Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 245 (2024) 971–978

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Computer Science and Computational Intelligence 2024 (ICCSCI 2024)

# Mental illness detection using sentiment analysis in social media

Kasimirus Derryl Odja[a,*], Jasson Widiarta[a], Eko Setyo Purwanto[a], Muhamad Keenan Ario[a,b]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
[b]*Mobile Application & Technology Program, Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

**Abstract**

This research tries to detect mental illness using sentiment analysis on Reddit data, as well as comparing the performance of the k-Nearest Neighbors (k-NN), Random Forest, and Neural Network models. Using text post data from the pre-pandemic and post-pandemic periods, we concluded that the Random Forest model had the highest overall performance with an F1 Score, accuracy, recall and precision of 80.6%, making it quite effective in detecting depression. Even though the Neural Network model shows slightly lower accuracy, namely 79%, in fact this model has the lowest error rate, namely 0.06496. The k-NN model showed the lowest accuracy and higher error rate. These findings highlight the potential of sentiment analysis and machine learning in identifying mental health issues on social media and suggest that better models can improve early detection and intervention efforts.



*Keywords:* Sentiment Analysis; Mental Illness; k-NN; Random Forest; Neural Network.

## 1. Introduction

Mental illness is a type of illness that affects someone's state of mind and their wellbeing. There are a lot of mental health problem types with various methods to detect it using machine learning technique [1]. but from all those types, the major problems are depression and anxiety [2]. Also, it is important to know that mental illness could become a severe problem because it can cause other dangerous act such as self-harm or even suicide [3]. Considering that, it is

* Corresponding author.
*E-mail address:* kasimirus.odja@binus.ac.id


10.1016/j.procs.2024.10.325

important to raise awareness of mental health problems so we can prevent those dangerous acts or even the long-term effect. Since there are a lot of mental health problem types, this paper will focus on several types of mental health problem with their corresponding method.

To raise awareness of mental health problems, it is needed to detect those symptoms. AI is known to be capable of diagnosing this and describing the state of someone's mental health [4]. One of the best methods to diagnose mental illness using AI by using sentiment analysis. Using AI also means that we need to have a dataset for training and testing the model. Those data can be obtained from various sources like social media, blog post/diary, etc. Most of the people with mental health illness need a media to express their emotions without a threat of being know by the other in real life of their mental disorder through social media, and currently the social media that is known to have good quality and quantity of those data is twitter or also known as Reddit [5]. Then from those data we can train a model to determine whether someone have a mental illness or not.

The purpose of this research is to compare the methods of classifying the mental health category using sentiment analysis. The methods that will be used are based on the literature review conducted later. By comparing those methods, the effectiveness of mental health classification can be seen and can be customized based on our needs.

This paper organized as follows: Section 1 provides overview of usage of AI especially sentiment analysis in detecting mental health problems; Section 2 presents a brief of literature review regarding related work on sentiment analysis including the methods they use and how well they performed; Section 3 presents the proposed methodology for detecting mental illness using sentiment analysis including some steps like data gathering, pre-processing, feature extraction, training, and evaluation; Section 4 explains the result of the experimentation; and finally, Section 5 provides the conclusion of the research and ideas for future works.

## 2. Literature Review

The level of depression in social media can be detected using sentiment analysis with Natural Language Processing (NLP). Whether a sentiment analysis NLP model is good or not is determined by the data set and training methodology used. Data set search and text preprocessing are very crucial stages for providing text training of the best quality. Conventional or widely adopted techniques are used for this purpose. These techniques include tweet deduplication, to remove redundant events, there is also converting all characters to lowercase, cleaning tweet data from redundant characters or terms, tokenization to convert text into individual words, and filtering to isolate terms. -an important term that comes from token [6].

Other methods such as Naïve Bayes can be used and achieve an accuracy of 80.9% which is quite effective for sentiment analysis [6]. The hybrid model or combination of SpacySM and SVM also has superior performance. Using Spacy small core English language for feature extraction and data set training using linear SVM with a data split ratio of 80% and the remaining 20% for testing. Another alternative, such as TextBlob, functions as a powerful and easy-to-use sentiment analyser, leveraging the popular NLTK and Patterns packages.

