khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Automatic Categorization of Mental Health Frame in Indonesian X (Twitter) Text using Classification and Topic Detection Techniques

Rizky Indrabayu[1], Nico Ardia Effendy[2], Setio Basuki[3*]

*setio_basuki@umm.ac.id

[1-3]Informatics Department

Universitas Muhammadiyah Malang

Malang, Indonesia

Abstract-This paper aims to develop a machine learning model to detect mental health frames in Indonesian-language tweets on the X (Twitter) platform. This research is motivated by the gap in automatically detecting mental health frames, despite the importance of mental health issues in Indonesia. This paper addresses the problem by applying classification and topic detection methods across various mental health frames through multiple stages. First, this paper examines various mental health frames, resulting in 7 main labels: Awareness, Classification, Feelings and Problematization, Accessibility and Funding, Stigma, Service, Youth, and an additional label named Others. Second, it focuses on constructing a dataset of Indonesian tweets, totaling 29,068 data, by filtering tweets using the keywords "mental health" and "*kesehatan mental*". Third, this paper conducts data preprocessing and manual labeling of a random selection of 3,828 tweets, chosen due to the impracticality of labeling all data. Finally, the fourth stage involves conducting classification experiments using classical text features, non-contextual and contextual word embeddings, and performing topic detection experiments with three different algorithms. The experiments show that the BERT-based method achieved the highest accuracy, with 81% in the 'Others' vs. 'non-Others' classification, 80% in the seven main label classifications, and 92% in the seven main labels classification when using GPT-4-powered data augmentation. Topic detection experiments indicate that the Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI) algorithms are more effective than the Hierarchical Dirichlet Process (HDP) in generating relevant keywords representing the characteristics of each main label.

Keywords: BERT, GPT-4, Mental Health, Topic Detection, Word-Embedding

Article info: *submitted November 28, 2023, revised April 23, 2024, accepted January 13, 2025*

1. Introduction

Mental health, as defined by the World Health Organization (WHO), is a condition of mental well-being that helps people to manage life's stressors, recognize their potential, perform successfully in learning and working, and contribute to their community [1]. This definition highlights the positive aspects of mental health, emphasizing its role in enabling people to lead fulfilling lives and engage constructively with others in their community. However, the COVID-19 pandemic led to a worldwide rise in mental health issues like depression and anxiety [2]. This trend was also evident in Indonesia, where the rate of mental health issues among self-check users increased from 70.7% in 2019 to 80.4% in 2020, and further to 82.5% by 2022 [3]. In handling this phenomenon, a comprehensive method to detect mental health issues is required. One of the solutions is using Machine Learning (ML) to detect and categorize mental health-related topics by analyzing the frames related to it.

In the past two decades, ML has gained prominence due to its effectiveness in addressing various detection problems. ML's utilization in the mental health field has surged for tasks such as identifying individual behaviors that aid in understanding mental health symptoms and risk factors [4], as well as detecting and diagnosing mental health conditions through the analysis of patient data [5]. The rapid expansion of social media platforms has facilitated the detection of mental health issues, with X (Twitter) being a popular platform for this purpose. Analyzing mental health issues on Twitter offers several advantages, including (a) predict and monitor mental health issues based on tweet activity [6]; (b) real-time data retrieval, enabling the tracking of mental health trends [7]; and (c) the ability to apply Natural Language Processing (NLP) techniques to analyze Twitter data, aiding in the identification of mental health-related patterns [8]. Considering these factors, it is not surprising that various research studies focused on mental health issues make use of Twitter data.

Numerous studies currently employ ML-based approaches for sentiment analysis on mental health topics in Indonesian Twitter text. The first research discusses mental disorders on Twitter using the Long Short Term Memory (LSTM) algorithm and achieved 70.89% accuracy in sentiment analysis with three sentiment labels:

positive, neutral, and negative [9]. The next research explored the impact of COVID-19 on mental health in Indonesia using Twitter data. It achieved 74.36% accuracy with the NB algorithm and 76.42% with the Support Vector Machine (SVM) for two classes: psychological problems and anxiety problems [10]. The next research employed the NB algorithm for sentiment analysis on Twitter replies, achieving an accuracy of 79% with three classes: positive, neutral, and negative [11]. Following this, research explored mental health sentiment on Twitter using the (K-Nearest Neighbor) KNN method. The classification results of this research achieved an accuracy of 58.39% [12]. Another study focused on identifying depression among Twitter users based on two classes, namely "yes" and "no", using Multinomial NB and SVM, which reached an accuracy of 88.07%, while the Convolutional Neural Network (CNN) algorithm achieved an accuracy of 91.23% [13]. The presence of these studies demonstrates that detecting mental health issues continues to attract significant attention.

