Moestopo International Review on Societies, Humanities, and Sciences (MIRSHuS)
Vol. 6, No. 1 (2026) pp. 114-131
ISSN: 2775-9601

Public Perceptions of HPV Vaccination Through Transformer-Based Social Media Sentiment Analysis

Desi Elfrida Silaban*
Universitas Bina Nusantara, Jakarta, Indonesia
Tuga Mauritsius
Universitas Bina Nusantara, Jakarta, Indonesia

*Correspondence: desi.silaban@binus.ac.id

ARTICLE INFO

Article History:
received: 02/02/2026
revised: 11/03/2026
accepted: 01/04/2026

Keywords:
Sentiment Analysis; Human Papillomavirus Vaccine; Social Media Discourse; Transformer Model; Public Health Communication

DOI:
10.32509/mirshus.v6i1.175

ABSTRACT

Public perception plays a crucial role in determining the success of vaccination programs, particularly for the human papillomavirus vaccine aimed at preventing cervical cancer. Despite the increasing implementation of vaccination initiatives, public opinions expressed in digital environments may influence the acceptance and effectiveness of such programs. This study aims to examine public sentiment toward the human papillomavirus vaccine by analyzing discussions on a social media platform widely used for public communication. A data mining framework was employed to guide the analytical process, including data collection, preprocessing, sentiment classification, and thematic exploration. Transformer-based language models were utilized to classify public sentiment expressed in social media posts, followed by topic modeling to identify key issues discussed by users. The findings reveal that public discourse is largely characterized by supportive attitudes toward vaccination, reflecting a growing awareness of its role in cervical cancer prevention. Nevertheless, several concerns related to vaccine cost, accessibility, and post-vaccination experiences continue to emerge in online discussions. These results highlight the importance of integrating digital discourse analysis into public health communication strategies in order to better understand societal perspectives and improve the effectiveness of vaccination programs.

INTRODUCTION

Human papillomavirus (HPV) infection is widely recognized as the primary cause of cervical cancer among women worldwide. Cervical cancer remains a major global public health concern, particularly in developing countries where access to early screening and preventive healthcare services remains limited. According to global health statistics, cervical cancer ranks as the fourth most common cancer affecting women worldwide (Mascarenhas et al., 2024). In Indonesia, the burden of the disease is particularly significant, as cervical cancer is the second

most frequently diagnosed cancer among women. Data reported by the Global Cancer Observatory (GLOBOCAN) indicate that Indonesia recorded 36,964 new cases of cervical cancer and 20,708 related deaths in 2022, reflecting the substantial public health impact of HPV infection. Despite its high mortality rate, cervical cancer is largely preventable through early preventive interventions, among which HPV vaccination has been widely recognized as one of the most effective strategies for reducing infection rates and preventing cervical cancer. Empirical evidence demonstrates that HPV vaccination provides strong protection against HPV-related diseases and significantly reduces the risk of cervical cancer, particularly when administered at younger ages (Ellingson et al., 2023). Furthermore, large-scale population studies confirm that prophylactic HPV vaccination offers substantial protection against invasive cervical cancer among vaccinated populations (Arbyn et al., 2024).

Recognizing the urgency of addressing this public health challenge, the Indonesian government has implemented several strategic initiatives to strengthen cervical cancer prevention programs. In 2023, the Ministry of Health introduced the National Action Plan (Rencana Aksi Nasional/RAN) for Cervical Cancer Elimination for the period 2023–2030, aiming to accelerate preventive interventions and improve vaccination coverage nationwide. As part of this strategy, a nationwide HPV vaccination program was launched in August 2023. The program is implemented in two phases to ensure broader population coverage. In the first phase, the government targets 90% vaccination coverage among girls aged 11–12 years, particularly those enrolled in elementary school or equivalent levels of education. In the second phase, the vaccination program is expanded to include both girls and boys aged 11–12 years with the same vaccination coverage target. This policy reflects a long-term strategy to strengthen population immunity and reduce the transmission of HPV infection. Expanding vaccination programs and increasing public awareness are therefore essential strategies for improving HPV vaccine uptake and strengthening national cervical cancer prevention efforts (Furuno et al., 2024; Sendekie et al., 2025).

Despite increasing awareness regarding the importance of HPV vaccination, challenges related to public perception and vaccine acceptance remain significant. In many communities, misconceptions, misinformation, and concerns regarding vaccine safety continue to influence public attitudes toward vaccination programs. As a result, some parents remain hesitant to allow their children to receive the HPV vaccine. Public perception and societal attitudes therefore play a critical role in determining the success of national vaccination initiatives. Understanding how individuals perceive and discuss HPV vaccination is essential for developing effective health communication strategies and strengthening public trust in vaccination programs. Previous studies indicate that vaccine hesitancy is often driven by misinformation, safety concerns, and insufficient knowledge regarding HPV and its vaccine, which can significantly reduce vaccination uptake (Sendekie et al., 2025). Similarly, parental concerns regarding vaccine safety and potential side effects have been identified as key determinants influencing the acceptance of HPV vaccination among children (Heyde et al., 2024). These findings highlight the importance of understanding public perceptions in order to design more effective communication strategies that address misinformation and improve vaccine acceptance.

Alongside the rapid expansion of digital communication technologies, social media platforms have emerged as influential spaces where individuals express opinions, exchange information, and discuss public health issues. Platforms such as X (formerly Twitter) generate large volumes of user-generated textual data that reflect real-time public discourse on a wide range of topics, including vaccination policies and health interventions. These digital conversations provide valuable opportunities to analyze public sentiment and identify emerging concerns within society. Advances in Natural Language Processing (NLP) and machine learning have enabled researchers to analyze large-scale textual data from social media to better understand patterns of public opinion and sentiment. Social media analytics using computational approaches has increasingly been applied to examine public attitudes toward HPV vaccination and to detect misinformation as well as shifts in vaccine confidence within online discussions (Liu et al., 2025; Singh et al., 2025).