In other studies, there are those that show variability in training methods, such as classification for feature selection and training with KNN in increasing efficiency in data analysis [7]. Additionally, there are studies that have highlighted the use of techniques such as TF-IDF and Word2Vec for data preprocessing and feature engineering. Deep Learning (DL) and Machine Learning (ML) models will be evaluated based on these features. For example, the use of the Extra Tree classifier combined with fabric techniques such as TF-IDF, which has proven successful in classifying depressive sentiment in tweets with an accuracy of 84.83% [8].

In other research, someone used the LSTM-CNN architectural model to detect depression and was able to achieve the highest accuracy of 97% which surpassed other models such as logistic regression, Naïve Bayes, Random Forest, and Decision Tree [9]. For multiclass classification, other studies use techniques such as TF-IDF combined with LIWC to extract features and classify emotions [10]. This kind of classification can also be carried out using Multinomial Naïve Bayes (MNB) and TF-IDF and with better performance [11].

Using a heavier model BERT with its variation claims to be great method for complex problem sentiment classification. This has been proven by research for multimodal and multilingual sentiment analysis of tweets in which they used 6 different models that consist of Multilingual-BERT (M-BERT), XLM-RoBERTa (XLM-R), XLM-RoBERTa-Sentiment-Multilingual (XLMR-SM), Vision Encoders, CLIP and DINOv2. Of all the models trained, XLM-RoBERTa Sentiment-Multilingual successfully achieved the overall best result for unimodal and multimodal

with combination of CLIP with highest F1-Score of 74.6 and 98.4 [12].

For sentiment analysis with easy and quick usage, VADER can be utilized to provide those sentiments. Research focused on twitter sentiment analysis states that VADER and NLTK are suitable for multi-classification system. However, there are some limitations at the presented paper, such as only small amount of data was used, general lexicon was used for categorizing specific data, and the resulted data were not trained. So, the effectiveness of using VADER and NLTK for sentiment analysis cannot be seen.

## 3. Methodologies

As the purpose defined in the first section of this paper, a methodology is required to achieve desired results. Steps that are carried out by this paper consist of data collection or data gathering, data preprocessing, model construction, model training dan model evolution. In this paper we are also going to compare each method of classification that will be specified in the model construction.

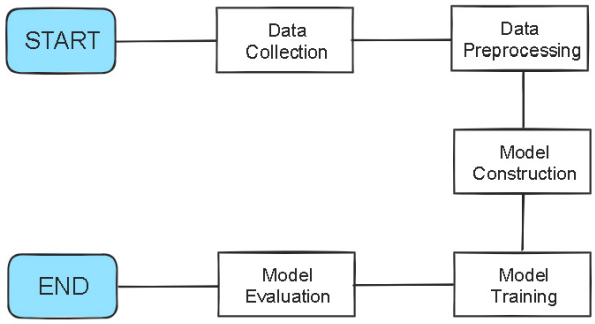


Fig. 1. Flowchart

### *3.1. Data Collection*

The data we collected is from Zenodo, authored by Low et al and published by JMIR Publication as a part of journal of medical Internet research [14]. The dataset gathered from social media platform called reddit which consist of several subreddits. The post was taken around January 2018 to April 2018, January 2019 to April 2019, December 2018 to Dec 2019, and January 2020 to April 2020. The dataset also comes with useful features like TF-IDF, LIWC and sentiment. In this paper, we will use the dataset combination from December 2018 to Dec 2019, and January 2020 to April 2020, which is the pre- and post-pandemic category. By using these combinations, it is possible to capture mental health problems and patterns from the effect of pandemics. The labels that will be used are addiction, anxiety, autism, depression, and schizophrenia. Each label consists of 1000 pre-pandemic rows and 1000 post-pandemic rows. The sum of all the labels will result in 10000 rows of data.

