Text Mining for Classifying Potentially Depressive Tweets on X Using IndoBERT

Salsabila Safirana Wibisono(1), Masy Ari Ulinuha(2), Siti Nur'aini(3)

[1,2,3]Department of Information Technology, Universitas Islam Negeri Walisongo

Jl. Prof. Hamka, Ngaliyan, Semarang, 50185, Central Java

e-mail: salsaswibisono@gmail.com(1), ulinuha@walisongo.ac.id(2), siti_nuraini@walisongo.ac.id(3)

ABSTRAK

Depresi merupakan masalah serius di Indonesia, di mana penderitanya sering kali enggan mencari bantuan profesional dan lebih memilih mengekspresikan diri melalui media sosial seperti X. Studi ini menggunakan *text mining* untuk mengklasifikasikan potensi depresi dari 5.000 *tweet* berbahasa Indonesia sejak Oktober 2024 hingga Januari 2025. Tahapan *preprocessing* meliputi *case folding, cleaning, normalization, dan stopword removal*. Pelabelan dataset menghasilkan dua kelas: berpotensi depresi dan normal, kemudian dibagi menjadi 80% data latih dan 20% data uji. Model *pre-trained* IndoBERT disesuaikan dengan parameter *learning rate* 2e-05, *batch size* 8, dan *epoch* 2 untuk tugas klasifikasi potensi depresi ini. Hasil evaluasi menunjukkan bahwa model IndoBERT memberikan performa yang baik dengan akurasi sebesar 87%, *precision* 87%, *recall* 87%, dan *f1 score* 87%. Namun, model mengalami ketimpangan data (*class imbalance*), sehingga cenderung lebih baik dalam memprediksi label mayoritas (normal) daripada minoritas (depresi). Oleh karena itu, disarankan melakukan rebalancing guna mencegah hal serupa. Model IndoBERT dalam penelitian ini diinisialisasi dari model klasifikasi emosi, pelabelan dataset dilakukan secara manual oleh peneliti didampingi psikiater untuk memastikan relevansi klinis, dan hasil akhir model diterapkan menjadi aplikasi berbasis *web* dengan Streamlit. Aplikasi ini dibuat sebagai alat skrining awal untuk membantu psikiater, bukan sebagai sistem diagnostik.

Kata kunci: Depresi; IndoBERT; Klasifikasi; Media Sosial X; Text Mining

ABSTRACT

Depression is a severe issue in Indonesia, where sufferers often do not seek professional help and prefer to express themselves through social media such as the X. This study uses a text mining approach to classify potential depression using a dataset of 5,000 Indonesian-language tweets from October 2024 to January 2025. The preprocessing steps involves case folding, cleaning, normalization, and stopword removal. The dataset was labeled into two classes: potentially depressive and normal, then divided into 80% training data and 20% test data. A pre-trained IndoBERT model was adjusted with a learning rate of 2e-05, batch size of 8, and epoch of 2 for this depression potential classification task. The evaluation results showed that the IndoBERT model performed well with an accuracy of 87%, precision of 87%, recall of 87%, and f1 score of 87%. However, the model's performance affected by class imbalance, so it tended to be better at predicting the majority label (normal) than the minority label (depression). Therefore, rebalancing is recommended to prevent similar occurrences. The IndoBERT model used in this study was initialized from an emotion classification model, manual labelling was conducted by researchers in collaboratoin with psychiatrists to ensure clinical relevance. Finally, the trained model was deployed into a web-based application using Streamlit. This application was created as a preliminary screening tool to assist psychiatrists, not as a diagnostic system.

Keywords: *Depression; IndoBERT; Classification; X Social Media; Text Mining*

INTRODUCTION

Depression is one of the most common mental disorders, in Indonesia itself, more than 12 million Indonesians aged 15 and older suffer from depression in 2018 [1]. Depression can arise from environmental, psychosocial, or cognitive factors that cause changes in physical, feelings, thoughts, and daily habits [2]. If left untreated for a long time, depression can lead to more serious health problems, including suicide [3]. Data shows that 1,023 people died by suicide due to depression in Indonesia from January to October of 2024 [4]. Despite the high risk, negative stigma causes more than 80% of sufferers to be reluctant to seek formal treatment, so many of them turn to platform X to express themselves anonymously [5], [6]. This makes social media a treasure trove of data, but unstructured data can pose analytical challenges, requiring text mining to analyze it [7].

