

Evaluating CNN Performance for Sentiment Classification of Mental Health-Related Social Media Text: A Comparative Study of Small and Large Datasets

Dr. Angela Deepa V R¹, K. Shwetha², Sneha S³

¹Faculty, Westford Education, India

²Assistant Professor, Women's Christian College

³PG Student, Women's Christian College

DOI: <https://doi.org/10.51244/IJRSI.2026.1308000203>

Received: 26 August 2026; Accepted: 31 August 2026; Published: 15 September 2026

ABSTRACT

Social media provides a large amount of user-generated text in which people share their thoughts, feelings, experiences, and opinions. This text can be studied using natural language processing and deep learning techniques. This study evaluates the performance of a Convolutional Neural Network (CNN) for sentiment classification of mental-health-related social media text. A small dataset and a larger dataset are compared to examine the classification performance of the CNN across datasets of different sizes.

The small dataset contains 732 social media text entries, while the large dataset contains 27,972 entries. The text data were prepared using lowercasing, tokenization, stopword removal, stemming, and lemmatization. TF-IDF, GloVe, and Word2Vec were used for text representation as described for the respective datasets. CNN models were developed using Keras and evaluated using accuracy, precision, recall, and F1-score.

The CNN achieved an accuracy of 97% on the small dataset and 90.56% on the large dataset. The reported F1-scores were 49% and 90%, respectively. The difference between accuracy and F1-score in the small dataset shows the importance of considering class distribution when evaluating classification performance. The results indicate that CNN can classify sentiment patterns in mental-health-related social media text. However, sentiment classification cannot be considered a clinical diagnosis or direct evidence of schizophrenia. The study therefore focuses on sentiment classification and on comparing CNN performance across datasets of different sizes.

Keywords: Sentiment Classification, Mental Health, Social Media Text, Convolutional Neural Networks, Deep Learning, Natural Language Processing, Text Classification

INTRODUCTION

Mental health is an important area of research because mental health conditions can affect a person's thoughts, emotions, behaviour, and daily life. Social media has become an important platform for people to express their thoughts, feelings, experiences, and opinions. As a result, social media text provides an opportunity to study language and sentiment using natural language processing and machine learning techniques.

Social media text is different from formal written text. It may contain informal language, slang, abbreviations, spelling variations, emojis, sarcasm, and expressions that depend on the context. These features can make automated text analysis difficult. Sentiment analysis can be used to classify text according to its emotional orientation and can help in understanding patterns in user-generated content.

Previous studies have examined social media data in relation to mental health. Bae, Shim and Lee (2021) investigated the use of machine learning with social media content related to schizophrenia. Their work shows

how language patterns in social media can be examined using computational methods. However, research involving a particular mental health condition requires careful consideration of how the target condition is defined and how the labels are obtained.

Birnbaum et al. (2017) used a combination of machine learning and clinical appraisal to study social media markers of schizophrenia. Their study used Twitter timelines from users who had self-disclosed a schizophrenia diagnosis, and clinicians assessed the authenticity of these disclosures. The study shows the importance of clinical information when computational models are used to study a specific mental health condition.

Chancellor and De Choudhury (2020) reviewed methods used to predict mental health status from social media. They discussed issues involving data collection, annotation, preprocessing, feature selection, model selection, and validation. They also highlighted concerns about construct validity and the way mental health status is defined in computational studies.

The present study has a different focus. It examines sentiment classification of mental-health-related social media text. The sentiment labels available in the datasets are used as the classification targets. They are not treated as clinical labels for schizophrenia.

Deep learning methods have been widely used for text classification. CNNs can learn local patterns from textual sequences and have been applied to sentiment analysis. Ouyang et al. (2015) examined CNN-based sentiment analysis using word representations and demonstrated the use of CNN for text classification.

The present study compares CNN performance using two datasets of different sizes. The first dataset contains 732 social media text entries, while the second dataset contains 27,972 entries. The comparison is used to examine the classification performance of CNN under different dataset conditions.

The text is processed through cleaning, tokenization, stopword removal, stemming, lemmatization, and vectorization. TF-IDF and Word2Vec are used for text representation, while GloVe embeddings are used in the CNN approach described for the small dataset.