Several previous studies have applied sentiment analysis techniques to examine public attitudes toward vaccination programs using data derived from social media platforms. These studies demonstrate that computational approaches can effectively capture large-scale public opinion and analyze digital discourse surrounding health interventions (Aljabar et al., 2024; Das et al., 2025). However, much of the existing literature primarily focuses on COVID-19 vaccination and often relies on conventional machine learning models or single deep learning architectures to classify public opinions expressed in online discussions. While these approaches provide valuable insights into public sentiment, they may not fully capture the contextual and linguistic complexity inherent in social media communication, particularly when analyzing informal language and evolving digital narratives. Furthermore, research specifically examining public discourse related to HPV vaccination, especially within the Indonesian context, remains relatively limited. This limitation highlights an important research gap that calls for more context-aware analytical methods capable of modeling nuanced expressions of opinion in online environments. In addition, only a limited number of studies integrate sentiment analysis with topic modeling techniques to explore the underlying factors shaping positive and negative perceptions of vaccination programs. Previous research has shown that social media data can be analyzed using computational methods to understand vaccine-related attitudes and identify misinformation influencing public opinion regarding HPV vaccination (Kim et al., 2022; Singh et al., 2025). Similarly, recent studies employing data-driven analyses of online discussions highlight the importance of advanced analytical techniques for identifying patterns of vaccine hesitancy and shifts in public sentiment within digital communication environments (Yoon et al., 2024).

To address these research gaps, this study proposes a transformer-based sentiment analysis approach to examine public discourse related to HPV vaccination on social media platform X. Public opinions and discussions regarding the HPV vaccine are collected and analyzed using transformer-based language models, specifically Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT), to classify sentiments expressed in social media posts. Furthermore, the results of sentiment classification are analyzed using word cloud visualization and BERTopic-based topic modeling

to identify key themes and factors shaping positive and negative public perceptions regarding HPV vaccination.

This study introduces a novel analytical perspective by integrating transformer-based sentiment classification with BERTopic-based topic modeling to examine public discourse related to HPV vaccination on social media. While previous research has explored public attitudes toward vaccination using social media analytics, many of these studies focus primarily on COVID-19 vaccination or rely on conventional machine learning approaches that may not fully capture the contextual complexity of social media language. By employing advanced transformer architectures, this research enables deeper contextual understanding of sentiment expressed in Indonesian social media texts. In addition, the integration of sentiment classification with topic modeling allows the analysis not only to identify sentiment polarity but also to reveal the underlying themes that shape public perceptions regarding HPV vaccination.

The contributions of this study can be summarized in three main aspects. First, this research provides an empirical examination of public sentiment toward HPV vaccination based on large-scale social media discussions, offering valuable insights into how vaccination programs are perceived within Indonesia's digital public sphere. Second, the study conducts a comparative evaluation of transformer-based models, namely IndoBERT and GPT-2, for sentiment classification tasks in Indonesian social media texts, thereby highlighting the importance of language-specific pre-trained models in health-related sentiment analysis. Third, by integrating sentiment analysis with BERTopic-based topic modeling, this research identifies key thematic factors that shape both positive and negative public perceptions regarding HPV vaccination. These findings offer practical implications for policymakers, healthcare institutions, and public health communicators in developing more responsive communication strategies aimed at improving vaccine acceptance and strengthening public trust in vaccination programs. Therefore, this study aims to analyze public perceptions of HPV vaccination by applying transformer-based sentiment analysis to social media discussions. Specifically, the study compares the performance of IndoBERT and GPT-2 models for sentiment classification and employs BERTopic modeling to identify the major themes influencing positive and negative public perceptions of HPV vaccination.

METHOD

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework as the primary methodological foundation for conducting sentiment analysis on public discourse related to HPV vaccination. CRISP-DM provides a structured and systematic workflow that supports analytical rigor, methodological transparency, and reproducibility throughout the research process. The framework consists of six sequential phases: *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* and *Deployment*. Each stage guides the analytical process from problem identification to result interpretation, ensuring that the analysis is conducted in a comprehensive and methodologically sound manner. The CRISP-DM framework has been widely applied in data mining and analytics research because it offers a structured approach for data exploration,

model development, and performance evaluation across various analytical contexts (Lundén et al., 2023). Recent studies further emphasize that CRISP-DM remains highly relevant for modern data analytics projects because it enables the integration of diverse analytical techniques while maintaining transparency and reproducibility throughout the analytical pipeline (Acuña-Cid et al., 2025)

Figure 1 illustrates the overall research framework implemented in this study. The framework integrates multiple stages, including data collection, data preprocessing, transformer-based sentiment classification, and BERTopic-based topic modeling, to systematically analyze public discourse related to the HPV vaccine on social media platform X.