There are many ways to detect or identify potential mental health issues. One method is to apply text mining techniques to analyze texts such as someone's social media confessions [8]. Text mining focuses on processing text-based data, which requires natural language processing (NLP) techniques to properly extract the data [9]. Various methods can be used for text mining; one example is the IndoBERT model, a BERT adaptation using transformers to determine contextual meaning in text, that has been trained on Indonesian texts [10]. Previous research has shown significant variation in mental health detection performance, with the RoBERTa model achieving 83% accuracy on Reddit data [11], outperforming the LSTM model which only achieved 70.89% on Twitter data [12]. Although Transformer-based models have proven to be superior, there is a noticeable gap in methodology, where most research is still limited to English-language data, lacks expert validation in labeling, and is rarely implemented as a practical tool.

Unlike previous studies, this study uses the X platform and focuses on analyzing tweets in Indonesian. The selected IndoBERT model has also been trained with emotion classification tasks and equipped with expert-assisted labelling. Furthermore, there is still limited research on X using transformer-based models for Indonesian-language potentially depressive tweet classifier that deployed as a practical web tool. Therefore, this study aims to develop and evaluate an IndoBERT-based model to classify Indonesian tweets as potentially depressive or normal and implement it in a Streamlit web application.

METHOD

Data Crawling

This study collected 5,000 tweets from X users between October 1st, 2024, and January 31st, 2025, using Tweet Harvest. The X platform is one of the most popular for expression and communication, it also provides an excellent API service, making it simple for anyone to retrieve data using API authentication tokens.

The main inclusion criteria applied include Indonesian language filters, the use of specific keywords, and certain time limits. Although effective, this keyword-based method has the potential for bias: selection bias and context bias. Selection bias risks ignoring relevant data that uses terms outside the keywords. Context bias risks capturing irrelevant content that has the same keywords, potentially compromising the analysis accuracy. To maintain data integrity, only original tweets are retained, while other columns containing replies, quotes, and usernames are removed before further analysis. Tweet searches were conducted by entering specific keywords below [10].

Table 1. Tweet Search Keyword

Potentially Depressed	Normal
Bersalah, Cemas, Depresi, Diabaikan, Dikucilkan, Frustasi, Gagal, Galau, Gelisah, Hilang kendali, Hampa, Kecewa, Kesepian, Lelah, Menyerah, Menyesal, Pasrah, Putus asa, Sedih, Sendirian, Stress, Tak berharga, Terasing, Terisolasi, Terluka, Terpuruk, Tertekan, Tersisih, Tidak berarti, Tidak berdaya, Tidak berguna, Tidak didengar, Tidak mampu	Antusias, Bahagia, Berharga, Berhasil, Bersama, Ceria, Dekat, Dihargai, Dipercaya, Diterima, Gembira, Mampu, Optimis, Positif, Riang, Semangat, Senang, Sukses, Tenang

Data Labelling

To maintain anonymity, all columns except the original tweet were deleted, then a "label" column was added for labelling results. Researchers, assisted by a psychiatrist, manually classified the data as depression (potential depression) or normal by interpreting the tweets contextually. The number of annotators was one of the limitations of this study, as only two people (including the psychiatrist) performed the labeling. Contextual interpretation is necessary because potentially depressive sentences cannot be identified solely by relying on specific keywords listed in Table 1 [13], but rather require an understanding of the complexity of language, context, and human emotions [14]. Additionally, not everyone who is potentially depressed uses words that directly reflect their condition. Some people may use more subtle language [15]. This limitation poses a risk of misclassification, where false positives (FP) may lead to stigmatization of healthy individuals, while false negatives (FN) may result in the undetection of individuals who are actually depressed. Therefore, manual labeling with psychological insight is essential to minimize this risk.

