Indonesian Journal of Innovation Studies
Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies

Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

Table Of Contents

Journal Cover 1
Author[s] Statement 3
Editorial Team 4
Article information 5
- Check this article update (crossmark) 5
- Check this article impact 5
- Cite this article 5

Title page 6
- Article Title 6
- Author information 6
- Abstract 6

Article content 7

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Copyright © Author(s). This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies

Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

EDITORIAL TEAM

Editor in Chief

Dr. Hindarto, Universitas Muhammadiyah Sidoarjo, Indonesia

Managing Editor

Mochammad Tanzil Multazam, Universitas Muhammadiyah Sidoarjo, Indonesia

Editors

Fika Megawati, Universitas Muhammadiyah Sidoarjo, Indonesia

Mahardika Darmawan Kusuma Wardana, Universitas Muhammadiyah Sidoarjo, Indonesia

Wiwit Wahyu Wijayanti, Universitas Muhammadiyah Sidoarjo, Indonesia

Farkhod Abdurakhmonov, Silk Road International Tourism University, Uzbekistan

Bobur Sobirov, Samarkand Institute of Economics and Service, Uzbekistan

Evi Rinata, Universitas Muhammadiyah Sidoarjo, Indonesia

M Faisal Amir, Universitas Muhammadiyah Sidoarjo, Indonesia

Dr. Hana Catur Wahyuni, Universitas Muhammadiyah Sidoarjo, Indonesia

Complete list of editorial team (link)

Complete list of indexing services for this journal (link)

How to submit to this journal (link)

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

Article information

Check this article update (crossmark)

Check this article impact (*)

Save this article to Mendeley

(*) Time for indexing process is various, depends on indexing database platform

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

Sentiment Analysis and Topic Modelling Using IndoBERTweet and BERTopic for Public Health Issues

Analisis Sentimen dan Pemodelan Topik Menggunakan IndoBERTweet dan BERTopic untuk Isu Kesehatan Publik

Wesly M Sihombing, wesly.m.2205356@students.um.ac.id, (1)
Department of Electrical and Informatics Engineering, State University of Malang, Indonesia

Triyanna Widiyaningtyas, triyannaw.ft@um.ac.id, ()
Department of Electrical and Informatics Engineering, State University of Malang, Indonesia

(1) Corresponding author

Abstract

Background: Public health challenges in Indonesia continue to expand across areas such as mental health, chronic diseases, vaccination debates, and environmental issues. **Specific background:** The rapid use of platform X provides large-scale public discourse that can be analyzed to understand real-time health discussions. **Knowledge gap:** Limited studies integrate advanced sentiment and topic modeling tailored to informal Indonesian social media language. **Aims:** This study analyzes public health conversations on platform X using IndoBERTweet for sentiment classification and BERTopic for topic extraction. **Results:** From 6,740 processed tweets, neutral sentiment dominated public discussions, while topic modeling produced 44 themes, with mental well-being, vaccination debates, chronic disease concerns, and regional disease reports emerging as key issues. IndoBERTweet demonstrated reliable performance (Weighted F1-Score 0.7822), and BERTopic produced coherent and diverse topics. **Novelty:** This research combines IndoBERTweet and BERTopic to generate a contextual, adaptive, and real-time mapping of public health discourse in Indonesia. **Implications:** The findings support data-driven health policymaking, enabling authorities to monitor public perceptions, strengthen communication strategies, and design region-specific interventions.

Highlights

- Public conversations emphasize mental health and lifestyle-related issues.
- Topic modeling identifies diverse clusters, including vaccination debates and endemic diseases.
- Integrated sentiment—topic analysis enables real-time mapping of health discussions in Indonesia.

Keywords

Public Health, Social Media Analysis, IndoBERTweet, BERTopic, Sentiment Classification

Published date: 2025-11-25

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

I. Introduction

Public health in Indonesia has faced increasingly complex challenges in recent years. The problems are no longer limited to infectious diseases but have expanded to include non-communicable diseases, nutrition, mental health, and environmental factors. Based on data from the Ministry of Health, the prevalence of stunting among children under five still reached 21.6% in 2022, a figure higher than the WHO standard [1]. Stunting not only reflects nutritional deficiencies but also indicates the social and economic conditions th the community that are not yet optimal. Therefore, addressing health problems requires a multidisciplinary and cross sectoral approach to ensure sustainable solutions. In addition, non communicable diseases have increased significantly. The 2023 Indonesia Health Profile recorded a hypertension prevalence of 34.1% and diabetes mellitus of 10.9% among the adult population [2]. This condition is related to modern lifestyles such as lack of physical activity, consumption of foods high in sugar and salt, and social pressures in urban areas [3]. If this trend continues, Indonesia could face an increase in healthcare costs and a decline in the productivity of the working-age population nationwide.

