I.J. Intelligent Systems and Applications, **2026, 4, 226-238**
Published Online on August 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2026.04.12

Topic Modeling and Sentiment Analysis on News Headlines Using BERTopic and IndoBERT Models

Bagas Yana Prayoga
Department of Information System, Universitas Islam Negeri Syarif Hidayatullah Tangerang Selatan, 15412, Indonesia
E-mail: bagasyana21@mhs.uinjkt.ac.id
ORCID iD: https://orcid.org/0009-0005-9579-2657

Qurrotul Aini*
Department of Information System, Universitas Islam Negeri Syarif Hidayatullah Tangerang Selatan, 15412, Indonesia
E-mail: qurrotul.aini@uinjkt.ac.id
ORCID iD: https://orcid.org/0000-0002-2722-4755
*Corresponding author

Fitroh Fitroh
Department of Information System, Universitas Islam Negeri Syarif Hidayatullah Tangerang Selatan, 15412, Indonesia
E-mail: fitroh@uinjkt.ac.id
ORCID iD: https://orcid.org/0000-0001-9548-0155

Received: February 9, 2026; Revised: March 24, 2026; Accepted: May 17, 2026; Published: August 8, 2026

Abstract: The high volume of information from online media in Indonesia poses challenges for manual analysis in identifying emerging themes and sentiments. News headlines, as the primary element seen by the public, play a crucial role in shaping opinion; however, their massive volume and diverse themes necessitate automated approaches to identify topics and underlying sentiments. To address this challenge, this study analyzed 30,329 news headlines from the online news portal detik.com for the entire year 2024. A quantitative Natural Language Processing (NLP) framework was applied, comprising data collection via web scraping, text preprocessing, transformer-based topic modeling using BERTopic, sentiment classification using IndoBERT, and a topic-sentiment intersection analysis. Preprocessing included case folding, text cleaning, normalization of informal words, and tokenization. For lexicon-based labeling, stopword removal and stemming were applied. In contrast, transformer-based models utilized text that underwent only case folding, cleaning, and normalization to preserve contextual information. Topic modeling was performed using BERTopic, while sentiment classification (positive, negative, and neutral) used the IndoBERT model. The main objective of this study was to evaluate the combined performance of the two models in mapping dominant issues and the sentiments contained in media reports. The results showed that BERTopic successfully identified 366 topics. An evaluation of the 10 most dominant topics yielded a coherence score of 0.5145, indicating a relevant topic clustering. IndoBERT demonstrated high agreement with lexicon-generated sentiment labels, with an accuracy of 94.78%, a precision of 95.04%, a recall of 94.79%, and an F1-score of 94.81%. These findings confirm that the combination of transformer-based models is effective for in-depth discourse analysis of political news headlines from detik.com, a major Indonesian online news portal.

Index Terms: Topic Modeling, Sentiment Analysis, News Headlines, BERTopic, IndoBERT.

1. Introduction

The digital era has fundamentally changed the way people access and consume information, with digital media now an integral part of everyday life. A 2024 report from We Are Social showed that the average internet user in Indonesia spent 7 hours and 38 minutes daily on various online activities [1]. Of that duration, the time allocated to reading news, both online and in print, reached 1 hour and 26 minutes. This significant duration indicates the enormous volume of textual data produced by online news portals and consumed by the public. This massive volume of data creates new challenges, where manual analysis to understand dominant issues and overall sentiment patterns is no longer efficient or

adequate. Therefore, a technological approach is needed that can systematically and objectively process and analyze large-scale data.