Data Preprocessing

Data preprocessing involves several stages, which are ment to prepare and refine textual data for further analysis. These stages typically include:

1. Case Folding
   Case folding refers to a text processing method that converts all letters to lowercase, for example, the word "Sedih" become "sedih".
2. Cleaning
   Cleaning is performed to remove characters or words that are considered disruptive to text data, including punctuation marks, hashtags, mentions, URLs, and emojis. For example, the phrase "viral #depression #life" becomes only "viral".
3. Normalization
   Normalization aims to convert non-standard words (abbreviations and slang) into standard words to reduce diversity. For example, the word "pdhl" would be converted into "padahal". To address this issue, researchers used an open-source file containing a list of slang Indonesian words from GitHub, namely colloquial-indonesian-lexicon.
4. Stopword Removal
   Stopword removal eliminates common words that are considered irrelevant to the analysis and do not contribute significantly to the meaning. Words such as "adalah", "di", "tapi", and

"jika" are examples of stopwords. The list source in this process uses the Satrawi library and a custom list contains a few additional words, such as “nih,” “sih,”, “kok”, and so on.

Data Splitting

Data splitting prevents overfitting and validates generalization on unseen data [16]. The data is split into two proportions in this study: train data for training and fine-tuning the model, and test data for evaluating the model's final performance. The data is divided into 80% training data and 20% testing data. This division has been stratified to maintain the same class proportions during the training and testing processes. The 80% of training data is also used during hyperparameter tuning.

IndoBERT Tokenization

IndoBERT is a derivative of BERT (Bidirectional Encoder Representation from Transformer) architecture used in NLP. IndoBERT is for Indonesian language processing and has a special tokenization technique, WordPiece Tokenization. The BERT model uses Transformer layers to analyze text, featuring an attention mechanism that emphasizes the contextual significance of words [17], [18]. BERT first converts sentences into numerical vectors, then converts them back into tokens. This step is called tokenization, which is very important for machine learning models. This tokenization process includes the input phase, addition of special tokens [CLS] and [SEP], and the encoding process. [CLS] is used to determine the input sentence category, [SEP] marks the end of a sentence, and [PAD] adds padding to shorter texts so all inputs are the same length [16], [19] with the truncation direction set to the default, which is head-only. The maximum length limit used is 128 tokens, as numerous studies have proven that this number provides good performance without requiring excessive resources [18], [20].

Pre-Fine-Tuning Model

This stage is the core of text processing using the IndoBERT model, which is done by creating a model from a pre-trained model. The base or pre-trained model used is the prediksi-emosi-indobert model that has been developed to classify six emotions, which is this model was trained using many Indonesian words, making it more accurate than the original model. The IndoBERT model architecture is modified at the output layer to accommodate binary (two classes) classification tasks. The last layer from the previous emotion classification layer was replaced with a new layer that had two output units, which were then passed to the softmax activation function to generate probabilities for the “normal” and “potentially depressed” classes. All parameter weights in the IndoBERT layer were not frozen during the fine-tuning process to allow the model to fully adapt to the linguistic characteristics of the dataset.

Hyperparameter optimization is performed automatically with the Optuna library, which Optuna is assigned to find the best combination of hyperparameters to improve the effectiveness of the model [21]. This experiment focused on a search space consisting of learning rate, batch size, and number of epochs. The objective function set in this search is to maximize the weighted f1 score on the validation set. The selection of this metric aims to maintain the balance of model performance between majority and minority classes amid class imbalance conditions. For computational efficiency, the pre-fine-tuning stage for hyperparameter search only uses a sub-

dataset of 20% of the entire training data and applies a maximum timeout of two hours. This strategy allows for extensive parameter exploration in a shorter time without significantly compromising data pattern representation.

Fine-Tuning Model

Fine-tuning process adjusts a trained model to improve performance and accuracy [22]. Hyperparameter values are also adjusted to optimize performance. Learning rate controls how fast a neural network learns from data, batch size is the number of data points used to adjust weights, and epoch shows the total number of times a network has processed all examples. The best-performing configuration used a learning rate of 2e-05, batch size 8, and 2 epochs. This process uses a whole dataset, applies a maximum timeout of two hours, as well as early stopping to immediately stop the entire process if the model's performance on the testing data does not show improvement for one consecutive epoch (after the last best performance).