Mental health issues have also become a global concern. The World Mental Health Report shows that more than one billion people worldwide experience mental disorders, with depression and anxiety being the most common [4]. In Indonesia, this issue is often not accurately recorded due to strong social stigma, making mapping and intervention difficult to implement [5]. Yet, mental health is an important factor in determining the quality of productive human resources in the future.

Environmental factors also worsen public health conditions. The Better Air, Better Indonesia report shows that $PM_{2.5}$ concentrations in major cities such as Jakarta are still far above the limits recommended by WHO [6]. This increases the risk of chronic respiratory diseases and cardiovascular disordes [7]. Therefore, the public health approach must be comprehensive, encompassing the dimensions of nutrion, non communicable diseases, mental health, and the environment.

The advancement of digital technology presents new opportunities in analyzing public health issues. Social media, particularly the social media platform X (Twitter), has become a public space where people express opinions, share experiences, and respond to health policies [8]. Based on the 2023 DataReportal report, the number of active users of platform X in Indonesia has reached millions, making it a potential data source for understanding public sentiment in real time [9]. Compared to conventional surveys, social media analysis offers speed and dynamism in capituring public perceptions.

Several previous studies have utilized data from social media C for social and health analysis. indoBERTweet, a pre trained language model developed specifically for data from the Indonesian-language X platform, has proven effective in sentiment classification and hate speech detection [10]. [11]. Meanwhile, the BERTopic method can extract dominant topics with high coherence from in depth analysis of public health issues. However, the characteristics of platform X, which are rich in informal language, abbreviations, and code mixing, remain a challenge in data processing [13].

IndoBERT weet was developed to handle informal language styles on social media, making it superior to general language models in understanding social contect [14]. On the other hand, BERTopic combines classical methods such as LDA [15]. The combination of the two enables a more comprehensive and socially contextual ana;ysis of public health trends in Indonesia,

Based on this background, this study aims to analyze public health trends in Indonesia using data from social media X through the fine tuning approach of IndoBERTweet and the BERTopic method. The research results are expected to provide insights for policymakers, academics, and health practitioners, as well as enrich Natural Language Processing (NLP) studies in the context of social media-based public health. The novelty value of this research lies in the integration of two advanced approaches, IndoBERTweet and BERTopic, to produce a contextual, adaptive, and real-time data-based mapping of public health trends.

Study of Literature

1. Text Mining

Text mining is the process of discovering patterns or information from large amounts of textual data, whether unstructured or semi-structured [16]. Text mining is a variation of data mining and is commonly used for text categorization, text clustering, and sentiment analysis [17]. The stages in the text mining process include:

a. Text Processing

The initial stage includes case folding, tokenization, stemming, and tagging.

b. Feature Selection

Selecting relevant features, including stopword removal to eliminate unimportant words.

c. Text Representation

Transforming textual data into a more processable form, for example using vector representation.

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo
Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

d. Application of Text Mining Techniques

Applying specific methods to extract information from the processed textual data.

2. Sentiment Analysis

Sentiment analysis is the process of identifying public opinions, judgments, attitudes, and emotions toward a specific object, such as policies, public services, or health issues, which are often discussed on social media [18]. As part of opinion mining, sentiment analysis is defined as an automated process to extract and process textual data to obtain sentiment information contained in opinions, which are classified as positive, negative, or neutral. One of the tools used in sentiment analysis is IndoBERTweet. IndoBERTweet is a pre-trained language model specifically designed to handle the characteristics of informal language, abbreviations, and code-mixing that frequently appear on social media [19]. This model is adapted to the context of sentiment analysis on specific issues to produce accurate sentiment classification.

3. Topic Modelling

Topic modeling is a text analysis process that uses statistical methods to uncover hidden themes and relationships among them [20]. Since it does not require prior data or document labeling, topic modeling is classified as unsupervised learning. Its main function is to help organize, search, and summarize large collections of unstructured documents. One effective modern method is BERTopic, which combines embedding, clustering, and class-based TF-IDF techniques to extract topics; for instance, this method has been successfully applied to Indonesian-language tweets [21].

II. Method

This research was conducted using a data-driven approach to analyze public health trends in Indonesia using data from social media platform X. The study was carried out using Colab as the main development environment because it supports large-scale data processing, GPU acceleration, and integration with Python libraries for Natural Language Processing (NLP) and topic modeling [22], [23]. The research framework consists of several stages, namely Data Collection, Data Preprocessing, Data Labeling, fine-tuning the IndoBERTweet model, Topic Modeling (BERTopic), and Model Evaluation.