Through literature review, several research gaps from existing research have been identified. The main gap is the lack of research on automatically detecting or classifying mental health frames in Indonesian Twitter (X) text, despite the importance of mental health issues in Indonesia. Here are the specific details of these issues. First, previous research was dominated by sentiment analysis that used a maximum of only 3 classes, i.e., positive, neutral, and negative. Second, most of the previous research used traditional ML as a classification method. Third, no research has implemented a topic detection method on mental health issues in Indonesia. To address these gaps, this paper aims to categorize more detailed frames related to mental health issues in Indonesian text using modern classification methods and topic detection algorithms.

The mental health frame used in this paper consists of 7 main labels, namely "Awareness", "Feelings and Problematization", "Classification", "Accessibility and Funding", "Stigma", "Service", "Youth", and an additional label "Others". Classification uses several scenarios, from combining ML and classical text representation to modern word-embedding techniques. The topic detection is implemented using 3 different types of algorithms. To be more specific, this paper provides several contributions:

1. Using more detailed mental health frames on Indonesia's Twitter based on classification and topic detection techniques, which were not covered in previous research.
2. In all stages of classification experiments, Bidirectional Encoder Representations from Transformers (BERT) with 'indobertweet-base-uncased' show superior performances compared to other techniques.
3. The classification experiments show diverse levels of accuracy: 81% in stage-1 for binary classification 'Other' versus 'non-Other', 80% in stage-2 for classifying the seven main labels, and an impressive 92% in stage-3, by adding augmented data for the same seven labels.
4. The obtained performances were considered competitive compared to the previous studies, as this paper used 7 labels while the previous research only employed a maximum 3 labels.
5. Topic detection experiments showed that Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI) generate a list of keywords relevant to all major labels' characteristics. In contrast, the Hierarchical Dirichlet Process (HDP) generates less relevant keywords.

2. Methods

The mental health-related frame detection system on the Indonesian Twitter platform is implemented through several stages, namely Twitter data collection, Twitter data labelling, data pre-processing, classification stages, and topic detection stages. The proposed system architecture is depicted in Figure 1.

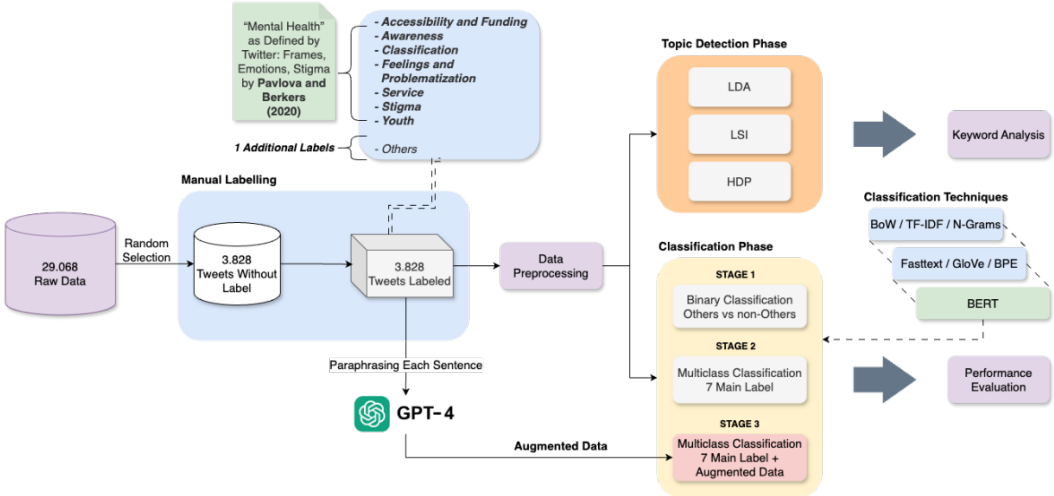

Figure 1. The architecture of mental health-related frame detection system on Indonesian-language twitter platform

a. Twitter Data Collection

This paper uses data from Indonesian-language Twitter tweets from 2020-2023. The data collection uses a Python library that supports the Twitter platform. The keywords "*kesehatan mental*" and "mental health" are employed for the crawling purpose. This data-gathering stage resulted in 29,068 data containing dates, usernames, number of likes, and tweets.