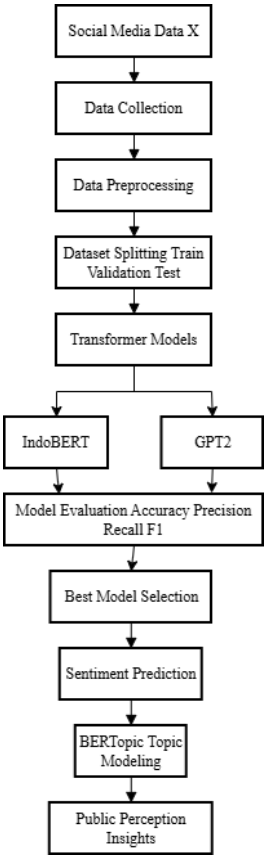

Figure 1. Research framework of transformer-based sentiment analysis for HPV vaccine discourse

Business Understanding

The Business Understanding phase focuses on defining the research objectives and clarifying the analytical problem addressed in this study. The primary objective of this research is to examine public perceptions of the HPV vaccine as expressed in discussions on social media platform X. Public opinion plays a significant role in shaping vaccination acceptance and influencing the effectiveness of national health policies. Understanding how individuals express support, concerns, or skepticism regarding HPV vaccination is therefore essential for strengthening public health communication strategies. Previous research demonstrates that public perceptions, attitudes, and trust in vaccination information sources significantly influence decisions related to HPV vaccine acceptance and uptake (Naoum et al., 2022).

In this stage, relevant literature related to HPV vaccination, public perception, and sentiment analysis methodologies is reviewed to establish a conceptual foundation for the study. Based on this literature review, the analytical objective is formulated as developing a transformer-based sentiment classification model capable of identifying positive, negative, and neutral sentiments expressed in user-generated social media content related to HPV vaccination.

Data Understanding

The Data Understanding phase involves collecting and exploring textual data obtained from public discussions on social media platform X. The dataset consists of user-generated posts containing opinions, comments, and personal experiences related to HPV vaccination and cervical cancer prevention. Social media platforms generate large volumes of publicly accessible textual data that reflect real-time public attitudes, discussions, and concerns regarding vaccination programs. Consequently, these platforms represent valuable sources of data for understanding public perceptions and behavioral tendencies toward health interventions (Singh et al., 2025).

Data collection in this study is conducted using the Advanced Search feature provided by the platform. This feature allows researchers to filter posts based on specific keywords, language preferences, and time ranges to ensure the relevance of the collected data. Only posts explicitly discussing HPV vaccination or related topics are included in the dataset. The collected data are then explored to understand their structure, linguistic characteristics, and overall distribution prior to further analysis. This stage ensures that the dataset accurately represents public discourse and provides a reliable foundation for subsequent sentiment analysis.

Data Preparation

The Data Preparation phase focuses on transforming raw textual data into a structured format suitable for transformer-based modeling. The dataset collected from social media is divided into training, validation, and testing subsets to enable robust model training and unbiased model evaluation.

Several preprocessing techniques are applied to improve data quality. First, text cleaning is performed to remove non-informative elements such as special characters, excessive punctuation, hyperlinks, and irrelevant symbols commonly found in social media text. Text preprocessing plays a critical role in sentiment analysis because raw social media data frequently contain noisy elements that can negatively affect model performance. Cleaning and normalization processes are therefore essential to improve the quality and consistency of textual data before further analysis (Papia et al., 2024).

Next, slang word normalization is conducted to convert informal or colloquial expressions into standardized Indonesian language forms. Social media users often express opinions using abbreviations or informal language, which may create inconsistencies in textual representation. By converting slang expressions into their formal equivalents, the dataset becomes more linguistically consistent and easier for machine learning models to interpret.

Previous studies emphasize that normalization of informal expressions significantly improves the accuracy of sentiment classification models in social media datasets (Saputra et al., 2025).

Finally**, tokenization** is applied to segment textual data into smaller linguistic units known as tokens. Transformer-based architectures rely on subword tokenization to capture contextual meaning at both word and subword levels. This process converts textual data into numerical representations that can be processed computationally by machine learning models. In natural language processing pipelines, tokenization is a fundamental step that enables models to capture contextual relationships between words more effectively (Dotan et al., 2024).

Modeling

The **Modeling** phase involves developing transformer-based sentiment classification models. In this study, two transformer architectures are implemented: IndoBERT and GPT-2. Transformer-based language models such as BERT and GPT have demonstrated strong performance in sentiment classification tasks because their attention mechanisms allow models to capture contextual relationships and semantic nuances within textual data more effectively than traditional machine learning approaches (Aljabar et al., 2024).

IndoBERT, a pre-trained language model optimized for Indonesian textual data, is fine-tuned for sentiment classification. The model processes preprocessed textual input through tokenization, contextual embedding generation, and a classification layer to predict sentiment categories. The overall workflow of the IndoBERT-based sentiment classification process is illustrated in Figure 2, which demonstrates the sequence of preprocessing, contextual embedding extraction, and sentiment prediction. Pre-trained transformer models such as BERT have been widely adopted for sentiment classification tasks because contextual embeddings enable the model to capture semantic relationships and contextual information within textual data more effectively (Talaat, 2023).

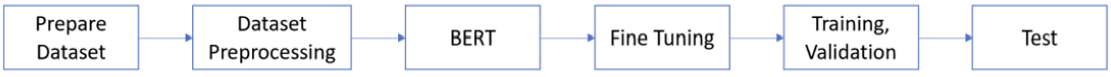

Figure 2. IndoBERT-based sentiment classification workflow

In addition to IndoBERT, a GPT-2-based transformer architecture is implemented for sentiment classification. GPT-2 utilizes an autoregressive modeling approach that predicts tokens sequentially based on previously generated tokens. The modeling pipeline for the GPT-2 sentiment classification process is presented in Figure 3, illustrating stages such as tokenization, contextual representation learning, and classification output. Previous research indicates that fine-tuned GPT-2 models can effectively perform sentiment classification by learning contextual representations from textual narratives and mapping them to sentiment polarity classes (Deng et al., 2025).