The model performance is evaluated using accuracy and F1-score. The results show a clear difference between accuracy and F1-score for the small dataset. This demonstrates the importance of considering more than accuracy when assessing classification performance.

It is also important to distinguish sentiment classification from clinical screening. Positive or negative sentiment alone cannot establish whether a person has schizophrenia. Therefore, the results in this study are interpreted only as sentiment classification results.

The main objective of this study is to evaluate CNN performance for sentiment classification of mental-health-related social media text and compare the results obtained from small and large datasets.

LITERATURE REVIEW

Several studies have examined the use of machine learning, deep learning, and natural language processing for analysing social media text related to mental health.

Bae, Shim and Lee (2021) studied schizophrenia detection using machine learning and social media content. Their study used social media information and examined language patterns associated with schizophrenia. The study reported high classification performance and demonstrated the potential of computational methods in this area. However, the data and labelling approach used in that study are different from the datasets used in the present research.

Lavanya and Angela Deepa V.R. developed an ensemble learning model for classifying tweets related to depression. The study used Random Forest and AdaBoost and also considered positive, negative, and neutral

sentiment categories. This work shows the use of machine learning and sentiment analysis for studying mental-health-related social media text.

Birnbaum et al. (2017) proposed a collaborative approach for identifying social media markers of schizophrenia by combining machine learning with clinical appraisal. Their study used Twitter data from users who had self-disclosed schizophrenia diagnoses. The disclosures were reviewed by clinicians before the data were used for classification. The study highlights the importance of combining computational methods with clinical information when the research objective involves a specific psychiatric condition.

Chekroud et al. (2021) examined the use of machine learning for predicting treatment outcomes in psychiatry. Their work discussed the growing use of predictive models in psychiatric research and emphasised the importance of evaluating models on new and unseen data.

Chancellor and De Choudhury (2020) reviewed predictive methods for mental health status using social media. Their review covered data annotation, data collection, preprocessing, feature selection, model selection, and validation. They identified concerns about construct validity and the way mental health status is operationalised in social media research.

Gkotsis et al. (2017) used informed deep learning to study mental health conditions in social media. Their study classified Reddit posts into different mental health themes and demonstrated the use of deep learning for mental-health-related text analysis.

Zomick, Levitan and Serper (2019) examined linguistic indicators of schizophrenia in Reddit posts. Their study used posts from users who reported having received a formal diagnosis of schizophrenia and compared their language with a control group.

Mitchell, Hollingshead and Coppersmith (2015) examined linguistic patterns associated with schizophrenia in social media. Their study focused on language differences in social media data and explored features that could distinguish schizophrenia-related users from control users.

Ouyang et al. (2015) investigated sentiment analysis using a CNN and word representations. Their work demonstrates the suitability of CNN-based methods for sentiment classification.

McManus et al. (2015) used Twitter data and machine learning methods to distinguish users who self-identified as having schizophrenia from control users. Their study used features such as emoticon use, posting time, and dictionary-based information. The SVM model achieved 92% precision and 71% recall on the held-out test data.

Srinivas et al. (2021) examined sentiment analysis using neural networks and LSTM. Their study used a large Twitter dataset and compared different approaches for sentiment classification.

The studies reviewed above show that machine learning and deep learning can be applied to social media text. They also show that the research objective and the quality of the labels are important. The present study therefore focuses on sentiment classification rather than clinical identification of schizophrenia.

Proposed System

Dataset

Two datasets are used to evaluate CNN performance. The first dataset is a small Kaggle dataset containing 732 social media text entries. It is used for sentiment analysis. The second dataset, `mental_health.csv`, contains 27,972 entries with the main fields `text` and `label`. The dataset includes mental-health-related topics such as anxiety and depression. The labels are described as `toxic` and `non-toxic`.

The two datasets differ in size and characteristics. They are used to examine how the CNN performs when working with a smaller and a larger collection of textual observations. The labels provided with the datasets are

used for the classification task. They are not treated as clinically validated schizophrenia labels. This distinction is important because the presence of positive or negative sentiment does not establish the presence or absence of a psychiatric disorder.