b. Twitter Data Labelling

This stage assigns a label to each retrieved tweet, primarily utilizing 7 frames as defined in the research by Pavlova and Berkers

[14]. Below are explanations for each of these mental health frames:

1. Awareness: Individual awareness of mental health, including mental illness.
2. Feelings and Problematization: Expressing feelings related to mental health, such as anxiety, stress, etc.
3. Classification: Classifying or categorizing mental health issues, such as identifying symptoms or diagnoses.
4. Accessibility and Funding: Accessibility and funding of mental health services.
5. Stigma: Society's stigma towards mental health problems.
6. Service: Mental health services, such as seeking or offering help.
7. Youth: Teenager's mental health problems.

This paper introduces an extra "Others" label for tweets not falling under the 7 primary labels. This additional label is crucial for accurately representing the real tweet distribution. In total, 8 labels were employed in this paper. Table 1 provides tweet examples for each label, including both the main and additional labels.

Table 1. Example of Tweets of Each Mental Health Frame

No	Tweets Examples	Frame
1	*Jangan remehkan mental health issue* (Do not underestimate mental health issues.)	Awareness
2	*di bilang ga sayang sama yang udah ngerusak mental health gua, wkwk. lu liatnya cuma luar tapi ga tau isinya kan, lawak!* ("They say I do not care about those who have harmed my mental health, haha. You only see the outside but do not know what's inside, it is funny!")	Feelings and Problematization
3	*Banyak kondisi kesehatan mental seperti bipolar, depresi, atau anxiety juga disebabkan karena oversharing.* ("Many mental health conditions such as bipolar, depression, or anxiety are also caused by oversharing.")	Classification
4	💚 *guys konsul kesehatan mental via halodoc tuh worth ga sih? aku blm pernah sama sekali ke psikolog, sbrnya uda ada rencana tp aku blm ada uang untuk konsul offline dan gapunya bpjs juga*😔 (💚 Guys, is the mental health consultant at Halodoc worth it? I've never been to a psychologist, so I have a plan, but I don't have the money for an offline consul and I don't have BPJS either😔)	Accessibility and Funding
5	*stop berpikir kl org yg ke psikolog itu punya masalah mental health* (Stop thinking that people who go to a psychologist have mental health issues)	Stigma
6	*Saran dr psikolog pro yg mendampingiku unt keluar dr mental health issue ini agar aku mencoba hal baru diluar dr yg biasa dilakukan. Inilah awalnya, aku putuskan unt daftar anggota vlive (amaze sih aku ngerti apa yg diomongin pas live pdhl bhs korea aku cetek bgt)* (The advice from the professional psychologist who accompanied me to get out of this mental health issue was for me to try new things outside of what I usually do. This is the beginning, I decided to register as a member of Vlive (amaze I understand what is being said during live even though I speak Korean very well))	Service
7	*banyak gen z yang memiliki pengetahuan kesehatan mental namun banyak juga yang rentan depresi karena setiap proses mereka yang cukup ambisi #MahasiswiGenz* (Many Gen Z individuals know about mental health, but some are still vulnerable to depression due to their ambitious nature throughout their processes. #GenZStudents)	Youth
8	*@ichahrnssa Terima kasih doa-doanya Icha. Glad to hear kamu udah di kantor yang lebih baik buat kesehatan fisik dan mental kamu sekarang* 😊😊 (@ichahrnssa Thank you for your prayers, Icha. Glad to hear that you are now in a better office environment for your physical and mental health 😊😊)	Others

Two people conducted the process of manual labelling. The initial step involved analyzing 2,000 tweets, followed by the next 2,000 tweets. Due to specific data containing extensive Indonesian vocabulary that could not be interpreted, excluding such data from the research was necessary. As a result, the final labelled dataset consisted of 3,828 tweets. The sample dataset will be used to develop the classification model, chosen due to the impracticality of labelling all data. The distribution of all labelled sample data is depicted in Figure 2.

