Our primary dataset aims to support the classification of online posts on social media platforms as either indicative of depression or suicide. We extracted this data from Reddit utilizing the Python Reddit API, specifically targeting posts from two subreddits: SuicideWatch and Depression. This dataset includes 6,628 posts in total, utilizing the text of each post for input and the source subreddit as categorical labels—posts from SuicideWatch are categorized as 'suicidal' and those from r/Depression as 'depressed.' We have made the dataset and the scraping tools used to gather it publicly accessible to ensure our methodology can be independently verified and reproduced.

#### *2.5.2. Reddit C-SSRS Dataset*

Additionally, we utilize the specialized C-SSRS dataset to test and refine our label correction methods. This dataset comprises 500 entries from r/depression, each meticulously labeled by experts like clinical psychologists, according to the Columbia Suicide Severity Rating Scale. The labels, which are assigned based on a graduated five-point scale, reflect the severity of depressive symptoms portrayed in each post. This clinically validated dataset is crucial, as it provides a dependable standard against which we can gauge the effectiveness of our label correction approach, ensuring that our findings are robust and valid within the mental health topic.

## 3. Experimental Results

### *3.1. Implementation Details*

For the dataset we crawled, we utilized data from Reddit as a training set and utilized data from IMDB as a test set. The deep learning framework utilizes TensorFlow and uses Bert for text-to-vector conversion. In the training process, we also use the Adam optimizer to train the deep learning model and use the binary cross-entropy loss function. For classification accuracy, we use five metrics: accuracy (Acc), precision (Prec), recall (Rec), F1 score (F1), and area under the curve (AUC).

### *3.2. Label Correction Performance*

#### *3.2.1. Clustering Performance*

To evaluate the effect of our optimization, we used the KMEANS clustering method as an example. We compared the clustering performance on data without features extracted by the feature extraction layer and with features extracted by the feature extraction layer.