Model Evaluation

Evaluating a model involves measuring its performance, which is determined through a 2x2 confusion matrix. This measurement requires True Positive (TP) which predictions are correct according to the facts, True Negative (TN) which predictions are negative according to the facts, False Positive (FP) which predictions are positive but the facts are negative, and False Negative (FN) which predictions are negative but the facts are positive. The following computations can then be performed using a confusion matrix:

1. Accuracy is a measure of how well a system classifies correctly, calculated as follows:
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$
2. Precision is the exactness between the number of correctly classified data and the total data predicted correctly by the system, defined as:
$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2)$$
3. Recall is the comparison between the success rate of the system in classifying data correctly and the total amount of data that is actually correct:
$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3)$$
4. F1 Score is a measure that combines recall and precision:
$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \times 100\% \quad (4)$$

This study sets the weighted f1 score as the main objective function in the hyperparameter optimization stage. This selection is based on the characteristics of the dataset, which experiences class imbalance between the “normal” and “potentially depressed” categories.

The macro f1 score gives equal weight to each class regardless of the number of samples, while the micro f1 score tends to be identical to the accuracy in binary classification. Meanwhile, the weighted f1 score calculates the average metric by considering the number of occurrences (support) of each label. This approach ensures that the hyperparameter search process produces a model that is fair to both the majority and minority classes, resulting in a more statistically stable model generalization.

Model Implementation to Streamlit

Finally, Streamlit web application is used to provide user with access to the model. Streamlit is a free Python toolkit for developing user-friendly web applications and dashboards from machine learning models [23]. Streamlit supports many Python libraries such as matplotlib, pandas, and Keras and is used to create this website's interface. The system is executed in a local environment with the following hardware specifications: an Intel Core i3-1005G1 CPU @ 1.20GHz, 4 GB of RAM, and an integrated graphics, which is utilized to evaluate the model's inference performance on standard consumer-grade hardware. It is important to remember that this model is only a supporting tool, not a diagnostic tool, and it does not replace a professional mental health assessment.

RESULT AND DISCUSSION

Tweet-Harvest and the Pandas library from Python were the tools used to collect data from X using token authentication API. This data was collected using Google Colab. The tweets collected met several criteria: they contained several keywords mentioned in Table I, limited to 5,000 data points, and within a specific time frame. The dataset has 15 attributes (columns), including full_text, as shown in Table 2 below.

Table 2. Crawling Data Results

full_text
aku sedih pdhl lg pgn di support
@muh_basra Keknya beban depresi pada cash for healing ga si kak (?) wkwkwk
gue udah di titik cape banget nutup nutupin image keluarga yang harmonis selama ini. iya gue harmonis tapi lebih banyaknya ga terasa. kontrol emosional nyokap yg ga pernah bisa dijaga peran support yang ga pernah gue dapet untuk gue ngelakuin apapun di kuliah
...
@iteramfs Gk usah kkn nder nanti rame2 bilang yg salah panitnya ngapain ngewajibin tapi lama

Of the 33 search keywords, the five words that occurred most frequently were "sedih" (1,104 times), "depresi" (385 times), "gagal" (379 times), "menyerah" (319 times), and "bersalah" (279 times). Then, the labeling of data involved adding label attributes, contain "depression" and "normal" classes as shown in Table 3.

Table 3. Data Labelling Results

full_text	label
aku sedih pdhl lg pgn di support	depression
@muh_basra Keknya beban depresi pada cash for healing ga si kak (?) wkwkwk	normal

The key features of this dataset are shown in Table 4 below. The distribution shows an imbalance between classes, with the majority class (normal) having almost twice the proportion of the minority class (potentially depressed) which can be observed more clearly in the confusion matrix and classification report.

Table 4. Key Dataset Characteristic

Key Features	Total
Keyword appearances	3,622 times
Average tweet length	20 words
Duplicate tweet	13
"Normal" label	3,287 (65.47%)
"Depressed" label	1,713 (34.26%)

Next come pre-processing to process unstructured and noisy from raw data, preparing it for the next stage, as shown in Table 5.