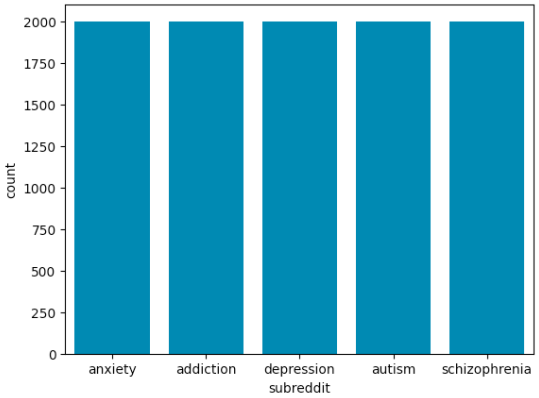


Fig. 2. Diagram

### *3.2. Data Preprocessing*

The dataset that we have collected and combined consists of 350 columns of data. The text from these datasets also contains a lot of special characters, similar words but one of them may be capitalized and there are also emojis which should be avoided for modelling because they usually cannot be coded during training and can cause inaccurate results. and less consistent. Therefore, these texts must be processed and go through a stage called pre-processing. In pre-processing, characters with capital letters will be reduced and made uniform, tokenized, and lemmatized. For example, you can see that the sentence "Feels like anxiety" was changed to "Feels", "like", and "anxiety", this is intended to increase understanding of the model.

### *3.3. Model Construction*

In our paper, we will use three different machine learning models for the classification task, each model can handle varying data complexities. The first model, the k-Nearest Neighbors (k-NN) algorithm, is famous for its non-parametric and instance-based learning approach. Trained on a cleaned and pre-processed dataset. Second, we will utilize the Random Forest algorithm, which is an optimal ensemble learning technique and is known for its robustness and reliability in solving classification problems.

The final model that we will try to develop is a neural network model. Using Keras, we combine dense layers with ReLU activation functions, dropout regularization, and batch normalization to improve performance and prevent overfitting. With Adam optimization and categorical cross-entropy loss function, it is hoped that the model can provide optimal classification performance. Through evaluating the training and testing data sets, as well as analysing the accuracy and loss plots, we will ensure the reliability and generalization of the model, so that it will also provide good classification results and strong conclusions.

### *3.4. Model Training*

In order to improve the performance of our models, the processed text data is directly fed into various learning architectures, namely k-NN, Random Forest, and Neural Network, for class classification. Training is carried out using 70% of the dataset. In the k-NN model, we will set n_neighbors=5. For the Neural Network model, we tried to use a Sequential model with five layers, each followed by a Batch Normalization and Dropout layer to reduce overfitting. The activation function used is ReLU, and L2 regularization with a regularization parameter of 0.01 which we try to apply to each layer. The model is compiled with the Adam optimizer and a categorical cross-entropy loss function. We also try early stopping to prevent overfitting during training. This neural network model will undergo training for 50 epochs with a batch size of 32, and we will monitor it through accuracy and loss plots. After training, the model

performance is evaluated on both the training and test datasets, providing insight into the classification accuracy of the three models.

### *3.5. Model Evaluation*

The last step we take is to evaluate the trained model. The metrics that will be evaluated are F1 score, accuracy, recall, precision, confusion matrix, and error. To see the error metric, we use is Mean Squared Error. The purpose of the evaluations that we carry out is none other than to find out whether the model we use has superior performance or not. The parameter is that if a model has a high accuracy value among other models, then that model will be considered the best model.

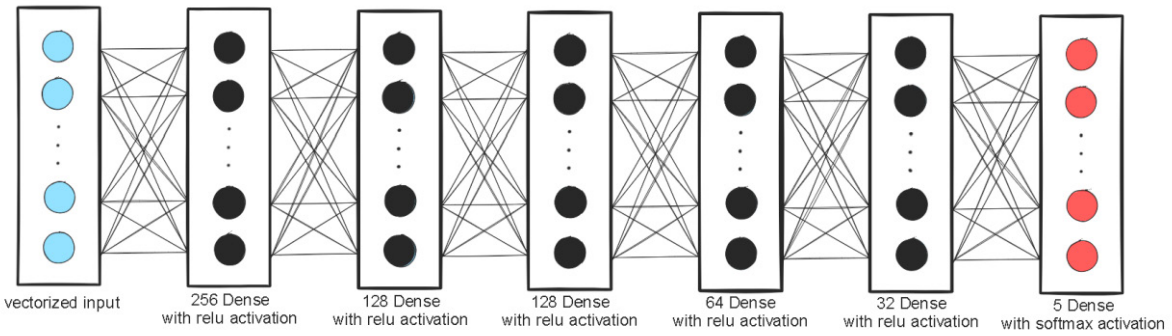


Fig. 3. Neural Network Architecture

## 4. Result & Discussions

In our experiments, we are using the mental illness Dataset from reddit. We carefully evaluated the accuracy of various models, as explained at the detail in the previous section. This section explores the comprehensive results of these experiments. Our research aims to compare methods for classifying mental health categories using sentiment analysis. The results of this experiment include metrics such as F1-Score, Accuracy, Recall, Precision and Error**.**