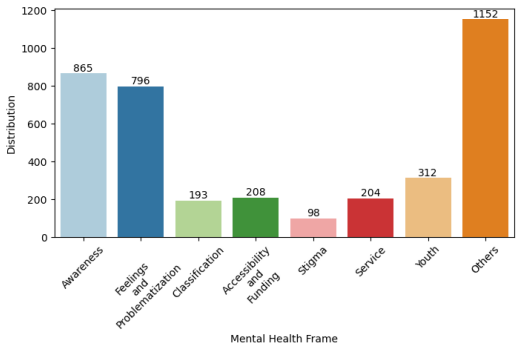

Figure 2. Distribution of all labelled sample dataset

c. Pre-processing Twitter Data

The data pre-processing stage is crucial to prepare the data for classification and topic detection. It involves three main steps: (a) Converting sentences to lowercase, eliminating tags, links, punctuation, and emojis; (b) Standardizing short words or slang terms frequently encountered in the data, such as replacing "*krn*" with "*karena*" (because), "*dmn*" with "*dimana*" (where), "*adlh*" with "*adalah*" (am/is/are), etc. The slang word dataset is sourced from a study on commonly used everyday language in Indonesia [15]; and (c) Removing stop words like "*aku*" (I), "*kamu*" (you), "*yang*" (that), and etc.

d. Classification Stages and Classification Scenario

This stage comprises three parts. In Stage 1, a binary classification process is performed to categorize tweets into "Others" and "non-Others" (7 main labels). This stage aims to segregate tweets into the main or "Others" labels. This step is crucial since tweets in the "Others" category hold valuable information and cannot be excluded. While stage 2 focuses on performing multiclass classification to categorize the data into the 7 main classes using original data distribution, stage 3 performs the same task using an

augmented dataset. Figures 3 and 4 present the data distributions for binary and multiclass classifications, respectively, while Figure 2 shows the distribution for stage 3.

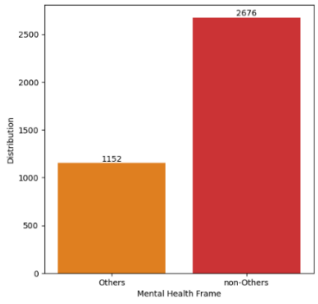

Figure 3. First stage data distribution (2 Classes)

The classification process uses a wide range of techniques, from classical methods like text representation to transformer-based models. (a) Classical text representations, such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF/IDF), and n-grams, are trained using neural network architectures. (b) Non-contextual word embeddings, including Fasttext (indonesian-embedding), Byte-Pair Encoding (BPE) (indonesian-embedding), and GloVe (multilingual-embedding), are employed. These embeddings are further processed using BiLSTM (Bidirectional Long Short-Term Memory) architecture to enhance performance. (c) Contextual word embeddings are utilized through pre-trained transformer-based models like Bidirectional Encoder Representations from Transformers (BERT). This paper specifically employs two pre-trained BERT models: 'indolem/indobert-base-uncased' and 'indolem/indobertweet-base-uncased' [16][17]. For all classification experiments, this paper specifies a 90% training and 10% testing data split.

This research suggests employing a data augmentation approach to handle the imbalance in the original labelled dataset for the stage 3 classification. The process is performed using the capabilities of GPT-4, a state-of-the-art Generative Pre-trained Transformer, version 4 [18]. Specifically, GPT-4 is prompted to paraphrase sentences across five out of the seven main labels with insufficient data distribution: Classification, Accessibility and Funding, Service, Stigma, and Youth.

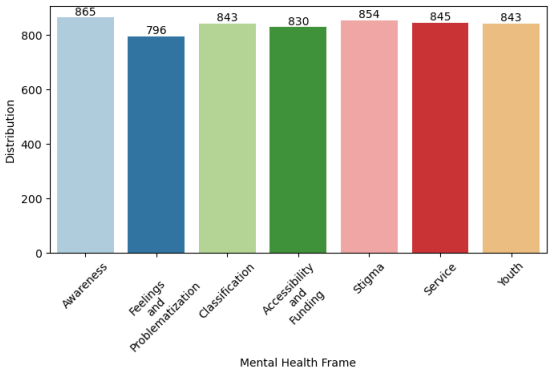

Figure 4. The data distribution used in the third stage (original data + augmented data)