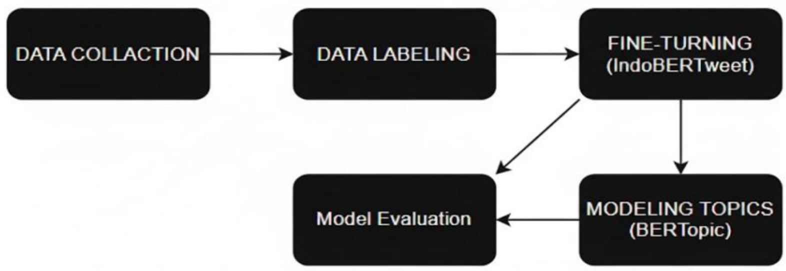

Figure 1. ***Research Framework***

1. Data Collection

The dataset used was obtained through the platform-X Application Programming Interface (API) for the period of January to July 2025 via data scraping, focusing on Indonesian-language tweets. The keywords used included "covid," "corona," "vaksin," "flu," "demam berdarah," "tuberkulosis," "obesitas," "mental health," and "kesehatan jiwa" The collected tweets were filtered to remove retweets and irrelevant content, then stored in CSV format for further analysis [24]. This method aligns with recent studies that utilize social media data to monitor health trends and public opinion.

2. Data Preprocessing

The preprocessing stage aims to clean and transform raw data into an analyzable form. This process includes case folding, text cleansing, tokenizing, normalization, and stopword removal. The implementation was carried out using Python libraries such as Sastrawi, Regex, and NLTK in Colab. The detailed stages are as follows:

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

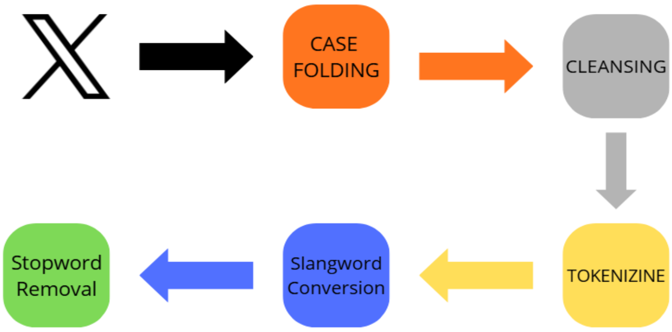

Figure 2. ***Data Preprocessing***

a. Case Folding — Converting all letters to lowercase for text consistency.

b. Cleansing — Removing URLs, mentions, hashtags, emojis, special characters, and duplicate tweets.

c. Tokenizing — Splitting sentences into words or tokens.

d. Normalization — Converting informal or slang words into standard words.

e. Stopword Removal — Removing less meaningful words using an Indonesian stopword list.

Raw data
pagi ini saya demam tdk jd perjalanan #rsj #fever#muterline @radioelshinta https://t.co/evjqkgf4pm
Cleansing result
pagi ini saya demam **tdk** jd perjalanan **#rsj #fever** @radioelshinta https://t.co/evjqkgf4pm
Cleansing result
pagi ini saya demam **tdk** jd perjalanan
Tokenizing result
'pagi', 'ini', 'saya', 'demam', 'tdk', 'jd', 'perjalanan'
Slangword conversion result
'pagi', 'ini', 'saya', 'demam', **'tidak'**, **'jadi'**, 'perjalanan'
Stopword removal result
'pagi', 'demam', 'tidak', 'jadi', 'perjalanan'

Table 1. *Data Processing Example*

3. Data Labeling

The data labeling stage was carried out to assign sentiment labels (positive, negative, and neutral) that serve as the foundation for the fine-tuning process of the IndoBERTweet model. This process fully utilized the w11wo/indonesian-roberta-base-sentiment-classifier model, which has been trained to analyze sentiment in Indonesian-language social media texts. The model automatically assigns a label to each tweet based on the sentence context and the emotional expressions contained within it. This automatic labeling approach was chosen because it accelerates the annotation process while providing consistent results for large-scale data. Furthermore, the w11wo model is known to be effective in understanding informal language characteristics, abbreviations, and code-mixing commonly used by Indonesian social media users [25].

4. Fine-Tuning (IndoBERTweet)

After preprocessing, the next step is fine-tuning the IndoBERTweet model, a transformer model pre-trained on Indonesian-language texts from platform X. Fine-tuning is performed to adapt the model to the context of sentiment analysis related to public health issues. This process includes dataset splitting (train-test), tokenization using the IndoBERTweet tokenizer, model training, and evaluation [26]. This model was chosen because it demonstrates superior performance in understanding informal language, abbreviations, and code-mixing that frequently appear on social media.