METHODOLOGY

The methodology includes data preprocessing, feature extraction, text vectorization, CNN model development, training, and evaluation. The general procedure is applied to both datasets. The main difference is the representation of the text and the CNN configuration used for each dataset.

Data Preprocessing

For the small dataset, the text is converted to lowercase, emojis are removed, and the text is tokenized. Stopwords are removed, and stemming and lemmatization are applied. Similar preprocessing is applied to the large dataset. Unnecessary symbols and stopwords are removed, the text is converted to lowercase, and the text is tokenized. Stemming or lemmatization is used to obtain a consistent form of words. These steps help prepare the text for feature extraction and classification.

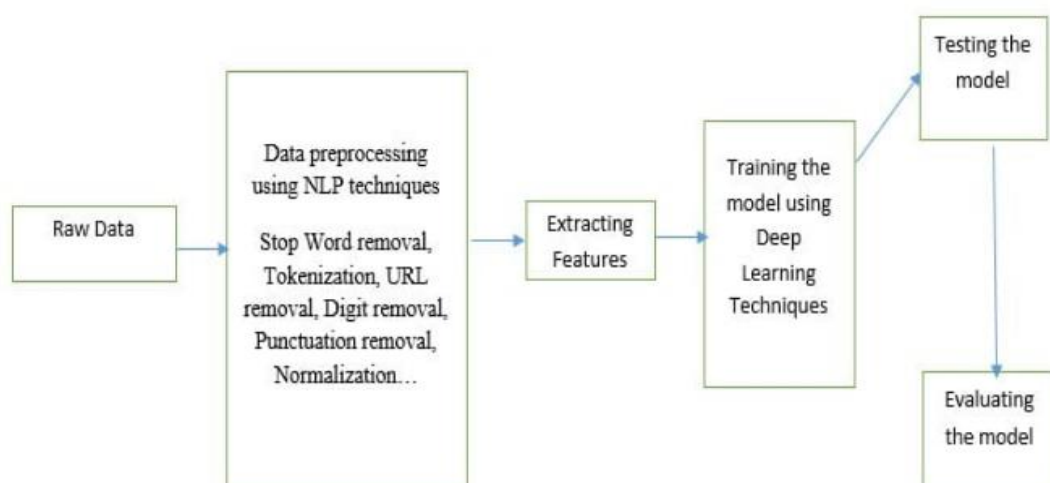


Figure 1: Methodology Diagram

Visualization of Data

Word clouds are used to provide a visual representation of commonly occurring words. For the small dataset, the word cloud contains terms related to topics such as hallucination, delusion, and negative symptoms. For the large dataset, the word cloud includes terms related to depression, anxiety, hallucination, and delusion. The word clouds are used to provide a general view of the textual content. Individual words are not treated as indicators of a clinical condition.