Moestopo International Review on Societies, Humanities, and Sciences (MIRSHuS)
Vol. 6, No. 1 (2026) pp. 114-131
ISSN: 2775-9601

Figure 3. GPT-2 sentiment classification workflow

Following the sentiment classification stage, topic modeling is performed to identify thematic structures within the classified sentiment data. This study employs BERTopic, a topic modeling technique that combines transformer embeddings with clustering mechanisms to generate interpretable topics from textual data. The BERTopic modeling process is illustrated in Figure 4, demonstrating how sentiment-labeled data are embedded, clustered, and transformed into coherent discussion themes. BERTopic leverages contextual embeddings generated by transformer models and integrates dimensionality reduction and clustering techniques to discover semantically meaningful topics within large textual (Ma et al., 2025).

Figure 4. BERTopic topic modeling process

Through this modeling framework, the study not only classifies sentiment polarity but also identifies the underlying themes influencing public discourse regarding HPV vaccination.

Evaluation

The Evaluation phase assesses the performance of the developed sentiment classification models. Model performance is evaluated using standard classification metrics, including Accuracy, Precision, Recall, and F1-Score. These metrics are widely used in sentiment analysis and text classification research to evaluate the effectiveness of machine learning and deep learning models in predicting sentiment categories (Abdulhakim Al-Absi et al., 2023).

Accuracy measures the overall correctness of predictions, while Precision and Recall evaluate the model's ability to correctly identify specific sentiment classes. The F1-Score provides a balanced evaluation metric by combining Precision and Recall, which is particularly useful when dealing with class imbalance. These evaluation metrics enable systematic comparison between transformer models to determine the most reliable architecture for analyzing HPV vaccine-related discourse. Previous studies commonly employ these metrics to compare sentiment classification performance and ensure reliable evaluation of classification outcomes (Das et al., 2025).

Deployment

The final phase, Deployment, involves applying the best-performing sentiment classification model to classify sentiments in previously unseen social media posts. The prediction results provide an overview of current public sentiment toward HPV vaccination and reveal emerging patterns in online discussions.

Furthermore, the sentiment classification results are integrated into the BERTopic modeling stage to generate thematic insights regarding public discourse. This integration enables deeper exploration of the factors influencing positive and negative public perceptions of HPV vaccination. Through this deployment process, the study produces actionable insights that may support policymakers, healthcare institutions, and public health communicators in designing more effective communication strategies to improve HPV vaccine acceptance.

Finally, to ensure methodological transparency and reproducibility, the analytical pipeline implemented in this research follows clearly defined preprocessing procedures, modeling configurations, and evaluation metrics. The use of established transformer architectures and standardized evaluation methods enables other researchers to replicate the analytical workflow and extend the findings to similar contexts in social media–based public health research.

RESULT AND DISCUSSION

Dataset Collection

The dataset used in this study was collected manually using the Advanced Search feature available on social media platform X. This feature allows researchers to retrieve posts based on specific keywords, language preferences, and time ranges. The data collection process focused on public discussions containing the keyword *"HPV vaccine"*, enabling the study to capture online conversations related to HPV vaccination from social media users. Previous studies indicate that data from platforms such as Twitter/X can be systematically collected using keyword-based queries or hashtags to analyze public discourse and online discussions related to health topics (Xue et al., 2023).

Data collection was conducted over a one-year period from June 2023 to June 2024 to observe variations in public opinion during the early implementation of HPV vaccination programs in Indonesia. After completing the data retrieval process, a total of 3,572 tweets were obtained and compiled as the initial dataset for analysis.

To support the training and evaluation of machine learning models, the dataset was divided into three subsets: training data, validation data, and testing data. The dataset partition followed a 60:20:20 ratio, a common strategy in machine learning experiments that ensures sufficient data for model learning while preserving independent datasets for validation and performance evaluation. Previous studies emphasize that dividing datasets into three subsets helps prevent overestimation of model performance and enables more reliable model evaluation during experimentation (Surya et al., 2023).

Table 1. Dataset Distribution

Dataset Type	Number of Data
Training Data	2143
Validation Data	714
Test Data	715

This distribution ensures that the model can learn effectively from the training dataset while maintaining independent validation and testing datasets to assess model generalization capability.

Data Preprocessing

Before the modeling stage, several preprocessing procedures were applied to improve the quality and consistency of the textual dataset. Social media data often contain informal expressions, abbreviations, emojis, hyperlinks, and special characters that introduce noise and reduce the effectiveness of machine learning models. Textual variations and non-standard language frequently produce out-of-vocabulary words and noisy inputs, making preprocessing an essential stage in natural language processing pipelines (Khan et al., 2025).

The preprocessing pipeline implemented in this study consisted of three main stages: text cleaning, slang word normalization, and tokenization.

1. First, text cleaning was performed to remove non-informative elements such as punctuation marks, hyperlinks, emojis, repeated characters, and special symbols commonly found in social media posts. Removing these elements helps reduce noise within the dataset and ensures that the textual input focuses on meaningful linguistic information. Previous research highlights that noisy elements in social media datasets can distort sentiment analysis results if not properly removed during preprocessing (Pencheva, 2025).
2. Second, slang word normalization was conducted to convert informal expressions commonly used in social media communication into standardized Indonesian vocabulary. Social media users frequently employ abbreviations or colloquial language that may create inconsistencies in textual representation. By normalizing slang expressions into their formal equivalents, the dataset becomes more linguistically consistent and easier for machine learning models to interpret.
3. Finally, tokenization was applied to segment textual data into smaller linguistic units known as tokens. Transformer-based architectures rely on subword tokenization to capture contextual meaning at both word and subword levels. These tokens are subsequently converted into numerical representations that allow models to process textual information computationally. Modern NLP models commonly utilize subword tokenization techniques such as Byte Pair Encoding or WordPiece to represent textual data while mitigating out-of-vocabulary problems (Zouhar et al., 2023).