Our research explores the potential of classification models using social media data combined with learning technology to detect mental health categories. Findings from our research show that among the models tested, Random Forest demonstrated the highest overall performance, achieving accuracy, F1-Score, Recall and Precision**.** This highlights the model's ability to predict depression. Moreover, the results show that Random Forest can effectively detect depression in textual data. However, if we look at the complexity and results, the Neural Network model also shows results that are no less good. With higher complexity accompanied by regularization, the training accuracy reaches 76%. While Random Forest obtains high score for F1-score, accuracy, recall and precision. In terms of error, random forest might not get the lowest error which is 0.08479999 followed by KNN which has the error of 0.1678. So, the lowest error obtained by the neural network developed which is 0.064961255.

The result metrics of the experiment can be seen in the table below.

Table 1. Evaluation metrics of the experiment.

| ***Model*** | ***F1-Score*** | ***Accuracy*** | ***Recall*** | ***Precision*** |
|---|---|---|---|---|
| KNN | 0.678 | 0.678 | 0.678 | 0.680 |
| Random Forest | 0.806 | 0.806 | 0.806 | 0.806 |
| NN | 0.801 | 0.79 | 0.79 | 0.80 |

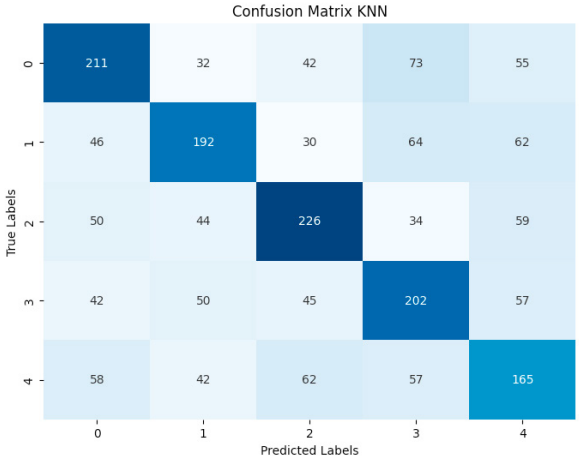


Fig. 4. K-Nearest Neighbourhood Confusion Matrix

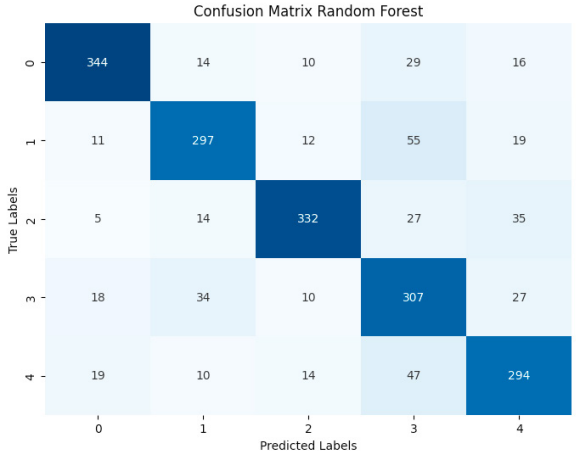


Fig. 5. Random Forest Confusion Matrix

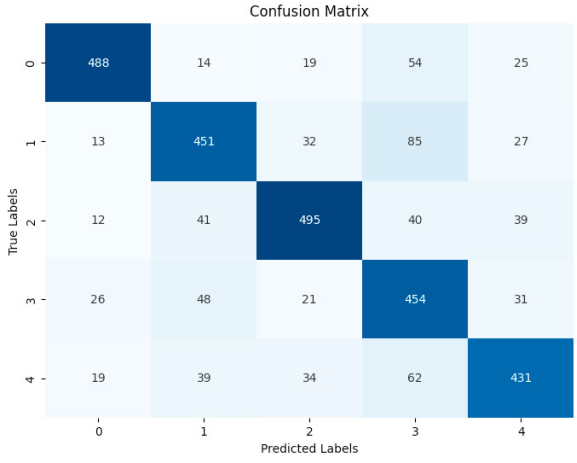


Fig. 6. Neural Network Confusion Matrix

The confusion matrix of KNN can be seen in Figure 4. Confusion matrix for Random Forest can be seen in Figure 5 and Figure 6 shows the confusion matrix of developed neural network.