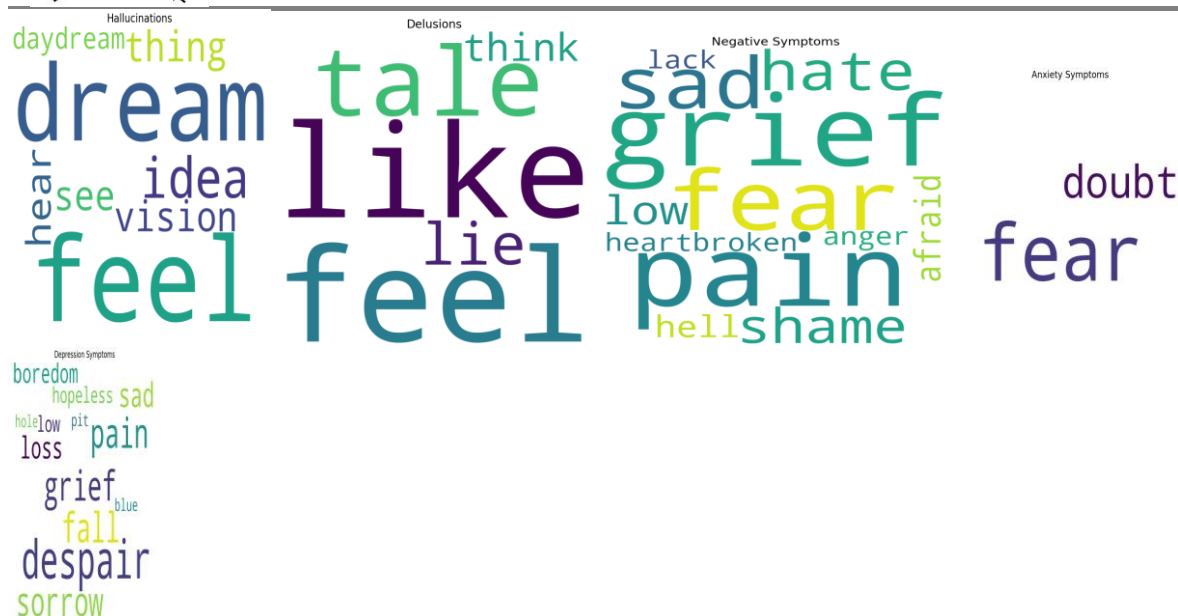


Figure 2: Word Cloud Visualization – Small Dataset

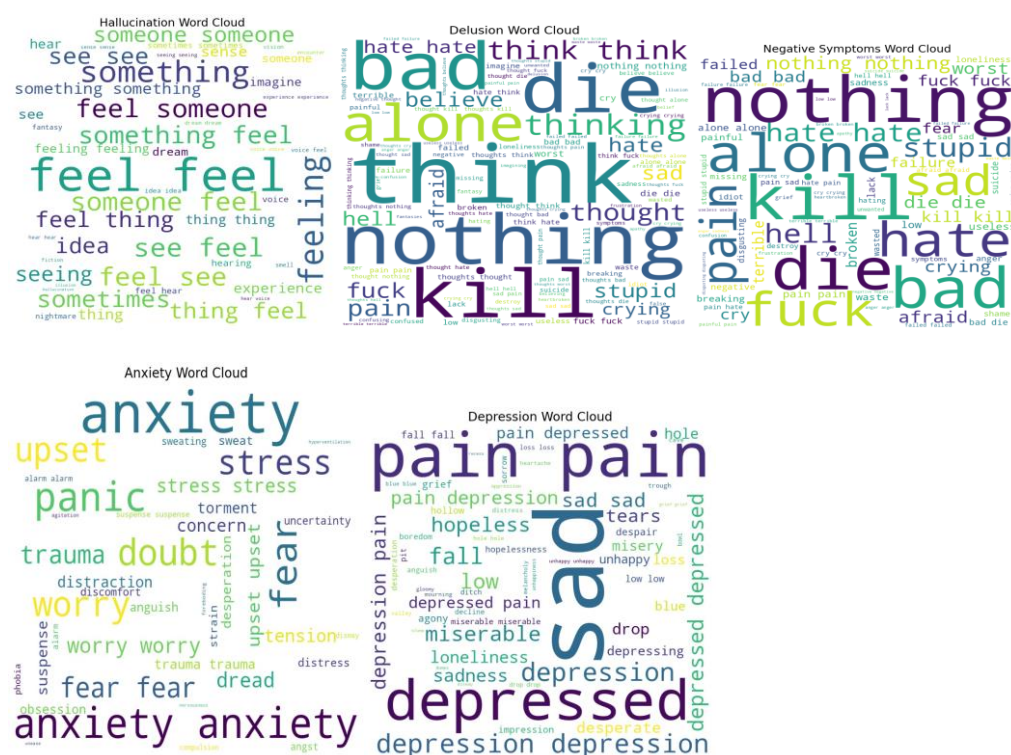


Figure 3: Word Cloud Visualization – Large Dataset

Feature Extraction and Selection

TF-IDF is used to convert text into numerical information. It gives greater importance to words that are useful for distinguishing text within the dataset. For the small dataset, TF-IDF is used for text representation. Feature selection methods such as chi-square or mutual information are considered for identifying informative features and reducing dimensionality. For the large dataset, TF-IDF is used together with Word2Vec. Word2Vec represents words as dense vectors and provides information about relationships between words. Feature extraction makes it possible to process textual information using the CNN model.