Table 5. Preprocessing Results

Step	Output
Raw	@muh_basra Keknya beban depresi pada cash for healing ga si kak (?) wkwkwk
Case Folding	@muh_basra keknya beban depresi pada cash for healing ga si kak (?) wkwkwk
Cleaning	keknya beban depresi pada cash for healing ga si kak wkwkwk
Normalization	kayaknya beban depresi pada cash for healing enggak si kak wkwkwk
Stopword Removal	kayaknya beban depresi cash for healing enggak kak wkwkwk

This process reduced the dataset to 4,987 clean data with 2 attributes (label and full_text), which 3,281 tweets are normal and 1,706 tweets are potentially depressed. A total of 13 tweets were removed because contained the duplicate content. Then, the labels are converted to numeric characters, where "depression" is converted to value 1 "normal" to 0.

The data is then used to train and measure the model. All data was divided into two groups: 80% for training and 20% for testing. There were 3,989 data points for training and 998 for testing. Then the data went into the tokenizing process. The first step is to access the pre-trained model, prediksi-emosi-indobert. The model's classification layers are modified to enable the production of two classes from the original six. This process handles padding (adding tokens) and truncation (cutting text) to ensure inputs are compatible and have the same length. Below is the IndoBERT tokenization steps. Table 6 below shows the conversion of text into numerical features such as "aku" becoming 304. All of these numbers come from the built-in IndoBERT vocabulary.

Table 6. IndoBERT Tokenizing Steps

Text	Input	Token
aku sedih padahal lagi pengin support	['[CLS]', 'aku', 'sedih', 'pd', '##hl', 'lg', 'pgn', 'di', 'support', '[SEP]']	['[CLS]', 'kek', '##nya', 'beban', 'depresi', 'pada', 'cash', 'for', 'he', '##aling', 'ga', 'si', 'kak', '(', '?', ')', 'wkwkwk', '[SEP]']

For further analysis, the model used a Trainer from Transformers with settings for evaluation, saving, and fixed seed for reproducibility. After adding settings and parameters, the model is tested on a test dataset. Prediction results are then converted into prediction labels, which are compared with the original labels. A weighted f1 score is used to evaluate the performance of the model during this stage. Researchers use a strategy to maintain computational efficiency with limited resources. This process used 20% of each dataset, with a two-hour maximum timeout. This approach uses the pre-trained IndoBERT model's transfer learning, where the model learns about language during pre-training.

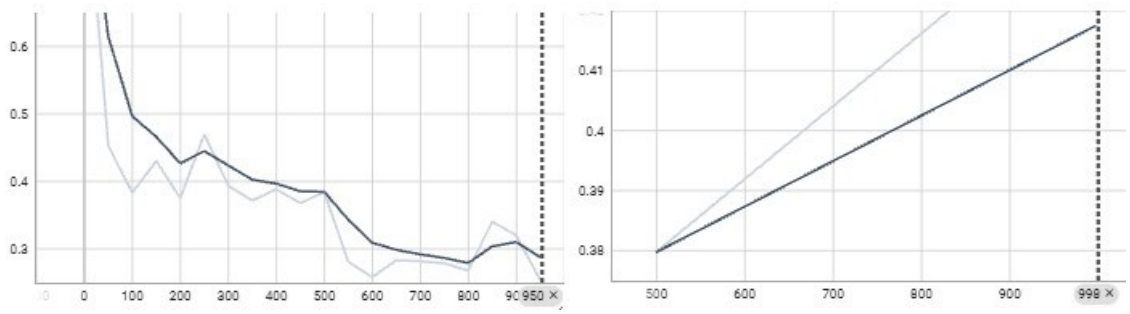
Furthermore, hyperparameter tuning and f1 score results will be saved for further analysis. Table 7 shows the results of the hyperparameters options that were selected by Optuna. From all the options, the best f1 score of 84% was obtained with a learning rate of 2e-05, a batch size of 8, and 2 epochs.