Through these preprocessing steps, the dataset was transformed into a structured format suitable for transformer-based sentiment classification.

Experimental Results Using the IndoBERT Model

The first experiment in this study involved the implementation of IndoBERT, a transformer-based language model specifically designed for Indonesian language processing. IndoBERT is based on the Bidirectional Encoder Representations from Transformers (BERT) architecture, which captures contextual relationships between words using bidirectional attention mechanisms. This architecture enables the model to learn deep contextual

representations by analyzing both preceding and succeeding tokens within a sentence, thereby improving performance in natural language processing tasks such as text classification and sentiment analysis (Li et al., 2023).

The IndoBERT model used in this research was obtained from the Hugging Face Model Hub, which provides pre-trained parameters optimized for Indonesian textual corpora. During the modeling process, input text was tokenized using the BERT tokenizer, which converts sentences into subword tokens that are subsequently transformed into numerical token identifiers.

The model was fine-tuned using the BertForSequenceClassification architecture to perform sentiment classification. The training process utilized the AdamW optimizer with several combinations of learning rates and dropout values in order to determine the optimal configuration for model training. Transformer-based classification models commonly use the AdamW optimizer because it improves training stability and model convergence when combined with appropriate hyperparameter settings (Lviv Polytechnic National University et al., 2025).

The experiment explored learning rates of 2e-5, 3e-5, and 3e-6, combined with dropout rates of 0.1, 0.2, and 0.3. Model performance was evaluated using validation data and assessed using Accuracy, Precision, Recall, and F1-Score.

The results show that the IndoBERT model achieved the best performance using a learning rate of 3e-5 and a dropout rate of 0.1, producing a training accuracy of 0.8752 and a validation accuracy of 0.7413. These results indicate that IndoBERT is capable of effectively capturing contextual semantic patterns in Indonesian social media text.

The strong performance of IndoBERT can be attributed to its bidirectional contextual representation, which enables the model to interpret semantic relationships between words from both directions within a sentence. This capability is particularly important for sentiment analysis in Indonesian language texts, where meaning is often influenced by contextual relationships between words rather than isolated lexical tokens.

Experimental Results Using the GPT-2 Model

The second experiment involved the implementation of GPT-2, a transformer-based language model that generates contextual representations using an autoregressive architecture. Unlike BERT-based models that analyze text bidirectionally, GPT-2 predicts tokens sequentially based on previously generated tokens. In autoregressive transformer models, the system learns contextual dependencies by predicting the next token in a sequence conditioned on preceding tokens (Boyd et al., 2022).

Similar to the IndoBERT experiment, the GPT-2 model was obtained from the Hugging Face Model Hub and adapted for sentiment classification using the GPT2ForSequenceClassification architecture. The input text was tokenized using the GPT-2 tokenizer, enabling the model to process textual sequences numerically.

The experimental configuration utilized the AdamW optimizer, with learning rate combinations of 2e-5, 3e-5, and 3e-6, as well as dropout values of 0.1, 0.2, and 0.3. In addition,

the model was trained using different numbers of epochs (5, 6, and 7 epochs) to determine the most effective training duration.

The results show that the best performance of the GPT-2 model was achieved using a learning rate of 3e-5 and a dropout rate of 0.1, producing a training accuracy of 0.6676 and a validation accuracy of 0.6651.

However, the training process revealed a noticeable difference between training loss and validation loss, indicating the presence of overfitting. This condition suggests that the model learned patterns specific to the training dataset but experienced difficulty generalizing when applied to unseen data. This limitation may be influenced by the relatively limited dataset size and potential imbalance in sentiment class distribution.

Model Performance Comparison

To better understand the performance differences between the two transformer-based models implemented in this study, the evaluation results of IndoBERT and GPT-2 were compared using standard classification metrics.

Table 2. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
IndoBERT	0.7413	0.74	0.73	0.73
GPT-2	0.6651	0.66	0.65	0.65

As shown in Table 2, the IndoBERT model consistently outperformed GPT-2 across all evaluation metrics. IndoBERT achieved an accuracy of 0.7413, which is significantly higher than the accuracy obtained by GPT-2 (0.6651).

This superior performance can be explained by several factors. First, IndoBERT employs bidirectional contextual embeddings, allowing the model to capture semantic relationships between words from both directions within a sentence. Second, IndoBERT was pre-trained specifically on Indonesian language corpora, making it more sensitive to linguistic patterns and vocabulary commonly found in Indonesian texts. In contrast, GPT-2 models are generally trained on multilingual or English-dominant datasets, which may reduce their ability to fully capture linguistic nuances present in Indonesian social media discourse.

These findings highlight the importance of using language-specific pre-trained models when analyzing sentiment in non-English textual datasets.

Sentiment Prediction and Topic Analysis

Based on the experimental comparison results, the IndoBERT model was selected as the best-performing model for the final sentiment prediction stage. The trained model was applied to previously unseen tweets discussing HPV vaccination to obtain insights into public sentiment regarding HPV vaccination in Indonesia.

Table 3. Sentiment Prediction Results

Sentiment	Number	Percentage
Positive	426	59.58%

Sentiment	Number	Percentage
Negative	133	18.60%
Neutral	156	21.81%

The results indicate that positive sentiment dominates public discourse, accounting for approximately 59.58% of the analyzed tweets. This suggests that most social media users express supportive attitudes toward HPV vaccination programs.