Data Vectorization

For the small dataset, TF-IDF is used to represent the text as numerical feature vectors. For the large dataset, TF-IDF and Word2Vec are used. TF-IDF represents word importance based on frequency, while Word2Vec provides word representations based on relationships between words. The resulting numerical representations are used as input for classification.

Model Building

A CNN model is developed using Keras for sentiment classification. For the small dataset, the CNN includes an embedding layer using GloVe word embeddings, followed by a convolutional layer, global max-pooling layer, and dense classification layer. For the large dataset, the CNN is also developed using Keras and includes embedding, convolutional, max-pooling, and dense layers. Binary cross-entropy is used as the loss function and Adam is used as the optimizer. CNN is used because it can learn local patterns in text and has been applied to sentiment classification tasks. Ouyang et al. (2015) also demonstrated the use of CNN for sentiment analysis.

Model Selection

The purpose of the study is to evaluate CNN performance across datasets of different sizes. Therefore, the experimental design focuses on CNN rather than on a comparison of several different architectures. Logistic Regression, SVM, Random Forest, LSTM, and transformer-based models are not included in the reported experiments. These models would provide useful comparisons, but adding numerical results without carrying out the corresponding experiments would not provide a valid comparison. They are therefore identified as suitable directions for future research.

Model Training and Evaluation

The CNN models are trained using the prepared text data. Validation data are used during training, and test data are used to evaluate the trained models. The model parameters are adjusted during training to reduce the loss. Binary cross-entropy is used as the loss function. Accuracy, precision, recall, and F1-score are used to evaluate classification performance.

Algorithm Description

The main steps are:

1. Load the datasets.
2. Prepare the text data.
3. Convert text to lowercase.
4. Tokenize the text.
5. Remove stopwords and unnecessary symbols.
6. Apply stemming or lemmatization.
7. Represent the text using TF-IDF, GloVe, or Word2Vec.
8. Build the CNN using Keras.
9. Train the CNN.
10. Evaluate the trained model.
11. Calculate accuracy, precision, recall, and F1-score.
12. Compare the results for the two datasets.

Evaluation Metrics

The models are evaluated using accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correctly classified observations. Precision measures the proportion of correctly predicted positive observations among the observations predicted as positive. Recall measures the proportion of correctly identified positive observations among all actual positive observations. F1-score combines precision and recall. Using several

measures is important because accuracy alone may not provide a complete view of classification performance when the classes are unevenly distributed. ROC-AUC, confusion matrices, and confidence intervals are not reported because the required prediction-level results are not available for calculating these measures reliably. They can be included in future experiments when the model predictions and complete evaluation outputs are retained.

RESULTS AND DISCUSSION

Sentiment Distribution

The two datasets contain different numbers of positive and negative sentiment observations.

Table 1. Sentiment Classification

Sentiment	Small Dataset	Large Dataset
Positive Sentiment	144	12,261
Negative Sentiment	588	15,716
Total	732	27,977

The small dataset contains 144 positive and 588 negative observations. The large dataset contains 12,261 positive and 15,716 negative observations. The small dataset therefore has a stronger difference between the two sentiment categories. This imbalance needs to be considered when interpreting the model results. The labels in this table refer to sentiment categories. They should not be interpreted as schizophrenia and non-schizophrenia groups.

CNN Accuracy

The CNN achieved an accuracy of 97% on the small dataset and 90.56% on the large dataset.

Table 2. CNN Model Accuracy

Model	Accuracy – Small Dataset	Accuracy – Large Dataset
CNN	97%	90.56%

The CNN achieved higher accuracy on the small dataset. However, accuracy alone does not provide enough information about the performance of each sentiment class. The large dataset produced a lower accuracy, but its F1-score was much higher than that of the small dataset.