Table 7. Hyperparameter Selection Experiment Results

Trial	Status	Value	Learning Rate	Batch Size	Epoch
0	1	0.83	1e-05	16	3
1	1	0.84	1e-05	8	3
2	1	0.83	2e-05	16	3
3	1	0.83	2e-05	8	2
4	1	0.84	2e-05	8	2

In Table 7 above, “status 1” indicates a successful experiment, and the “value” column contains the F1 score value. It can also be seen that a learning rate of 2e-05 and a batch size of 8 work better than other configurations. This is because the value 2e-05 is suitable for small datasets so that the model can find generalization patterns. In addition, a small batch size provides a higher frequency of weighting, which is crucial for the model in understanding the semantic features of the text, while mitigating the risk of overfitting.

From the experiment before, best hyperparameters are reused. In the pre-fine-tuning stage, 20% of the dataset was used for exploration and resource savings. Here, in the fine-tuning stage, the whole dataset is used for better model performance and generalization. This step reveals training and validation loss values. Training loss measures how well the model learns, while validation loss measures performance on unseen data. Figure 1 below is a graph of training and validation loss.



(a) (b)

Figure 1. Model Performance (a) Decreased Training Loss, (b) Increased Validation Loss

The loss curve on the training loss graph decreases from 0.37 to 0.25, indicating that the model learns better the longer it is trained. The validation loss curve rises from 0.38 to 0.44, showing the model's decline in predicting validation data, despite falling in training loss. The decreasing training loss and increasing validation loss indicate overfitting after a certain number of steps. To address class imbalance, future relevant research should explore text data augmentation techniques or the use of weighted loss functions so that models have more pattern references and sensitivity to depressive texts. The overall performance of the model is then measured using a 2x2 confusion matrix. Figure 2 shows the results of the confusion matrix with 282 TP, 584 TN, 59 FP, and 73 FN.

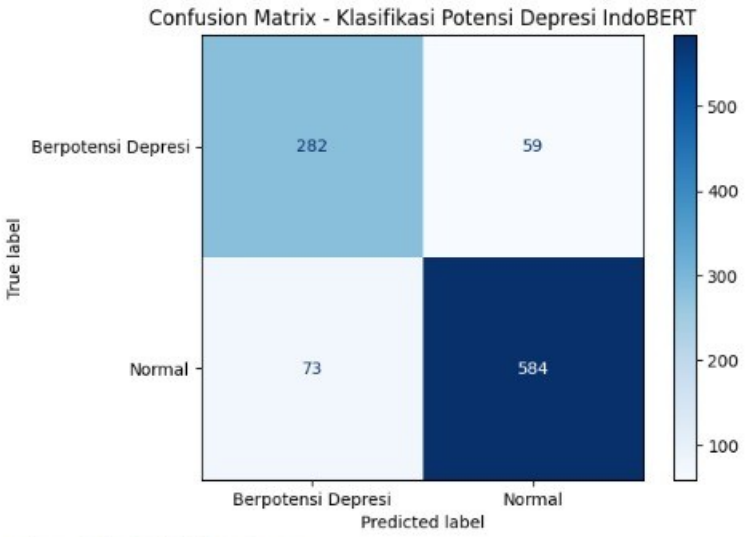


Figure 2. Confusion Matrix Result

Those four metrics above are used to measure accuracy, precision, recall, and f1 score, which are presented in Table 8 below.

Table 8. Classification Report

Result	Precision	Recall	F1 Score	Support
Normal	0.91	0.89	0.90	657
Potentially Depressed	0.79	0.83	0.81	341
Accuracy			0.87	998
Macro Avg	0.85	0.86	0.85	998
Weighted Avg	0.87	0.87	0.87	998

Further explanation is provided in Table 9 below.

Table 9. Metrics Analysis and Practical Implications

Class	Metrics Explanation	Practical Implications
Normal	The precision, recall, and f1 score are very high, meaning that the model understands the patterns of this class.	There are 59 FP, indicating that mentally healthy users are incorrectly identified as "potentially depressed." This may cause anxiety in users.
Potentially Depressed	Lower precision, recall, and f1 scores indicate that the model is less trained in identifying tweets with potential depression.	There are 79 FN, indicating that users who actually experience symptoms of depression are not detected by the system. This can inhibit them from getting support and treatment.