To further explore the factors influencing public sentiment, BERTopic modeling was applied to the sentiment-labeled dataset. This approach enables the identification of thematic patterns by grouping semantically similar discussions into interpretable topics.

Three dominant themes emerged within negative sentiment discussions:

1. concerns regarding vaccination policies and access inequality
2. personal experiences following HPV vaccination, including pain or soreness
3. financial barriers related to vaccine costs

Meanwhile, positive sentiment discussions were associated with:

1. increasing public awareness of the importance of HPV vaccination
2. improved access to vaccination services through digital health platforms

Overall, these findings suggest that although public discourse on social media tends to reflect supportive attitudes toward HPV vaccination, concerns regarding cost, accessibility, and policy implementation continue to shape public perceptions.

Interpretation, Policy Implications, and Future Research

The predominance of positive sentiment indicates that public awareness regarding HPV vaccination may have improved as a result of increased exposure to vaccination campaigns and health information. Previous studies also report that digital communication platforms can contribute to shaping public attitudes toward vaccination programs by facilitating the dissemination of health information and personal experiences (Kim et al., 2022; Liu et al., 2025).

In digital contexts, social media platforms function as important arenas where health information, experiences, and opinions circulate and shape collective understanding of vaccination programs. Online communication spaces therefore play a crucial role in influencing how individuals interpret vaccine risks and benefits, which ultimately affects vaccine acceptance experiences (Kim et al., 2022; Liu et al., 2025). Previous research also highlights that social media engagement and digital communication strategies can significantly influence how audiences receive and interpret information within online communities (Sari & Handayani, 2021).

From a theoretical perspective, these findings can be further interpreted through the lens of public perception and health communication frameworks. Public perception theory emphasizes that individuals construct attitudes toward health interventions not solely from scientific information, but also from social narratives, personal experiences, and the broader communication environment in which such information circulates. In contemporary digital ecosystems, social media platforms function as influential communication spaces where

health-related information, public opinions, and experiential narratives interact dynamically, shaping how individuals interpret the risks and benefits of vaccination. Within this context, online discussions about HPV vaccination can influence collective understanding of vaccine safety, accessibility, and effectiveness, thereby contributing to the formation of public attitudes toward vaccination programs. Previous studies have shown that digital communication environments can significantly shape vaccine-related perceptions and influence individuals' willingness to accept vaccination through the circulation of both supportive information and vaccine-related concerns (Kim et al., 2022; Liu et al., 2025). In addition, recent research highlights that social media discourse can amplify both positive narratives and hesitancy-related discussions, which ultimately affects public trust and decision-making regarding vaccination programs (Yoon et al., 2024).

Nevertheless, the persistence of discussions related to vaccine cost and accessibility highlights the existence of structural barriers that may influence vaccination uptake. These concerns align with previous research identifying economic constraints, safety concerns, and limited access to reliable health information as major factors contributing to vaccine hesitancy (Heyde et al., 2024; Sendekie et al., 2025).

From a policy perspective, the findings suggest that effective vaccination campaigns should not rely solely on communication strategies but should also address structural barriers such as vaccine affordability and equitable access to vaccination services. Monitoring social media discourse using computational techniques such as sentiment analysis and topic modeling can therefore provide valuable insights into public concerns and emerging issues related to vaccination programs.

Despite the valuable insights generated in this study, several limitations should be acknowledged. First, the dataset was limited to textual data obtained from a single social media platform, which may not fully represent the diversity of public opinion expressed across other digital platforms. Second, the dataset size remains relatively limited compared with large-scale social media analytics studies. Third, the analysis focused primarily on textual sentiment without incorporating multimodal data such as images or videos that may also influence online discussions.

Future research may address these limitations by integrating data from multiple social media platforms and exploring multimodal analytical approaches that combine textual, visual, and network-based data. In addition, future studies may investigate how misinformation spreads within digital networks and how social media interactions influence public trust in vaccination programs.

Overall, by integrating sentiment analysis and topic modeling, this study provides a comprehensive analytical framework for understanding public discourse surrounding HPV vaccination. The findings contribute not only to methodological developments in computational sentiment analysis but also to broader discussions on how digital public opinion can inform public health communication strategies and vaccination policies.

CONCLUSION

This study examined public sentiment toward the Human Papillomavirus (HPV) vaccine by analyzing discussions on social media platform X through the application of transformer-based language models. The experimental results indicate that the IndoBERT model demonstrates superior performance compared with the GPT-2 model in sentiment classification tasks. In particular, IndoBERT shows stronger stability and better generalization capability when applied to relatively small and imbalanced datasets. In contrast, the GPT-2 model exhibited signs of overfitting during the training process, suggesting limitations in its ability to generalize effectively when analyzing Indonesian social media text. These findings highlight the advantage of language-specific bidirectional transformer models in capturing contextual semantic relationships within Indonesian-language discourse.

The sentiment analysis results further reveal that public discourse regarding the HPV vaccine is predominantly positive, indicating that many social media users express supportive attitudes toward HPV vaccination programs. This positive orientation suggests that public awareness regarding the benefits of HPV vaccination as a preventive measure against cervical cancer has gradually increased. Nevertheless, the analysis also identifies several concerns expressed within online discussions, reflecting the complexity of public perceptions related to vaccination policies and practices.