e. Topic Detection Stages

A topic detection algorithm extract keywords representing specific topics in the text. Three algorithms are employed for topic detection: Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), and Hierarchical Dirichlet Process (HDP). LDA assumes that each document comprises a mix of topics, and each topic consists of word combinations. It determines a document's most likely topic by analyzing word frequencies [19]. LSI is a technique that identifies underlying topics by analyzing word-document relationships and identifying common occurrence patterns [20]. HDP is a non-parametric Bayesian approach for clustered data, serving as a robust mixed-membership model for unsupervised data grouping. Unlike LDA, which requires specifying the number of topics, HDP determines the number of topics from the data [21]. In this paper, the topic detection is conducted in a supervised manner, meaning the algorithms work on original data that has already been labelled.

3. Result

This section presents experiment results for classification and topic detection. We compare the performance of various techniques in the first, second, and third-stage classification, primarily using classification accuracy as the key performance metric. Considering the data class imbalance in both stages, we also utilize indicators like macro-average precision, macro-average recall, and macro-average F1 score to compare models [22]. For topic detection, we assess the results by evaluating the top 10 keywords that best represent each of the 7 main labels.

a. The Result on The First Stage Classification (2 Classes)

The top-performing model in the first-stage classification is BERT, using the 'indolem/indobertweet-base-uncased' model trained on large Indonesian-language Twitter data. This model achieved a macro average precision of 78%, macro average recall of 71%, macro average F1 of 73%, and an accuracy of 81%. To achieve this, BERT has been fine-tuned using AdamW optimizer with a learning rate of 9e-5, epsilon of 1e-8 and epoch of 3. Machine learning models combined with classic features like BoW, TF-IDF, and n-Grams performed comparably with non-contextual word-embedding models such as Fasttext, BPE, and GloVe, as well as contextual word embedding BERT, especially when using 'indolem/indobertweet-base-uncased' in almost all performance indicators. The performance comparison between all text representation methods of the first stage is shown in Table 2.

Table 2. First Stage Classification Result (Others vs non-Others)

Model	Macro Avg. Precision	Macro Avg. Recall	Macro Avg. F1	Acc.
BoW	0.62	**0.71**	0.62	0.72
TF-IDF	0.68	**0.71**	0.69	0.77
2-Grams	0.61	0.63	0.62	0.72
3-Grams	0.62	0.62	0.62	0.70
Fasttext (Pretrained: Indonesian)	0.67	0.66	0.66	0.72
BPE (Pretrained: Indonesian)	0.68	0.69	0.69	0.75
GloVe (Pretrained: Twitter Multilingual)	0.63	0.62	0.63	0.71
BERT: indobert-base-uncased	0.63	0.58	0.58	0.70

BERT: indobertweet-base-uncased	**0.78**	**0.71**	**0.73**	**0.81**

b. The Results on The Second and Third Stages Classification (7 Classes)

The second stage and third classifications are more complex than the first since they involve seven labels. The best second-stage classification performance is achieved by BERT using the 'indolem/indobertweet-base-uncased' model, with a macro-average precision of 75%, macro-average recall of 76%, macro-average F1 of 73%, and an accuracy of 80%. Using augmented data combined to balance the data distribution on the third stage classification improves all performance indicators significantly. Utilizing the same BERT model, we demonstrated substantial improvements in all performance indicators: a 16% rise in Macro Average Precision (from 75% to 91%), a 15% increase in Macro Average Recall (from 76% to 91%), an 18% increase in Macro Average F1 (from 73% to 91%), and a 12% improvement in overall accuracy (from 80% to 92%). These performances were achieved by using AdamW optimizer with a learning rate of 1e-4 (second stage) and 8e-5 (third stage), epsilon of 1e-8, and epoch of 3. A performance comparison between the second stage and the third stage is presented in Table 3.