## 5. Conclusion

This research aims to compare several classification models for several types of mental health using sentiment analysis on data from Reddit text posts. In this research we also used three machine learning models, namely k-Nearest Neighbours (k-NN), Random Forest, and Neural Network. These three models will be trained and evaluated using the Reddit data set mentioned previously.

From the results of this experiment, we conclude that the Random Forest model outperforms 2 other models in terms of F1-Score, accuracy, recall, and precision. Therefore, Random Forest is the most effective model in detecting mental health problems, especially depression. On the other hand, the performance of the random forest is quite good, even though it cannot reach the lowest error rate yet. The Neural Network model, although much more complex, it performs well and has the lowest error rate among the other 2 models.

For the k-NN model, unfortunately it shows the worst/lowest accuracy and has a higher error rate than the other 2 models. So, from all these facts, the use of sentiment analysis in combination with machine learning techniques can effectively identify mental health problems from social media data. The superior performance provided by the Random Forest model in predicting mental health problems shows its strength and reliability in real-world applications.

From our experiments and research, we hoped that future research will further consider and explore more sophisticated neural network architectures and hybrid models to further improve classification accuracy and reduce the error rates. In addition, newer and more diverse social media datasets can provide a better understanding of mental health trends and improve the model's predictive capabilities.

In conclusion, this research provides valuable knowledge regarding the use of sentiment analysis to detect mental health using machine learning models, especially Random Forest in classifying mental health problems from reddit post text. We also hoped that our findings would pave the way for the development of more sophisticated models to help detect and intervene early in mental health disorders.

## References

[1] Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: A scoping review of methods and applications. *Psychol Med* 2019;49:1426–48. https://doi.org/10.1017/S0033291719000151.

[2] Gao J, Zheng P, Jia Y, Chen H, Mao Y, Chen S, et al. Mental health problems and social media exposure during COVID-19 outbreak. *PLoS One* 2020;15. https://doi.org/10.1371/journal.pone.0231924.

[3] Hinduja S, Afrin M, Mistry S, Krishna A. Machine learning-based proactive social-sensor service for mental health monitoring using twitter data. *International Journal of Information Management Data Insights* 2022;2:100113. https://doi.org/10.1016/J.JJIMEI.2022.100113.

[4] Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Curr Psychiatry Rep* 2019;21. https://doi.org/10.1007/s11920-019-1094-0.

[5] Herdiansyah H, Roestam R, Kuhon R, Santoso AS. Their post tell the truth: Detecting social media users mental health issues with sentiment analysis. *Procedia Comput Sci*, vol. 216, 2022. https://doi.org/10.1016/j.procs.2022.12.185.

[6] Wongkar M, Angdresey A. Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter. 2019 *Fourth International Conference on Informatics and Computing (ICIC)*, IEEE; 2019, p. 1–5. https://doi.org/10.1109/ICIC47613.2019.8985884.

[7] Isnain AR, Supriyanto J, Kharisma MP. Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)* 2021;15:121. https://doi.org/10.22146/ijccs.65176.

[8] Muñoz S, Iglesias CA. A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations. *Inf Process Manag* 2022;59:103011. https://doi.org/10.1016/j.ipm.2022.103011.

[9] Sharma J, Tomer V. Depression detection using sentiment analysis of social media data. *AIP Conf Proc* 2022;2481:020044. https://doi.org/10.1063/5.0104192.

[10] Mustafa RU, Ashraf N, Ahmed FS, Ferzund J, Shahzad B, Gelbukh A. A Multiclass Depression Detection in Social Media Based on Sentiment Analysis, 2020, p. 659–62. https://doi.org/10.1007/978-3-030-43020-7_89.

[11] Abbas M, Ali K, Memon S, Jamali A, Memon S, Ahmed A. Multinomial Naive Bayes Classification Model for Sentiment Analysis. 2019. https://doi.org/10.13140/RG.2.2.30021.40169.

[12] Thakkar G, Hakimov S, Tadić M. M2SA: Multimodal and Multilingual Model for Sentiment Analysis of Tweets 2024.

[13] Elbagir S, Yang J. Analysis Using Natural Language Toolkit and VADER Sentiment, n.d.

[14] Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. *J Med Internet Res* 2020;22:e22635. https://doi.org/10.2196/22635.