Through the application of BERTopic-based topic modeling, this study further identifies several key factors shaping public perceptions of HPV vaccination. Negative sentiments are primarily associated with concerns regarding the high cost of the vaccine and the physical discomfort experienced after vaccination, such as pain, soreness, or temporary numbness. Meanwhile, positive sentiments are largely influenced by increasing awareness of the health benefits of vaccination, as well as considerations related to accessibility and the availability of vaccination services. These findings demonstrate that public attitudes toward vaccination are shaped not only by health awareness but also by practical considerations related to access and service delivery.

Overall, this research contributes to the growing body of literature on social media–based health sentiment analysis by demonstrating the effectiveness of transformer-based models, particularly IndoBERT, in analyzing Indonesian public discourse on vaccination. The integration of sentiment analysis with topic modeling also provides a more comprehensive understanding of the issues and concerns expressed in digital discussions. The insights generated in this study may support policymakers, healthcare institutions, and public health communicators in designing more responsive communication strategies and improving the implementation of HPV vaccination programs.

Future research may extend this work by incorporating larger and more diverse datasets, expanding the analysis to multiple social media platforms, or integrating multimodal data sources such as images and videos. Such approaches may provide a more comprehensive understanding of how digital discourse shapes public attitudes toward vaccination and other public health initiatives.

REFERENCE

Abdulhakim Al-Absi, A., Kang, D.-K., & Abdulhakim Al-Absi, M. (2023). Sentiment Analysis and Classification Using Deep Semantic Information and Contextual Knowledge. *Computers, Materials & Continua*, *74*(1), 671–691. https://doi.org/10.32604/cmc.2023.030262

Acuña-Cid, H. A., Ahumada-Tello, E., Ovalle-Osuna, Ó. O., Evans, R., Hernández-Ríos, J. E., & Zambrano-Soto, M. A. (2025). CRISP-NET: Integration of the CRISP-DM Model with Network Analysis. *Machine Learning and Knowledge Extraction*, *7*(3), 101. https://doi.org/10.3390/make7030101

Aljabar, A., Ali, I., & Karomah, B. M. (2024). Sentiment Analysis Using Transformer Method. *Journal of Informatics Information System Software Engineering and Applications (INISTA)*, *6*(2), 90–97. https://doi.org/10.20895/inista.v6i2.1383

Arbyn, M., Rousta, P., Bruni, L., Schollin Ask, L., & Basu, P. (2024). Linkage of individual-patient data confirm protection of prophylactic human papillomavirus vaccination against invasive cervical cancer. *JNCI: Journal of the National Cancer Institute*, *116*(6), 775–778. https://doi.org/10.1093/jnci/djae042

Boyd, A., Showalter, S., Mandt, S., & Smyth, P. (2022). *Predictive Querying for Autoregressive Neural Sequence Models* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2210.06464

Das, U. K., Ani, R. S., Datta, N., Fahad, I., Sikder, J., Sara, U., & Chakraborty, A. (2025). Enhancing sentiment analysis accuracy on social media comments using a tuned BERT model. *Discover Computing*, *28*(1), 198. https://doi.org/10.1007/s10791-025-09599-x

Deng, Y., Van Der Meer, J., Tzovara, A., Schmidt, M., Bassetti, C., & Denecke, K. (2025). Analyzing Sleep Behavior Using BERT-BiLSTM and Fine-Tuned GPT-2 Sentiment Classification: Comparison Study. *JMIR Medical Informatics*, *13*, e70753–e70753. https://doi.org/10.2196/70753

Dotan, E., Jaschek, G., Pupko, T., & Belinkov, Y. (2024). Effect of tokenization on transformers for biological sequences. *Bioinformatics*, *40*(4), btae196. https://doi.org/10.1093/bioinformatics/btae196

Ellingson, M. K., Sheikha, H., Nyhan, K., Oliveira, C. R., & Niccolai, L. M. (2023). Human papillomavirus vaccine effectiveness by age at vaccination: A systematic review. *Human Vaccines & Immunotherapeutics*, *19*(2), 2239085. https://doi.org/10.1080/21645515.2023.2239085

Furuno, A., Sukegawa, A., Ohshige, K., Suzuki, Y., Yamaguchi, M., Miyagi, E., Ueda, Y., Sekine, M., & Mizushima, T. (2024). Three-year questionnaire study on human papillomavirus vaccination targeting new female college school students: Follow-up to a 2021 report to reveal the impact of a policy change in Japan. *Journal of Obstetrics and Gynaecology Research*, *50*(9), 1640–1648. https://doi.org/10.1111/jog.16049

Heyde, S., Osmani, V., Schauberger, G., Cooney, C., & Klug, S. J. (2024). Global parental acceptance, attitudes, and knowledge regarding human papillomavirus vaccinations for their children: A systematic literature review and meta-analysis. *BMC Women's Health*, *24*(1), 537. https://doi.org/10.1186/s12905-024-03377-5

Khan, J., Ahmad, K., Jagatheesaperumal, S. K., & Sohn, K.-A. (2025). Textual variations in social media text processing applications: Challenges, solutions, and trends. *Artificial Intelligence Review*, *58*(3), 89. https://doi.org/10.1007/s10462-024-11071-z

Kim, S. J., Schiffelbein, J. E., Imset, I., & Olson, A. L. (2022). Countering Antivax Misinformation via Social Media: Message-Testing Randomized Experiment for Human Papillomavirus Vaccination Uptake. *Journal of Medical Internet Research*, *24*(11), e37559. https://doi.org/10.2196/37559

Li, Z., Yang, C., & Huang, C. (2023). A Comparative Sentiment Analysis of Airline Customer Reviews Using Bidirectional Encoder Representations from Transformers (BERT) and Its Variants. *Mathematics*, *12*(1), 53. https://doi.org/10.3390/math12010053