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

5. Modelling Topics (BERTopic)

The next stage is topic modeling using BERTopic, which combines transformer embeddings, dimensionality reduction (UMAP), and clustering (HDBSCAN) to identify dominant topics from tweets. The preprocessed tweets are transformed into embeddings using IndoBERTweet, then processed with BERTopic to extract public health themes and trends. The resulting topics are visualized using Plotly and Matplotlib to facilitate the analysis of public discussion patterns.

6. Model Evaluation

To ensure the reliability of the developed model, a comprehensive evaluation was conducted, divided into two parts: the evaluation of the sentiment classification model (IndoBERTweet) and the evaluation of the topic modeling quality (BERTopic).

a. Sentiment Classification Evaluation (IndoBERTweet)

The performance of the sentiment model was measured using standard classification metrics derived from the Confusion Matrix (TP, TN, FP, FN). These metrics were reported as Weighted Average values for multi-class classification (Positive, Negative, Neutral) [27]:

$$Accuracy = \frac{TP + TN}{TP + TN + TP + FP} \quad (1)$$

Accuracy aims to measure the ratio of total correct predictions (both True Positives and True Negatives) to the total number of tested data.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Precision aims to measure the model's reliability in predicting a specific class, which is the ratio between correct predictions for that class and the total predictions made for that class.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Recall aims to measure the model's ability to identify all actual cases belonging to a specific class, which is the ratio of correctly captured cases to the total actual cases of that class.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

F1-Score aims to measure the balance and combined performance between Precision and Recall, calculated as the harmonic mean of the two metrics, which is highly relevant for imbalanced datasets.

b. Topic Modeling Quality Evaluation (BERTopic)

The quality of topics is measured using standard intrinsic metrics to assess topic interpretability and distinctiveness [28]:

$$Coherence\ Score = C_v \quad (5)$$

Coherence Score aims to measure the semantic clarity of keywords within a topic. This metric evaluates how closely related the words are within the same topic. A high Coherence Score indicates that the topics generated by the model are easy to interpret.

$$Diversity\ Score = \frac{The\ number\ of\ unique\ words\ in\ the\ top\ N\ words\ of\ all\ topics.}{The\ total\ number\ of\ words\ in\ the\ top\ N\ words\ of\ all\ topics} \quad (6)$$

Diversity Score aims to measure the uniqueness and distinctiveness between topics. This metric is calculated as the ratio of the number of unique words across all top topics to the total number of words (including duplicates) across all topics. A high Diversity Score (close to 1.0) indicates that the topics generated by the model are diverse and not merely repetitions of the same keywords.

III. Result and Discussion

1. Dataset Description and Preprocessing Results

The results of public conversation data collection from platform X related to public health issues, conducted between January and July 2025, show that a total of 8,131 tweets were initially obtained. After undergoing the preprocessing steps described in

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

the previous section (Data Preprocessing and Data Labeling), 6,740 tweets were deemed valid and used as the final dataset for modeling. The sentiment labeling distribution of these 6,740 data points is summarized in Figure 3.

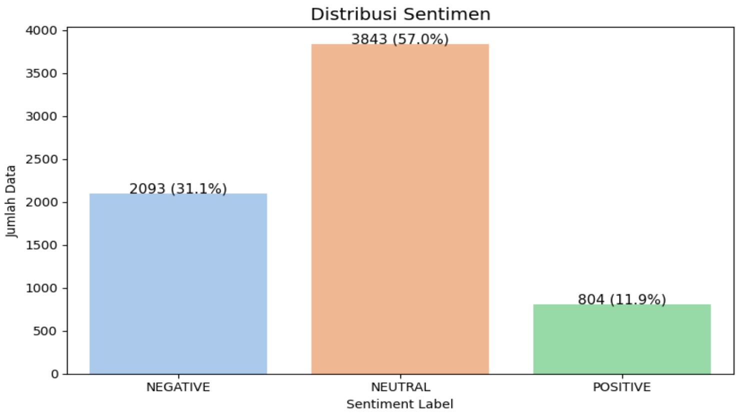

Figure 3. ***Sentiment Label Distribution***

Based on Figure 3, the NEUTRAL sentiment class is the most dominant, accounting for 57.02% (3,843 tweets) of the entire dataset. This dominance implies that most public interactions on X regarding health issues tend to focus on sharing factual information and asking questions. The NEGATIVE sentiment class holds the second-largest portion, at 31.05% (2,093 tweets), indicating that about one-third of the discussions carry significant tones of criticism, disappointment, or public concern. Meanwhile, the POSITIVE sentiment represents the smallest portion, only 11.93% (804 tweets), suggesting that explicit expressions of support or appreciation are relatively rare. The imbalanced distribution, particularly the dominance of Neutral sentiment, is an important factor considered during the fine-tuning process of the IndoBERTweet sentiment classification model.