Table 3. Comparison Classification Result (Second and Third Stages)

Model	Macro Avg. Precision		Macro Avg. Recall		Macro Avg. F1		Accuracy	
	Original Data	Original + Augmented Data	Original Data	Original + Augmented Data	Original Data	Original + Augmented Data	Original Data	Original + Augmented Data
BoW	0.45	0.85	0.49	0.85	0.46	0.85	0.61	0.85
TF-IDF	0.61	0.85	0.69	0.86	0.64	0.85	0.71	0.86
2-Grams	0.46	0.79	0.70	0.82	0.49	0.80	0.63	0.80
3-Grams	0.30	0.67	0.64	0.76	0.31	0.68	0.48	0.68
Fasttext (Pretrained: Indonesian)	0.60	0.80	0.63	0.81	0.60	0.80	0.72	0.81
BPE (Pretrained: Indonesian)	0.58	0.83	0.58	0.82	0.57	0.82	0.66	0.83
GloVe (Pretrained: Twitter-Multilingual)	0.63	0.85	0.61	0.59	**0.85**	0.58	**0.85**	0.85
BERT: indobert-base-uncased	0.56	0.88	0.61	0.87	0.57	0.87	0.67	0.87
BERT: indobertweet-base-uncased	**0.75**	**0.91**	**0.76**	**0.91**	0.73	**0.91**	0.80	**0.92**

c. Experiment Result on Topic Detection Algorithm

Three topic detection algorithms, LDA, LSI, and HDP, are implemented based on supervised approaches using Python's Gensim library. Each technique extracts the top 10 keywords for each of the 7 main labels. Evaluation is conducted manually by assessing the relevance of the generated keywords to the definition of each mental health frame. The evaluation follows these steps: (a) Examining the definition of each label and assessing the correlation between the generated keywords and the definitions; (b) Categorizing these correlations into three groups: "related" (take care, depression, psychologist, child, symptom, etc.), "unrelated" (like that, i don't know, he, salad, etc.), and "neutral" (work, few, lecture, fictional, choose, etc.); (c) Calculating the number of correlated categories for each algorithm and each label.

In general, HDP is less competitive than LDA or LSI in generating keywords representing the definition of each label. The top performance on this task is LDA since it produced the highest number of "related"-keywords and zero "unrelated"-keywords. Even though LSI generated almost a similar number of "related"-keywords as LDA, it has a lot of "neutral"-keywords. HDP obtained the lowest performance by showing the highest "neutral" and "unrelated"-keywords. While the distribution of these three types of associated keywords is illustrated in Figure 5, the LDA-generated keywords for wordcloud visualization are presented in Figure 6.

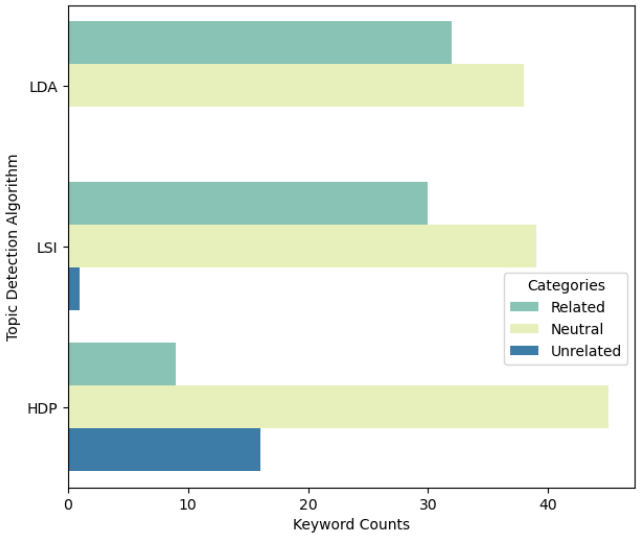

Figure 5. Distribution of keywords falls into related, unrelated, and neutral on the LDA, LSI and HDP

(6.1). Label: Awareness (6.2). Label: Stigma (6.3). Label: Feelings and Problematization (6.4). Label: Accessibility and Fundings