The results of this study achieved an accuracy rate of 87%, surpassing the performance of the LSTM (70.89%) and RoBERTa (83%) models in previous studies. The f1 score (0.81) was influenced by the initialization of the emotion classification model, which deepened the model's understanding, as well as labeling by experts, which ensured the quality of the training data. However, there are limitations in the form of an unbalanced data distribution and a limited number of expert annotators, which resulted in 79 cases of false negatives (FN). Therefore, future development requires a more balanced dataset to improve detection sensitivity.

Finally, Streamlit is run to connect localhost to the public internet. The main features for the website interface are text input, word count (maximum 100 words), prediction button, prediction results displayed along with confidence scores, and warnings if the text is empty or exceeds the limit. To use this website, users must input text, and then click the "prediksi" button. Then, the classification results will appear in the form of "normal" or "potentially depressed" along with the confidence score. The confidence score is a score that shows how confident the model is in its prediction, where the score ranges from 0% to 100%. The confidence score can guide prioritazion for human review but should not be interpreted as probability of having depression. The following are the Streamlit user interface results in Figure 3.

(a) (b)

Figure 3. Streamlit's User Interface (a) Normal Result, (b) Potentially Depressed Result

It is known that the IndoBERT model assesses the text in (a) as a normal sentence with a confidence level of 93%. Meanwhile, the second text in (b) is assessed as a potentially depressive sentence with a confidence level of 86%. This demonstrates that the website-based depression prediction application with Streamlit is running well.

CONCLUSION

The conclusion of this study is that the IndoBERT-based classifier developed works well for text mining in classifying potentially depressive and normal tweets, with an accuracy of 87%, precision of 87%, recall of 87%, and an f1 score of 87%. In addition, this model has been successfully deployed as a Streamlit web application. This model will play a more significant role as an early screening tool to assist psychiatrists, rather than a diagnostic tool. There are weaknesses in this study, which include class imbalance and a lack of annotators. Therefore, improvements that can be made for relevant future studies are to focus more on class balance, both for the potentially depressed and the normal class, and to add annotators to make the labeling process more efficient.

REFERENCE

[1] Kemenkes RI, "Cegah Bunuh Diri, Kemenkes Ajak Remaja Bicara Soal Kesehatan Mental," 2024. https://sehatnegeriku.kemkes.go.id/baca/umum/20240917/2446492/cegah-bunuh-diri-kemenkes-ajak-remaja-bicara-soal-kesehatan-mental/ (accessed Nov. 01, 2024).

[2] Y. A. Beo *et al.*, *Ilmu Keperawatan Jiwa dan Komunitas*. PENERBIT MEDIA SAINS INDONESIA, 2022.

[3] S. Aloysius and N. Salvia, "Analisis Kesehatan Mental Mahasiswa Perguruan Tinggi X Pada Awal Terjangkitnya Covid-19 di Indonesia," *J. Citizsh. Virtues*, vol. 1, no. 2, pp. 83–97, 2021, doi: 10.37640/jcv.v1i2.962.

[4] GoodStats Data, "Angka Kasus Bunuh Diri di Indonesia Meningkat 60% dalam 5 Tahun Terakhir," 2024. https://data.goodstats.id/statistic/angka-kasus-bunuh-diri-di-indonesia-meningkat-60-dalam-5-tahun-terakhir-2FzH6 (accessed May 05, 2025).

[5] A. Muhawarman, "Memutus Rantai Stigma Kesehatan Jiwa," *Kemenkes*. 2024.

[6] Z. Maritska, A. B. Prananjaya, S. P. Nabila, and N. Parisa, "Promosi Kesehatan Jiwa Berbasis Media Sosial (Instagram Live) bagi Masyarakat di Masa Pandemi COVID-19," *Wal'afiat Hosp. J.*, vol. 04, no. 01, pp. 13–22, 2023.

[7] E. Safitri, W. A. Syukrilla, and I. N. L. Fitriana, "Logistic Regression for Sentiment Analysis of Insecurity Phenomena on Platform X," *J Statistika*, vol. 18, no. 1, pp. 948–956, 2025.