2. Sentiment Classification Evaluation Results (IndoBERTweet)

Quantitative evaluation of the IndoBERTweet model, which has been fine-tuned in the context of public health sentiment issues, demonstrates strong predictive performance on the testing data. The evaluation metrics presented in Table 2 use a Weighted A verage approach to ensure objective measurement, considering the imbalanced distribution of sentiment classes as discussed in (Dataset Description and Preprocessing Results). The model achieved an Accuracy of 0.7574 (75.74%) and a Weighted F1- Score of 0.7822. the higher Weighted Precision (0.7980) compared to Weighted Recall (0.7782) indicates that the model tend to produce accurate positive predictions. Overall, these results confirm that IndoBERTweet is effective in understanding the informal language characteristics and code-mixing commonly used on the Twitter platform.

Metrix	Nilai
Accuracy	0.7574
Weighted Precision	0.7990
Weighted Recall	0.7574
Weighted F1-Score	0.7574

Table 2. *IndoBERTweet Model Perfomance Metrics (Weighted Average)*

The per-class perfomances analysis presented in Table 3 and visualized thorugh the Confusion Matrix in Figure 4 shows significant performance differences across classes. The Neutral class achieved the highest performance with an F1-Score of 0.823, Precision of 0.897, and Recall of 0/760, supported by the largest number of samples (3,437 support). This indicates that the model is highly capable of identifying the dominant neutral expressions in the data.

Kelas	Precision	Recall	F1-Score	Support
NEGATIVE	7,069	8,520	7,727	419
NEUTRAL	8,971	7,604	8,231	768

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo
Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Kelas	Precision	Recall	F1-Score	Support
POSITIVE	5,625	6,708	6,119	161
Macro Avg	7,222	7,611	7,359	1348
Weighted Avg	7,980	7,782	7,822	1348
Accuracy	-	-	7,574	1348

Table 3. *IndoBERTweet Model Performance Classification*

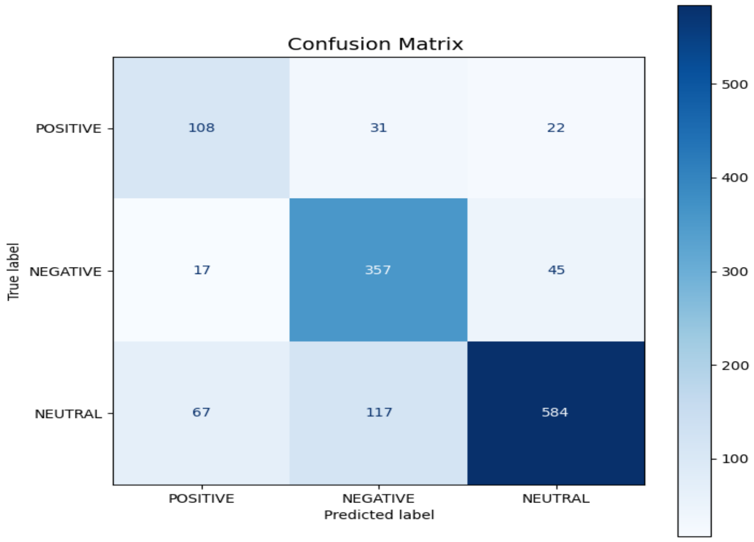

Figure 4. ***Confusion Matrix***

Conversaly, minority classes such as NEGATIVE and POSITIVE experienced performance drops due to limited sample sizes complexity of emotional expressions. tn the NEGATIVE class, a high Recall value (0.852) indicates the model's ability to capture most negative tweets. However, the ower Precision (0.707) suggest that the model tends to overpredict negatives, leading to some incorrect classifications. The challenges is more evident in the POSITIVE class, which has only 161 samples. This class recorded the lowest F1 score if 0.457, with Precision of 0.563 and Recall of 0.671, showing the model's difficulty in recognizing positive sebtiments that re often implicit or expressed indirectly.

The Confusion Matrix In Figure 4 reveals the most dominant prediction error pattern the model's tendency to classify extreme sentiments (NEGATIVE and POSITIVE) as NEUTRAL. A total of 275 negative tweets and 131 positive tweets were misclassified as neutral. This is attributed to linguistic ambiguity, sarcasm, or passive phrasing that obscures emotions. Nevertheless, the rate of misclassification between opposite polarities (positive predicted as negative or vice versa) is very low, indicating that the model can fundamentally distinguish sentiment direction. Overall, these evaluation results show that IndoBERTweet has adequate and reliable performance for further analysis.