(6.5). Label: Service (6.6). Label: Classification (6.7). Label: Youth

Figure 6. The worldcloud of the top-10 LDA-generated keywords

4. Discussion

This section presents a discussion of two tasks covered in this paper. In the first task, the classification of all stages, BERT demonstrated the highest results and outperformed classical text representation methods and non-contextual word embedding techniques. The BERT superior's performance is caused by several factors. Contextual word embedding effectively understands word meanings and relationships by representing words in a continuous vector space and through a self-attention mechanism [23][24]. Contextual word embeddings are more effective in handling imbalanced data since it has been trained on extensive corpora [25]. Our experiment has also shown that the performance achieved by the BERT model that has been fine-tuned using the Indonesian Twitter text "indobertweet-base-uncased" performed better than the generic Indonesian text "indobert-base-uncased". Twitter-based BERT model is more familiar with the tweet's short and informal characteristics. There were significant improvements for all performance metrics when adding GPT-4-powered augmented data on five of seven labels. We want to point out that, although a direct comparison is not possible due to varying classification scenarios, tasks, and datasets, our paper still attained the highest accuracy even considering the quantity of labels and complexity of classifying when compared to previous studies that focused on sentiment analysis (2-3 labels). Table 4 shows the performance comparison with previous studies on the same topic.

Table 4. Performance Comparison with Previous Studies

	B. Kholifah et al., (2020) [9]	N. Fatimah et al., (2021) [10]	K. Yan et al., (2022) [11]	M. L. Wicaksono et al., (2022) [12]	B. Nurfadhila et al., (2023) [13]	**This Paper**
Number of Labels	3	2	3	2	2	**7**
Accuracy	70.89%	76.42%	79%	58.39%	91.23%	**92%**

In the topic detection task, HDP generates less relevant keywords than LDA and LSI. The LDA-generated keywords are still dominated by the terms "mental" and "health" along with label-characterized terms such as "keep", "take", or "care" (Awareness), "counseling", "psychologist" (Service), etc. This situation shows that while ML models effectively categorize tweets, analyzing each word individually can lead to misunderstandings in topic detection. For future research, we plan to develop phrase or n-Grams level topic modeling to provide better insight into each label characteristics.

5. Conclusion

This research aims to build a mental health frame detection system on the Indonesian-language Twitter platform, using classification techniques and topic detection algorithms. This paper is motivated by the global importance of mental health topics and the lack of related research in Indonesia. Given the importance of mental health issues in Indonesia, there has been no research on the automatic detection or classification of mental health frames in Indonesian Twitter (X) text. Our experiment shows that leveraging a combination of the BERT model fine-tuned with Indonesian tweets and GPT-4-powered augmented dataset through paraphrasing achieved the best results compared with other techniques. The topic detection approach on mental health issues helps provide insight into two aspects, namely visualizing each mental health topic and providing relevant keywords on each topic. Additionally, this paper introduces a prediction system that not only automates detection but also aims to increase awareness, attention, and respect towards mental health issues in Indonesia.

Reference

[1] World Health Organization, *Mental Health Atlas 2020*. 2020. [Online]. Available: https://www.who.int/publications/i/item/9789240036703

[2] World Health Organization, "Mental Health and COVID-19 : Early evidence of the pandemic ' s impact," *Sci. Br.*, vol. 2, no. March, pp. 1–11, 2022, [Online]. Available:

https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1

[3] Ellyana Dwi Farisandy, Azzahra Asihputri, and Jennifer Shalom Pontoh, "Peningkatan Pengetahuan Dan Kesadaran Masyarakat Mengenai Kesehatan Mental," *Disem. J. Pengabdi. Kpd. Masy.*, vol. 5, no. 1, pp. 81–90, 2023, doi: https://doi.org/10.33830/diseminasiabdimas.v5i1.5037.

[4] A. Thieme, D. Belgrave, and G. Doherty, "Machine Learning in Mental Health: A systematic review of the HCI literature to support the development of effective and implementable ML Systems," *ACM Trans. Comput. Interact.*, vol. 27, no. 5, 2020, doi: https://doi.org/10.1145/3398069.

[5] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychol. Med.*, vol. 49, no. 9, pp. 1426–1448, 2019, doi: https://doi.org/10.1017/S0033291719000151.

[6] N. H. Di Cara, V. Maggio, O. S. P. Davis, and C. M. A. Haworth, "Methodologies for Monitoring Mental Health on Twitter: Systematic Review," *J. Med. Internet Res.*, vol. 25, pp. 1–21, 2023, doi: https://doi.org/10.2196%2F42734.

[7] L. Liu and B. K. P. Woo, "Twitter as a mental health support system for students and professionals in the medical field," *JMIR Med. Educ.*, vol. 7, no. 1, 2021, doi: https://doi.org/10.2196/17598.