Natural Language Processing (NLP), a branch of artificial intelligence that enables machines to process and understand human language, offers a solution to this challenge. In the context of news analysis, two key approaches in NLP are particularly relevant: topic modeling, which identifies dominant themes in news reports [2], and sentiment analysis, which explores the opinions or emotional polarities contained within them [3]. For topic modeling, this study utilized BERTopic, a modern deep learning-based model that integrates embeddings from Bidirectional Encoder Representations from Transformers (BERT) with advanced clustering algorithms. BERTopic demonstrates significant advantages over traditional methods such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) due to its ability to understand word meaning contextually, rather than simply based on their frequency of occurrence. Several studies have demonstrated BERTopic's superiority in analyzing COVID-19-related Twitter data [4], while other studies [5, 6] also confirmed its better performance than LDA in terms of topic coherence and accuracy in chatbot and news analysis.

Meanwhile, for sentiment analysis, this study relies on IndoBERT, a language model developed specifically for Indonesian [7]. As a model trained using a massive corpus of Indonesian language data, IndoBERT can capture contextual nuances, language variations, and even informal language with high accuracy. Its superiority has been demonstrated in various studies, such as [8], which showed that IndoBERT achieved an F1-score of 84% in sentiment analysis of PPKM policies on Twitter, outperforming SVM (70%) and Naïve Bayes (83%). Similarly, another study reported an accuracy of 82.19% for IndoBERT in sentiment analysis of COVID-19 news reports [9]. Furthermore, the effectiveness of IndoBERT was confirmed in a study that combined IndoBERT and BERTopic to analyze 32,985 tweets [10]. In this study, the sentiment classification component achieved an accuracy of 96%, demonstrating the model's superior ability to process Indonesian text within a topic modeling framework.

Although BERTopic and IndoBERT have proven effective across various domains, their combined application to news headlines remains limited. News headlines are uniquely concise and formal, yet play a crucial role in shaping the public's initial perception of an issue [11]. This study chose political news as a case study, as it is one of the most popular topics among Indonesian netizens, accounting for 40.56% of interest according to a report from Goodstats [12]. Data for this study were taken from detik.com, one of the top online news portals in Indonesia, with a weekly reach of 50% according to the Digital News Report 2024 from the University of Oxford [13].

Thus, this study aims to evaluate the effectiveness of combining BERTopic and IndoBERT for topic modeling and sentiment analysis of Indonesian political news headlines. This research is expected to provide empirical evidence regarding the ability of transformer-based models to analyze concise and formal news texts and serve as a reference for future studies. This study treats the 2024 news corpus as a single dataset to evaluate the effectiveness of transformer-based models, rather than focusing on temporal topic evolution.

2. Related Works

2.1. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a branch of artificial intelligence that enables computers to understand, interpret, and generate natural human language [14]. NLP has diverse applications, including chatbots and virtual assistants that interpret user queries, and machine translation systems that preserve meaning and context across languages.

Technological developments in the field of NLP have made significant progress with the emergence of transformer-based models such as BERT and GPT [15]. These models have transformed the way NLP systems understand context through attention mechanisms that allow the models to focus on the most relevant parts of the text.

2.2. Topic Modeling

Topic modeling is a statistical method that aims to identify hidden themes or topics in a collection of text documents. Topic modeling algorithms group documents into thematic clusters. This method, which is included in unsupervised machine learning, uses a soft/fuzzy clustering approach where each object can belong to more than one cluster [16]. Topic modeling works by identifying groups of words that frequently appear together and considering each document as a mixture of several topics. Each topic is then represented by a set of words with a certain probability of occurrence.

2.3. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a revolutionary language model introduced by Google researchers in 2018 [17]. This model uses *a* transformer architecture that allows for bidirectional understanding of the context of words in a sentence. In contrast to traditional language models that process text sequentially from left to right or right to left, BERT can understand the context of words by considering words that appear before and after simultaneously. This ability is made possible through a self-attention mechanism on the transformer architecture, which allows the model to assign different weights to each word in a sentence based on its relevance to the word being processed.