[8] D. K. A. Astutik, A. Indrasetianingsih, and F. Fitriani, "Penerapan Text Mining pada Analisis Sentimen Pengguna Twitter Layanan Transportasi Online Menggunakan Metode Density Based Spatial Clustering of Applications with Noise (DBSCAN) dan K-Means," *J Statistika*, vol. 15, no. 1, pp. 184–194, 2022.

[9] X. Liu *et al.*, "Emotion classification for short texts: an improved multi-label method," *Humanit. Soc. Sci. Commun.*, vol. 10, no. 1, pp. 1–9, 2023, doi: 10.1057/s41599-023-01816-6.

[10] G. F. Situmorang and Purba, "Deteksi Potensi Depresi dari Unggahan Media Sosial X Menggunakan Teknik NLP dan Model IndoBERT," *Build. Informatics, Technol. Sci.*, vol. 6, no. 2, pp. 649–661, 2024, doi: 10.47065/bits.v6i2.5496.

[11] I. Ameer, M. Arif, G. Sidorov, H. Gòmez-Adorno, and A. Gelbukh, "Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning," *arXiv Prepr. arXiv2207.01012*, 2022.

[12] B. Kholifah, I. Syarif, and T. Badriyah, "Mental disorder detection via social media mining using deep learning," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 5, no. 4, pp. 3019–316, 2020, doi: https://doi.org/10.22219/kinetik.v5i4.1120.

[13] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th Int. Conf. Comput. Linguist. Proc. Conf.*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[14] J. de J. Titla-Tlatelpa, R. M. Ortega-Mendoza, M. Montes-y-Gómez, and L. Villaseñor-Pineda, "A profile-based sentiment-aware approach for depression detection in social media," *EPJ Data Sci.*, vol. 10, no. 1, 2021, doi: 10.1140/epjds/s13688-021-00309-3.

[15] F. Alhamed, R. Bendayan, J. Ive, and L. Specia, "Monitoring Depression Severity and Symptoms in User-Generated Content: An Annotation Scheme and Guidelines," *Proc. 14th Work. Comput. Approaches to Subj. Sentim. Soc. Media Anal.*, pp. 227–233, 2024, [Online]. Available: https://aclanthology.org/2024.wassa-1.18

[16] W. A. Hidayat and V. R. S. Nastiti, "Perbandingan Kinerja Pre-trained IndoBERT-Base dan IndoBERT-Lite pada Klasifikasi Sentimen Ulasan TikTok Tokopedia Seller Center dengan Model IndoBERT," *J. Sist. Inf.*, vol. 11, no. 2, pp. 13–20, 2024, doi: 10.30656/jsii.v11i2.9168.

[17] X. Luo, H. Ding, M. Tang, P. Gandhi, Z. Zhang, and Z. He, "Attention Mechanism with BERT for Content Annotation and Categorization of Pregnancy-Related Questions on a Community QA Site," *Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020*, pp. 1077–1081, 2020, doi: 10.1109/BIBM49941.2020.9313379.

[18] K. Zeberga, M. Attique, B. Shah, F. Ali, Y. Z. Jembre, and T.-S. Chung, "A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model," *Comput. Intell. Neurosci.*, vol. 2022, 2023, doi: 10.1155/2022/7893775.

[19] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *Int. J. Electr. Comput.*

Eng., vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[20] T. Oswari, M. Murniyati, T. Yusnitasari, N. Nurasiah, and S. Wijay, "Sentiment Analysis of Indonesian Youtube Reviews About Lesbian, Guy, Bisexual and Transgender (LGBT) using IndoBERT Fine Tuning," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 15, no. 1, p. 26, 2024, doi: 10.24843//lkjiti.2024.v15.i01.p03.

[21] L. H. Lai *et al.*, "The Use of Machine Learning Models with Optuna in Disease Prediction," *Electron.*, vol. 13, no. 23, pp. 1–20, 2024, doi: 10.3390/electronics13234775.

[22] H. Imaduddin, F. Y. A, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," vol. 14, no. 8, pp. 113–117, 2023.

[23] S. P. Revathy, M. Sindhuja, and R. Jayashree, "Streamlit-based Web Application for Parkinson ' s Detection using Machine Learning," no. January, 2025, doi: 10.36548/jaicn.2024.4.006.