[8] A. Le Glaz, Y. Haralambous, D. Kim-Dufor, P. Lenca, R. Billot, T. Ryan *et al.*, "Machine learning and natural language processing in mental health: Systematic review," *J. Med. Internet Res.*, vol. 23, no. 5, 2021, doi: https://doi.org/10.2196/15708.

[9] 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. 4, pp. 309–316, 2020, doi: https://doi.org/10.22219/kinetik.v5i4.1120.

[10] N. Fatimah, I. Budi, A. B. Santoso, and P. K. Putra, "Analysis of Mental Health During the Covid-19 Pandemic in Indonesia using Twitter Data," *Proc. - 2021 8th Int. Conf. Adv. Informatics Concepts, Theory, Appl. ICAICTA 2021*, pp. 1–6, 2021, doi: https://doi.org/10.1109/ICAICTA53211.2021.9640265.

[11] K. Yan, D. Arisandi, and T. Tony, "ANALISIS SENTIMEN KOMENTAR NETIZEN TWITTER TERHADAP KESEHATAN MENTAL MASYARAKAT INDONESIA," *J. Ilmu Komput. dan Sist. Inf.*, vol. 10, no. 1, Mar. 2022, doi: https://doi.org/10.24912/jiksi.v10i1.17865.

[12] M. L. Wicaksono, R. Rusdah, and D. Apriana, "Sentiment Analysis Of Mental Health Using K-Nearest Neighbors On Social Media Twitter," *Bit (Fakultas Teknol. Inf. Univ. Budi Luhur)*, vol. 19, no. 2, p. 98, 2022, doi: http://dx.doi.org/10.36080/bit.v19i2.2042.

[13] B. Nurfadhila and A. S. Girsang, "Identifying Indication of Depression of Twitter User in Indonesia Using Text Mining," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 2, pp. 523–530, 2023, [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/2663

[14] A. Pavlova and P. Berkers, "'Mental Health' as Defined by Twitter: Frames, Emotions, Stigma," *Health Commun.*, vol. 37, no. 5, pp. 637–647, 2022, doi: https://doi.org/10.1080/10410236.2020.1862396.

[15] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," *Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018*, pp. 226–229, 2019, doi: https://doi.org/10.1109/IALP.2018.8629151.

[16] 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: http://dx.doi.org/10.18653/v1/2020.coling-main.66.

[17] F. Koto, J. H. Lau, and T. Baldwin, "INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," *EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 10660–10668, 2021, doi: http://dx.doi.org/10.18653/v1/2021.emnlp-main.833.

[18] OpenAI, "GPT-4 Technical Report," vol. 4, pp. 1–100, 2023, doi: https://doi.org/10.48550/arXiv.2303.08774.

[19] R. K. Gupta, R. Agarwalla, B. H. Naik, J. R. Evuri, A. Thapa, and T. D. Singh, "Prediction of research trends using LDA based topic modeling," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 298–304, 2022, doi: https://doi.org/10.1016/j.gltp.2022.03.015.

[20] B. Ogunleye, T. Maswera, L. Hirsch, J. Gaudoin, and T. Brunsdon, "Comparison of Topic Modelling Approaches in the Banking Context," *Appl. Sci. 2023, Vol. 13, Page 797*, vol. 13, no. 2, p. 797, Jan. 2023, doi: https://doi.org/10.3390/app13020797.

[21] S. Koltcov, V. Ignatenko, M. Terpilovskii, and P. Rosso, "Analysis and tuning of hierarchical topic models based on Renyi entropy approach," *PeerJ Comput. Sci.*, vol. 7, pp. 1–35, 2021, doi: https://doi.org/10.7717/peerj-cs.608/supp-1.

[22] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," pp. 1–17, 2020, doi: https://doi.org/10.48550/arXiv.2008.05756.

[23] A. Nugaliyadde, K. W. Wong, F. Sohel, and H. Xie, "Enhancing Semantic Word Representations by Embedding Deeper Word Relationships," 2019, doi: https://doi.org/10.48550/arXiv.1901.07176.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017, doi: https://doi.org/10.48550/arXiv.1706.03762.

[25] M. M. Taamneh, S. Taamneh, A. H. Alomari, and M. Abuaddous, "Analyzing the Effectiveness of Imbalanced Data Handling Techniques in Predicting Driver Phone Use,"