2.4. BERTopic

BERT (Bidirectional Encoder Representations from Transformers), introduced by a Google research team, is a

revolutionary model that uses a transformer architecture to process text bidirectionally [17]. BERT's power in producing semantically rich text representations is further exploited by modern topic modeling methods such as BERTopic. Introduced by Grootendorst [18], BERTopic was designed to address the limitations of traditional models like LDA and NMF, which often fail to capture nuanced semantic relationships between words. This method integrates transformer-based document embedding with clustering techniques such as HDBSCAN and topic representation using a variation of Class-based TF-IDF (c-TF-IDF). By leveraging BERT's contextual embedding as its foundation, BERTopic can generate more accurate and semantically coherent topics, making it a superior choice for in-depth text analysis.

2.5. Sentiment Analysis

Sentiment analysis is a branch of Natural Language Processing (NLP) that aims to identify, extract, and classify the opinions, attitudes, or emotions contained in text [19]. Basically, sentiment analysis focuses on grouping text into positive, negative, or neutral sentiments. In information systems, sentiment analysis plays an important role in helping organizations understand user sentiment patterns towards digital services or products [20].

2.6. IndoBERT

IndoBERT is a language model based on the BERT architecture that is specially trained for the Indonesian language [21]. This model was developed to overcome the limitations of pre-trained language models, which are mostly focused on English and other Indo-European languages. IndoBERT is trained using a large corpus of Indonesian texts, covering a wide range of sources such as news articles, Indonesian Wikipedia, and other web content [7]. The IndoBERT architecture adopts the basic structure of the BERT model with some modifications to accommodate the special characteristics of the Indonesian language. In its implementation, IndoBERT has shown superior performance compared to the multilingual language model for various Indonesian processing tasks [9]. This advantage is particularly evident in its ability to understand the local context and language nuances specific to Indonesian, including the use of informal and mixed languages commonly found in modern Indonesian texts.

3. Methodology

This study adopts a quantitative approach with a descriptive and exploratory research design. The research process follows a structured Natural Language Processing (NLP) workflow consisting of problem formulation, data collection, dataset description, text preprocessing, transformer-based modeling, and topic sentiment intersection analysis. The overall research framework is illustrated in Fig. 1.

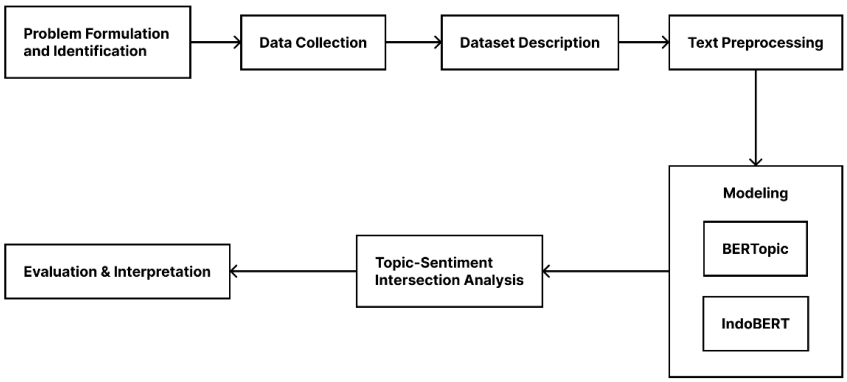

Fig.1. Research framework

3.1. Problem Formulation and Identification

This stage defines the research problem and establishes the analytical approach adopted to address it. This study examines the analysis of political news headlines to reveal topics and sentiments, considering the role of digital media in shaping public opinion. By combining BERTopic for topic modeling and IndoBERT for sentiment analysis, the research focuses on online news portals detik.com as the main data source to provide an in-depth understanding of the presentation and reception of political news in Indonesia.

3.2. Data Collection

This stage includes data collection activities through web scraping techniques from online news portals detik.com. The data collected is based on the period from January 1, 2024, to December 31, 2024, related to Indonesian politics. This scraping process was implemented using the Python programming language and web scraping libraries such as BeautifulSoup.

3.3. Dataset Description

This stage provides a descriptive analysis of the scraped political news headlines, including temporal distribution to identify reporting intensity across different periods. The distribution of the data is then visualized on a time-based basis to identify periods with high news volumes and uncover trends in political reporting.