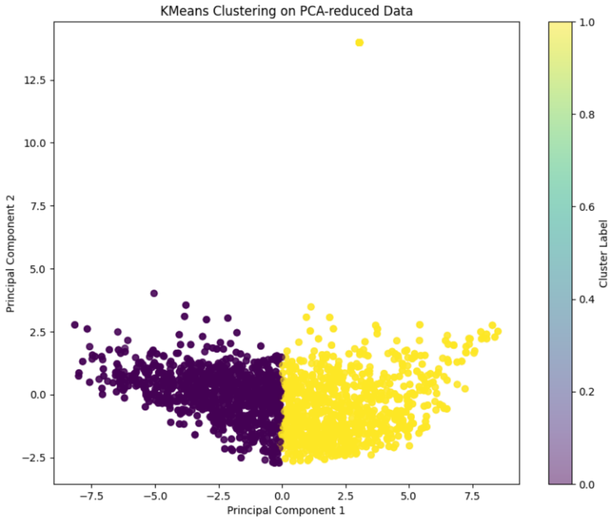

**Figure 2.** The clustering results of the data without feature extraction.

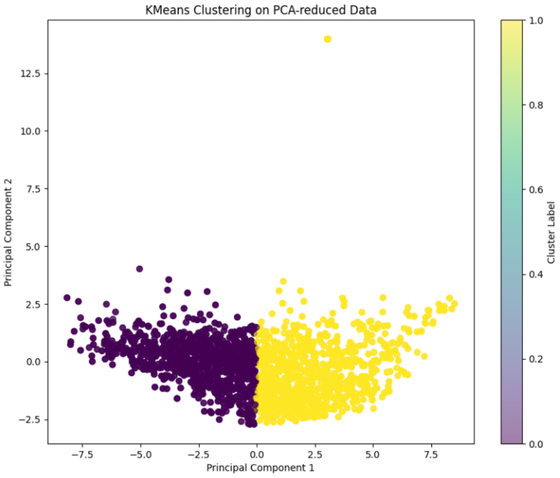

**Figure 3.** The clustering results after adding the feature extraction layer.

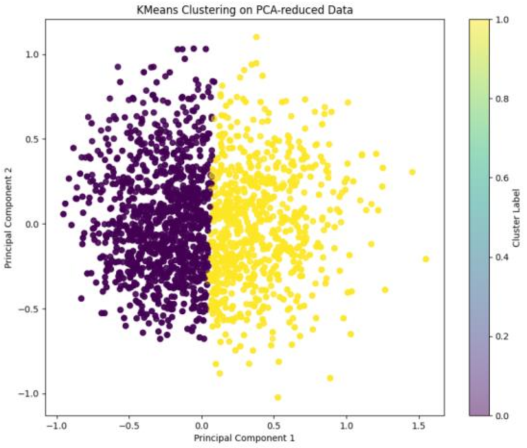

**Figure 4.** The clustering results after adjusting the parameters of the feature extraction layer.

In Figure2, the results are not very satisfactory, as most of the data are labeled as non-suicidal. It is evident that the clustering results significantly improve after incorporating the feature extraction layer in Figure4.

In Figure 2, the data distribution is more extensive, with a large vertical range. In contrast, Figure 3 shows a more compact data distribution, mainly concentrated in the central area of the horizontal axis. The clustering effect in Figure 2 is relatively weak, with more mixed points, whereas the clustering effect in Figure 3 is more substantial, with a more precise boundary between the two clusters.

### *3.2.2. Classification Performance after Label Correction*

We used four models to train the classifier for final evaluation: DNN, CNN, Logistic Regression, and Random Forest. We used the IMDB data for measurement and evaluated the final results based on metrics such as AUC.

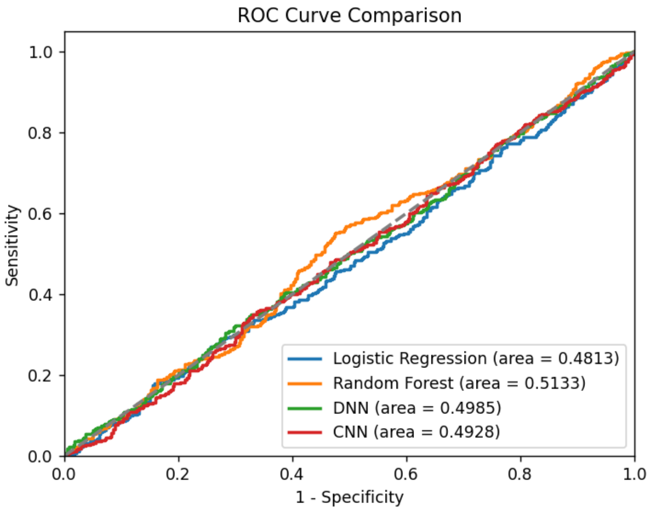

**Figure 5.** ROC Curve Comparison.

### *3.2.3. Final Evaluation*

The results show that feature extraction followed by clustering has achieved the goal of label correction, demonstrating practical effectiveness. We used Reddit data as the training set and IMDB data as the test set, training a binary classification model to observe the correction effect. It is clear that after applying feature selection, the final model's performance improved significantly, allowing for better judgment on the tendency of purple shirts.