3. Results Topic & Evaluation Results Modeling(BERTopic)

Topic modeling using BERTopic successfully extracted 44 topics from the entire social media X dataset analyzed, reflecting the diversity of public health issues discussed by the public. Of this number, the analysis focused on the nine most dominant topics (ID topic 1 - topic 9) because these topics had the highest frequency contributions to the overall discussion and were considered the most representative in describing public perceptions. The model quality validation is shown in Table 4, where the Coherence Score (C_v) is recorded at 0.7154. This value indicates that the words forming the topics have strong internal semantic relationships so that the generated topics are coherent and relevant to the context of public health.

Metrik	Nilai
Coherence Score (Cv)	7,154
Diversity Score	7,139

Table 4. *Topic Modeling Quality Evaluation Results (BERTopic)*

In addition, the Diversity Score of 0.7139 shows that the formed topics have clear differences in keywords from one another.

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo
Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Indonesian Journal of Innovation Studies
Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

This validates that the BERTopic model not only produces internally coherent topics but is also able to capture different and unique variations of discussion themes. The visualization of the frequency of the nine dominant topics in Figure 5 shows that Mental and Psychological Health is the most prominent topic, reflecting the high public attention to aspects of well-being and mental health as an integral part of public health

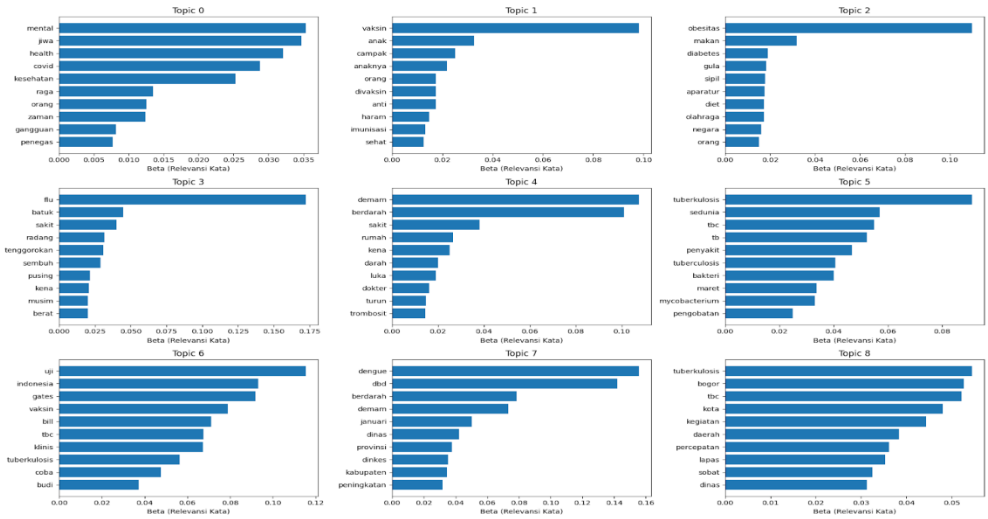

Figure 5. ***Visualization Of Topic Modelling Result***

Topic #1 shows that the most discussed health issue is centered on the category of Mental Health and Lifestyle. Where Topic #1 (Mental and Psychological Health) with a total of 2,152 tweets becomes the most dominant theme and is marked by keywords such as mental, jiwa, and health. The dominance of this topic shows an increase in public awareness of psychological and emotional health. This category is reinforced by Topic #3 (Obesity and Sugar Prevention) which focuses on chronic disease and unhealthy lifestyle issues through keywords such as obesity, eating, and diabetes, indicating public concern about consumption patterns and long-term health risks.

In addition, Topic #4 (Acute Respiratory Tract Diseases) describes public concern about common diseases related to the respiratory system, such as flu, cough, and inflammation, which are often associated with weather changes and decreased body immunity. Meanwhile, Topic #2 (Child Vaccination Controversy) with 805 tweets shows a sensitive issue that triggers divisions of public opinion, marked by keywords such as haram and immunization which indicate religious debates and public Trust in government health policies.

The issue of Dengue Hemorrhagic Fever (DBF) appears in two discussion clusters, namely Topic #5 (General Clinical DHF Disease) and Topics #8 (Regional DHF reports) with 97 tweets. Topic #5 reflect general discussions and initial handling actions of DHF, while Topics #8 leads to more specific and localized discussions, such as case reports by district health offices. This topics separation confirms the ability of BERTopic to distinguish between general clinical issues and regional-scale public health issues.

The issue of Tuberculosis (TBC) also appears at various levels of discussion. Topic #6 highlights education and general knowledge about TBC, while Topic #7 (Vaccine Research and Clinical Trials) with 131 tweets raises global discourse related to the involvement of figures such as Bill Gates in the development of the TBC vaccine. Furthermore, Topic #9 (Local TBC Intervention in Bogor City) with 91 tweets raises discussions about TBC management at the local level, including health interventions in the Bogor prison area.