Liu, J., Niu, Q., Nagai-Tanima, M., & Aoyama, T. (2025). Understanding Human Papillomavirus Vaccination Hesitancy in Japan Using Social Media: Content Analysis. *Journal of Medical Internet Research*, *27*, e68881. https://doi.org/10.2196/68881

Lundén, N., Bekar, E. T., Skoogh, A., & Bokrantz, J. (2023). Domain Knowledge in CRISP-DM: An Application Case in Manufacturing. *IFAC-PapersOnLine*, *56*(2), 7603–7608. https://doi.org/10.1016/j.ifacol.2023.10.1156

Lviv Polytechnic National University, Podolchak, N., Tsygylyk, N., Lviv Polytechnic National University, Petlovanyi, M., & Lviv Polytechnic National University. (2025). Mathematical modeling of multi-label classification of job descriptions using transformer-based neural networks. *Mathematical Modeling and Computing*, *12*(3), 767–778. https://doi.org/10.23939/mmc2025.03.767

Ma, L., Chen, R., Ge, W., Rogers, P., Lyn-Cook, B., Hong, H., Tong, W., Wu, N., & Zou, W. (2025). AI-powered topic modeling: Comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. *Experimental Biology and Medicine*, *250*, 10389. https://doi.org/10.3389/ebm.2025.10389

Mascarenhas, A. K., Kelekar, A., Lucia, V. C., & Afonso, N. M. (2024). The receipt of the human papillomavirus vaccine's influence on future human papillomavirus vaccine recommendations by medical and dental students. *JADA Foundational Science*, *3*, 100029. https://doi.org/10.1016/j.jfscie.2023.100029

Naoum, P., Athanasakis, K., Zavras, D., Kyriopoulos, J., & Pavi, E. (2022). Knowledge, Perceptions and Attitudes Toward HPV Vaccination: A Survey on Parents of Girls Aged 11–18 Years Old in Greece. *Frontiers in Global Women's Health*, *3*, 871090. https://doi.org/10.3389/fgwh.2022.871090

Papia, S. K., Khan, M. A., Habib, T., Rahman, M., & Islam, M. N. (2024). DistilRoBiLSTMFuse: An efficient hybrid deep learning approach for sentiment analysis. *PeerJ Computer Science*, *10*, e2349. https://doi.org/10.7717/peerj-cs.2349

Pencheva, D. (2025). Profiling Noisy Social Media Data for Sentiment Applications: A Visual and Analytical Framework. *SAR Journal - Science and Research*, 213–224. https://doi.org/10.18421/SAR83-01

Saputra, A. N. A., Saputro, R. E., & Saputra, D. I. S. (2025). Enhancing Sentiment Analysis Accuracy Using SVM and Slang Word Normalization on YouTube Comments. *Sinkron*, *9*(2), 687–699. https://doi.org/10.33395/sinkron.v9i2.14613

Sari, Y., & Handayani, M. (2021). ANALYSIS OF THE "SOME" MODEL (SHARE, OPTIMIZE, MANAGE, ENGAGE) INSTAGRAM ACCOUNT @tnlkep Kepulauanseribu IN THE FRAMEWORK OF DIGITAL PROMOTION OF A THOUSAND ISLANDS MARINE PARK AS AN ECO-TOURISM DESTINATION FOR THE MILLENIAL GENERATION. *Moestopo International Review on Social, Humanities, and Sciences*, *1*(1), 7–15. https://doi.org/10.32509/mirshus.v1i1.5

Sendekie, A. K., Abate, B. B., Adamu, B. A., Tefera, A. M., Mekonnen, K. T., Ashagrie, M. A., Tadesse, Y. B., Dagnaw, A. D., Melaku, M. S., & Bizuneh, G. K. (2025). Human papillomavirus vaccination hesitancy among young girls in Ethiopia: Factors and barriers to uptake. *Frontiers in Public Health*, *13*, 1507832. https://doi.org/10.3389/fpubh.2025.1507832

Singh, G., Dash, N. R., Shaju, A., & Chakkalakkunnan, S. S. (2025). Perceptions and sentiments associated with HPV vaccine uptake among Indian Reddit users: A qualitative social media analysis. *BMC Public Health*, *25*(1), 4037. https://doi.org/10.1186/s12889-025-25418-w

Surya, J., Kashyap, H., Nadig, R. R., & Raman, R. (2023). Developing a Risk Stratification Model Based on Machine Learning for Targeted Screening of Diabetic Retinopathy in the Indian Population. *Cureus*. https://doi.org/10.7759/cureus.45853

Talaat, A. S. (2023). Sentiment analysis classification system using hybrid BERT models. *Journal of Big Data*, *10*(1), 110. https://doi.org/10.1186/s40537-023-00781-w

Xue, J., Zhang, B., Zhang, Q., Hu, R., Jiang, J., Liu, N., Peng, Y., Li, Z., & Logan, J. (2023). Using Twitter-Based Data for Sexual Violence Research: Scoping Review. *Journal of Medical Internet Research*, *25*, e46084. https://doi.org/10.2196/46084

Yoon, S., Kim, H., An, J., & Jin, S. W. (2024). Exploring human papillomavirus vaccine hesitancy among college students and the potential of virtual reality technology to increase vaccine acceptance: A mixed-methods study. *Frontiers in Public Health*, *12*, 1331379. https://doi.org/10.3389/fpubh.2024.1331379

Zouhar, V., Meister, C., Gastaldi, J. L., Du, L., Sachan, M., & Cotterell, R. (2023). *Tokenization and the Noiseless Channel* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2306.16842