F1-Score

Table 3. CNN F1-Score

Model	Dataset	F1-Score
CNN	Small Dataset	49%
CNN	Large Dataset	90%

CNN obtained an F1-score of 49% for the small dataset and 90% for the large dataset. The difference between the accuracy of 97% and F1-score of 49% for the small dataset is important. The small dataset contains considerably more negative than positive observations. As a result, a high accuracy value does not necessarily mean that the model performs equally well across the two classes. For the large dataset, the accuracy was 90.56%

and the F1-score was 90%. The two measures are much closer. This shows why classification performance should be examined using more than one metric.

Discussion of the Results

The CNN was able to classify sentiment in both datasets. The small dataset produced a higher accuracy but a lower F1-score. The large dataset produced a slightly lower accuracy but a much higher F1-score. The difference can be considered together with the class distribution. The small dataset has a larger difference between the number of positive and negative observations. This may affect accuracy and explain why accuracy and F1-score provide different views of performance. The word cloud visualisations provide additional information about the textual content. The small dataset contains words associated with feelings and mental-health-related topics, while the large dataset contains words related to anxiety, depression, hallucination, and delusion. These observations are useful for understanding the text, but they should not be used to make a clinical judgement about an individual.

Relationship to Previous Studies

Previous studies have examined social media text using different types of labels and different machine learning approaches. Bae et al. (2021) used machine learning to study schizophrenia-related social media content. Birnbaum et al. (2017) combined computational analysis with clinical appraisal. Zomick et al. (2019) examined linguistic characteristics in Reddit posts, while McManus et al. (2015) used an SVM approach with Twitter data. These studies differ from the present study because they used schizophrenia-related cohorts or self-reported information. The present study uses sentiment classification labels. Other research has focused specifically on sentiment classification. Ouyang et al. (2015) used CNN for sentiment analysis, while Srinivas et al. (2021) examined neural network and LSTM approaches for Twitter sentiment classification. The present study is closer to this second group because its main task is sentiment classification. A direct comparison with Logistic Regression, SVM, Random Forest, LSTM, and transformer models is not made because these models were not evaluated using the datasets in this study. Such comparisons would be useful in future research.

Ethical And Privacy Considerations

The datasets used in the study were obtained from publicly available Kaggle sources. The study uses existing datasets rather than directly recruiting or interacting with social media users. Although publicly available data can be used for research, mental-health-related social media information needs to be handled carefully. Text posted online may contain personal or sensitive information. Researchers should therefore avoid identifying individual users or making unsupported claims about their mental health. The sentiment labels in this study are used only for classification. They are not treated as clinical diagnoses. Research involving clinically relevant schizophrenia data would require clear information about the source of the clinical labels, data collection procedures, privacy protection, demographic information where appropriate, and relevant ethical approval or exemption.

This approach is consistent with concerns raised in the literature about construct validity, data annotation, and responsible use of social media data in mental health research (Chancellor and De Choudhury, 2020).

Limitations

The study has several limitations.

First, the datasets are obtained from Kaggle and contain text classification labels rather than clinically validated schizophrenia labels. Therefore, the findings cannot be used to identify or diagnose schizophrenia. Second, the two datasets differ in size and in their label definitions. The comparison therefore focuses on CNN classification performance rather than on a controlled clinical comparison. Third, the small dataset has an uneven distribution between positive and negative sentiment. This needs to be considered when interpreting the difference between accuracy and F1-score. Fourth, the study focuses on CNN. Logistic Regression, SVM, Random Forest, LSTM,

and transformer-based models were not evaluated. Fifth, the available model description does not provide all numerical hyperparameters, such as the exact number of filters, kernel sizes, batch size, number of epochs, learning rate, and random seed. Finally, ROC-AUC, confusion matrices, and confidence intervals are not reported because the necessary prediction-level outputs are not available for calculating them reliably.

These limitations provide useful directions for further work.

Future Work

Future studies can improve the research in several ways. First, datasets with clinically relevant and appropriately validated labels can be used when the research objective is to study schizophrenia-related classification. Second, future studies can provide clear details about data collection, annotation, demographic characteristics where appropriate, ethics, and privacy. Third, CNN can be compared with Logistic Regression, SVM, Random Forest, LSTM, and transformer-based models using the same datasets and evaluation procedure. Fourth, future experiments can report the complete CNN architecture and all hyperparameters. These can include the number of filters, kernel size, embedding settings, learning rate, batch size, number of epochs, dropout, and random seed. Future studies can also include confusion matrices, ROC-AUC where suitable, confidence intervals, and class-specific precision, recall, and F1-score. These additions would provide a stronger evaluation and make the experimental process easier to reproduce.