Overall, the topic modeling results show that public conversations cover various dimensions of health, ranging from global policy issues, regional program implementation, endemic disease spread, to basic community needs related to well-being and lifestyle. These findings provide comprehensive insights into the health issues that are the main concern of the public and have the potential to become strategic input in formulating public health policies and interventions that are more responsive and targeted.

4. Results Topic & Evaluation Results Modeling(BERTopic)

The integration of BERTopic and IndoBERTweet sentiment analysis produced strategic insights indicating that mental health issues have become the public's top priority. The dominance of Topic #1 (Mental and Psychological Health) emphasizes the importance for the government to expand access to psychological services, enhance well-being literacy, and utilize social media as a Sustainable channel for emotional support. In addition, the findings of topic #2 (Child Vaccination Controversy) indicates significant social polarization, highlighting the need for collaboration between health authorities and religious leaders to develop evidence based communication narratives aimed at reducing oublic misunderstanding and increasing trust in immunization programs.

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo
Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Furthermore, local topics such as Topic #9 (TBC Intervention in Bogor City) and Topic #8 (Regional DHF Reports) show that public health issues are heterogeneous and require place based intervention tailored to the specific context of each region. Regional based interventions and the trengthening of endemic disease surveillance at the district/ city level become key strategies to enhance the effectiveness of health policies. Overall, the utilization of topic and sentiments modeling provides evidence based, actionable feedback, enabling stakeholders to design health policies that are more adaptive proactive, and responsive to the real needs of society.

Overall, the integration of sentiment analysis and topic modeling in this study provides a comprehensive overview ot the dynamics of the public perceptions of health issues. The novelty of this research lies in the application of integrated analytical approach that combines sentiment and topic analysis simultaneously to generate a more in depth, contextual, and data driven understanding of public health issues in Indonesia.

Furthermore, this research contributes to the development of Natural Languange Processing (NLP) methods, particularly in the Indonesian context, by demonstrating the effectiveness of the IndoBERTweet model in recognizing informal language styles and code mixing on social media, as well as the ability of BERTopic to group public discourse semantically and thematically. The applied hybrid approach can be used as a new methodological framework for the analysis of dynamic, unstructured, and contextual Indonesian texts, thus opening up opportunities for the development of more applicable NLP research in the fields of health, social sciences, and public policy.

IV. Conclusion

This study successfully identified public health trends in Indonesia through sentiment analysis using IndoBERTweet and topic modeling with BERTopic on social media X data. The model evaluation demonstrated reliable performance with a Weighted F1-Score of 0.7822, indicating the model's effectiveness in understanding informal language and code-mixing commonly used on digital platforms. Topic modeling generated 44 topics, with 9 main topics representing crucial issues such as mental health, child vaccination controversy, chronic diases, and region based policy discussions. These findings demonstrate that a communicative approach can provides a comprehensive representation of public opinion in real time.

However, this study has several limitations that should be noted. Keywords based data collcection may cause selection bias and might not fully capture all converstation related to health issues. Additionally, centiment analysis was limited to three categories, which does not account for emotional dimensions such as fear, anger, or anxietly that are important for understanding public reactions. The absence of geolocation data also limits the ability to analyze issue distribution across administrative regions.

For future research, it is recommended to integrate emotion classification to produce deeper psychological analysis and utilize geospatial metadata to support region-based health issue mapping. The topic modeling approach can also be extended to dynamic topic modeling to capture the evolution of public discourse over time. With these developments, future studies are expected to provide more precise, adaptive, and responsive health policy recommendations that align with societal needs

Practically, the results of this study have important implications for the formulation of data-driven public health policies. The resulting insights can be used by governments and health institutions to monitor public perceptions in real time, identify issues requiring rapid response, and design communication strategies and policy interventions that are more adaptive, participatory, and oriented toward community needs.

ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to all parties who contributed to the completion of this research, especially to the academic supervisors for their guidance, advice, and valuable insights throughout the research process. The author also extends appreciation to everyone who assisted in the collection and processing of data from social media platform X. It is hoped that the results of this study will provide positive contributions to the development of technology-based public health data analysis in Indonesia.