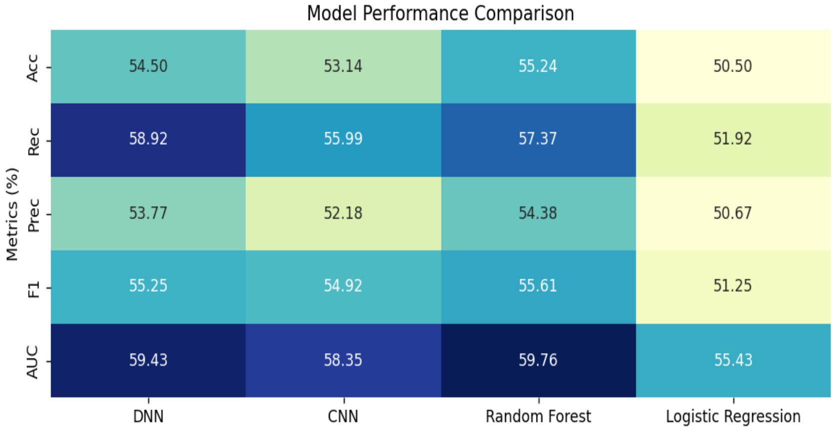

**Figure 6.** Model Performance Comparison.

## 4. Conclusion

### *4.1. Main Findings*

The team ran pre-research several times before the formal research and found that the clustering methods from the original papers actually could not handle the noisy points from real-world data. We consider that the reasons for the excellent performance of the real-world dataset in the original paper might be a coincidence and that the testing dataset is not mainly focused on mental health. Since there is still no testing dataset for mental health, we decided to refine the research method, adding feature extraction to control the noise. The results of the research show that our label correction is successful in the real-world dataset.

### *4.2. Significance of the research*

As the research result shows, our refined method can deal with the noise in the real-world dataset well, though its accuracy is not as good as we expected. In general, our research ways can be applied to databases that do not have standardized references and where users lack the corresponding subject knowledge to conduct supervised learning. In detail, this method can be used to analyze and classify people with suicidal tendencies or depressive tendencies.

### *4.3. Limitations*

The first limitation comes from the noisy points. Although we tried to use multiple methods to deal with the noisy points, there are too many in the primary dataset, which still influences the research result. The second limitation comes from the testing dataset. Since the mental health subject still lacks a standardized label dataset about suicide and depression, and the team lacks members with expertise in mental health studies, we are limited in combining unsupervised learning with supervised learning.

### *4.4. Recommendations for Future Research*

Due to the limitations, we recommend two directions for future research. The first direction is to control the noisy points in the raw datasets from the real world. The team tried several feature extraction methods to reduce the noises before clustering. There might be other steps that can be added to the method to deal with the noises. Another direction is to combine unsupervised learning with supervised learning. The original paper team and the team lack members with knowledge of mental health studies. We hope colleagues with knowledge in mental health studies can join in future research so that supervised learning can be applied better. We also hope the standardized label dataset about mental health will come out soon.

### *4.5. Conclusion*

The team ended up with methods that can successfully analyze and classify whether people have suicidal tendencies or depressive tendencies. Moreover, this would contribute to preventing suicide and depression diagnoses. The team hopes that more people will pay attention to the online mental health problem.

## References

[1] World Health Organization.(2021). Depression and Other Common Mental Disorders: Global Health Estimates.

[2] Vahia, I. V. (2013). Diagnostic and statistical manual of mental disorders 5: A quick glance.Indian Journal of Psychiatry, 55(3), 220. doi:10.4103/0019-5545.117131.

[3] Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications.Psychological Medicine, 49(9), 1426-1448. doi:10.1017/S0033291719000151.

[4] Saha, K., Torous, J., Ernala, S. K., Rizuto, C., Stafford, A., & De Choudhury, M. (2019). A computational study of mental health awareness campaigns on social media.Translational behavioral medicine, 9(6), 1197–1207.https://doi.org/10.1093/tbm/ibz028

[5] Chancellor, S., Nitzburg, G., Hu, A., Zampieri, F., & De Choudhury, M. (2019). Discovering alternative treatments for opioid use recovery using social media.Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-27. doi:10.1145/3359200.

[6] Zhang, Z., & Sabuncu, M. R. (2018). Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels.Advances in Neural Information Processing Systems, 31, 8778-8788.

[7] Li, Y., & Yang, T. (2018). Word embedding for understanding natural language: a survey.Guide to big data applications. Springer. doi:10.1007/978-3-319-53817-4_8.

[8] Fan, H., & Qin, Y. (2018). Research on Text Classification Based on Improved TF-IDF Algorithm.Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018). https://doi.org/10.2991/ncce-18.2018.79

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171-4186. doi:10.18653/v1/N19-1423.

[10] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal Sentence Encoder for English.Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 169-174. doi:10.18653/v1/D18-2029.