CONCLUSION

This study evaluated CNN performance for sentiment classification of mental-health-related social media text using two datasets of different sizes. The small dataset contained 732 observations, while the large dataset contained 27,972 observations. The text was processed using lowercasing, tokenization, stopword removal, stemming, and lemmatization. TF-IDF, GloVe, and Word2Vec were used for text representation as described for the respective datasets. The CNN achieved 97% accuracy on the small dataset and 90.56% accuracy on the large dataset. The F1-scores were 49% and 90%, respectively. The difference between accuracy and F1-score for the small dataset shows the importance of considering class distribution and using several evaluation measures. The large dataset produced a closer relationship between accuracy and F1-score. The findings indicate that CNN can be used to classify sentiment patterns in mental-health-related social media text. However, the results do not provide evidence of schizophrenia or any other clinical diagnosis. Positive or negative sentiment should not be treated as an indicator of schizophrenia. The study therefore focuses on computational sentiment classification and on understanding CNN performance across datasets of different sizes. Future research can use clinically validated datasets where appropriate and can include detailed annotation procedures, ethical and privacy safeguards, complete model specifications, baseline algorithms, and additional evaluation measures. These improvements would provide a stronger basis for computational research involving mental-health-related social media text.

REFERENCES

1. Bae, Y.J., Shim, M. and Lee, W.H. (2021). Schizophrenia detection using machine learning approach from social media content. *Sensors*, 21(17), 5924. DOI: 10.3390/s21175924.
2. Lavanya, R. and Angela Deepa, V.R. (2020–2021). An ensemble learning model to classify tweets based on depression. *Journal of Education: Rabindra Bharati University*, XXIII(6-II). DOI: Not identified.
3. Birnbaum, M.L., Ernala, S.K., Rizvi, A.F., De Choudhury, M. and Kane, J.M. (2017). A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of Medical Internet Research*, 19(8), e289. DOI: 10.2196/jmir.7956.
4. Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D. and Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, 20(2), 154–170. DOI: 10.1002/wps.20882.
5. Chancellor, S. and De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3, 43. DOI: 10.1038/s41746-020-0233-7.

6. Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T.J.P., Dobson, R.J.B. and Dutta, R. (2017). Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports*, 7, 45141. DOI: 10.1038/srep45141.
7. Zomick, J., Levitan, S.I. and Serper, M. (2019). Linguistic analysis of schizophrenia in Reddit posts. *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp. 74–83. Association for Computational Linguistics. DOI: 10.18653/v1/W19-3009.
8. Mitchell, M., Hollingshead, K. and Coppersmith, G. (2015). Quantifying the language of schizophrenia in social media. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 11–20. Association for Computational Linguistics. DOI: 10.3115/v1/W15-1202.
9. Ouyang, X., Zhou, P., Li, C.H. and Liu, L. (2015). Sentiment analysis using convolutional neural network. *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pp. 2359–2364. IEEE. DOI: 10.1109/CIT/IUCC/DASC/PICOM.2015.349.
10. McManus, K., Mallory, E.K., Goldfeder, R.L., Haynes, W.A. and Tatum, J.D. (2015). Mining Twitter data to improve detection of schizophrenia. *AMIA Joint Summits on Translational Science Proceedings*, 2015, 122–126. DOI: Not identified.
11. Srinivas, A.C.M.V., Satyanarayana, C., Divakar, C. and Sirisha, K.P. (2021). Sentiment analysis using neural network and LSTM. *IOP Conference Series: Materials Science and Engineering*, 1074(1), 012007. DOI: 10.1088/1757-899X/1074/1/012007.

AI Use Statement: Generative AI tools were used during the preparation of this manuscript for language editing, grammar improvement, and refinement of academic wording.