References

1. Kementerian Kesehatan Republik Indonesia, Buku Saku Hasil Studi Status Gizi Indonesia (SSGI) 2022. Jakarta: Kementerian Kesehatan RI, 2022.
2. Kementerian Kesehatan Republik Indonesia, Profil Kesehatan Indonesia 2023. Jakarta: Kementerian Kesehatan RI, 2023.
3. A. Prasetyo and S. Nurjanah, "Hubungan Gaya Hidup Modern dengan Risiko Penyakit Tidak Menular di Indonesia," Jurnal Kesehatan Masyarakat Nasional, vol. 17, no. 1, pp. 25–34, 2023.
4. World Health Organization, World Mental Health Report: Transforming Mental Health for All. Geneva: WHO, 2022.
5. A. Ningsih and R. Wardani, "Stigma Sosial dan Penanganan Gangguan Mental di Indonesia," Jurnal Psikologi Sosial dan Budaya, vol. 9, no. 2, pp. 115–124, 2023.
6. Systemiq, Better Air, Better Indonesia. London: Systemiq Earth, 2025.
7. L. Suryani and T. Hartono, "Dampak Pencemaran Udara terhadap Kesehatan Pernapasan di Kota Besar," Jurnal Lingkungan dan Kesehatan, vol. 11, no. 2, pp. 66–75, 2024.
8. DataReportal, Digital 2023: Indonesia. [Online]. Available: https://datareportal.com/reports/digital-2023-indonesia. [Accessed: May 2024].

Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Vol. 26 No. 4 (2025): October
DOI: 10.21070/ijins.v26i4.1833

9. R. Hidayat and D. Putra, "Pemanfaatan Media Sosial untuk Analisis Isu Publik di Indonesia," Jurnal Komunikasi Digital, vol. 6, no. 1, pp. 12–21, 2023.
10. J. F. Kusuma and A. Chowanda, "Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter," International Journal on Informatics Visualization (JOIV), vol. 7, no. 1, pp. 45–52, 2023.
11. F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in Proc. EMNLP 2021, pp. 1057–1070, 2021.
12. N. N. Hidayati and S. Shaleha, "BERTopic Analysis of Indonesian Biodiversity Policy on Social Media," ECTI Transactions on Computer and Information Technology, vol. 18, no. 3, pp. 244–252, 2024.
13. A. Rahmawati and D. Sari, "Tantangan Analisis Bahasa Informal di Media Sosial Menggunakan NLP," Jurnal Teknologi Informasi dan Komunikasi, vol. 10, no. 1, pp. 19–27, 2023.
14. I. Setiawan and T. Gunawan, "Penerapan IndoBERTweet untuk Analisis Sentimen pada Isu Sosial," Jurnal Data Sains Indonesia, vol. 4, no. 2, pp. 88–97, 2023.
15. M. Grootendorst, "BERTopic: Neural Topic Modeling with Contextual Embeddings," arXiv preprint arXiv:2203.05794, 2022.
16. [16] Feldman, R., Techniques and Applications for Sentiment Analysis, Communications of the ACM, 2020.
17. Mining, Text. Foundations and Applications, Springer, 2021.
18. A. Prasetyo, R. Hidayat, dan D. Putra, "Analisis Sentimen Media Sosial untuk Pemantauan Isu Kesehatan Masyarakat di Indonesia," Jurnal Teknologi Informasi dan Kesehatan Digital, vol. 5, no. 2, pp. 44–53, 2023.
19. Koto, F., et al., "IndoBERTweet: Pretrained Language Model for Indonesian Twitter," EMNLP 2021.
20. Blei, D. M., Ng, A. Y., & Jordan, M. I., "Latent Dirichlet Allocation," Journal of Machine Learning Research, 2021.
21. Grootendorst, M., "BERTopic: Neural Topic Modeling with Contextual Embeddings," arXiv:2203.05794, 2022.
22. M. A. Oktaviani dan R. Suryana, "Implementasi Teknik NLP untuk Analisis Sentimen Menggunakan Google Colab," Journal of Data Science Research, vol. 5, no. 2, pp. 45–53, 2022.
23. T. Nurzaman et al., "Eksplorasi Model Deep Learning untuk Analisis Teks di Platform Cloud," Indonesian Journal of Artificial Intelligence, vol. 7, no. 1, pp. 22–30, 2023.
24. P. Wibowo dan A. N. Lestari, "Monitoring Tren Kesehatan Publik melalui Data Twitter di Indonesia," Procedia Computer Science, vol. 213, pp. 550–558, 2022.
25. w11wo,"Indonesian-RoBERTa-Base-Sentiment-Classifier," Hugging Face, May 12, 2023. [Online]. Available: https://huggingface.co/w11wo/indonesian-roberta-base-sentiment-classifier
26. C. Wibisono et al., "Fine-Tuning IndoBERTweet untuk Analisis Sentimen Bahasa Indonesia," Applied Computing and Informatics, vol. 20, no. 1, pp. 25–34, 2024.
27. Sokolova, M., & Lapalme, G., "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 57, no. 1, pp. 102–129, 2020.
28. Purwarianti, A., & Crisdayanti, I., "Improving Indonesian Sentiment Analysis using Multilingual BERT," Proceedings of the International Conference on Asian Language Processing (IALP), 2019.

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo
Copyright © Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).