3.4. Text Preprocessing

The raw data then underwent a systematic preprocessing stage, including case folding, data cleaning, tokenization, normalization, stopword removal, and stemming, to ensure data quality and consistency [23]. For the transformer-based models (IndoBERT and BERTopic), the original sentence structure was preserved without stemming or stopword removal, as excessive preprocessing may remove contextual information required by the self-attention mechanism. Stopword removal and stemming were applied only during the lexicon-based sentiment labeling stage to improve dictionary matching accuracy.

3.5. Modeling

The cleaned data was processed using two models. First, BERTopic was applied to identify topics, using embeddings from the indobenchmark/indobert-base-p1 model [22]. Second, the IndoBERT model was used for sentiment classification after the data was automatically labeled using a lexicon-based approach [23] and balanced with the Random Oversampling (ROS) technique [24].

The model's performance was evaluated quantitatively. The topic quality of BERTopic was assessed using the coherence score, whereas the IndoBERT classification performance was evaluated based on accuracy, precision, recall, and F1-score metrics derived from the confusion matrix. The model's performance was validated using 5-Fold Cross-Validation, following the approach presented in [26, 27].

3.6. Topic–Sentiment Intersection Analysis

To obtain deeper insights into the distribution of sentiment across thematic clusters, a topic sentiment intersection analysis was conducted. This stage integrates the dominant topics generated by BERTopic with sentiment predictions produced by IndoBERT. Each news headline was mapped to its corresponding topic and sentiment label, enabling the calculation of sentiment distribution within each topic cluster. This analysis aims to identify how sentiment polarity (positive, negative, and neutral) is distributed across major political themes, thereby revealing the dominant emotional tone associated with specific political narratives in Indonesian online media.

3.7. Evaluation

The interpretation stage focuses on analyzing and reflecting on the results obtained through three main aspects. First, the performance metrics are examined and compared with related studies to validate the robustness of the findings. Second, the study's contribution is highlighted by emphasizing the integration of BERTopic and IndoBERT in the domain of political news headlines, which remains relatively underexplored. Third, the practical and academic implications of the findings are discussed, particularly their relevance for researchers, political analysts, and media practitioners.

4. Result

4.1. Dataset Description

The data collection process via web scraping from detik.com for the period January 1 to December 31, 2024, yielded 30,709 raw news headlines. Each data entry collected includes headline text, source URL, and publication date. An exploratory analysis of the temporal distribution of data shows fluctuations in news volume throughout the year, with the highest peaks identified in January (3,623 headlines) and August (3,389 headlines). This monthly distribution is presented visually in the form of a bar chart in Fig. 2.

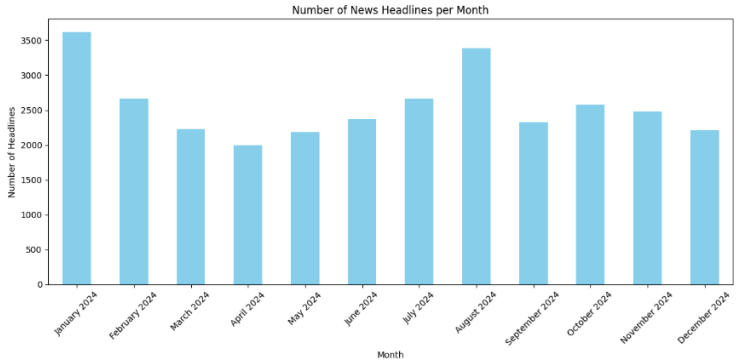

Fig.2. Monthly frequency of news headlines in the dataset (January–December 2024)

4.2. Preprocessing Data Results

Before text preprocessing, duplicate entries were identified and removed based on the news headline column. Consequently, out of the 30,709 collected records, 30,329 remained for analysis. Furthermore, text preprocessing was carried out to reduce noise and inconsistencies, ensuring the data was clean, structured, and ready for analysis. The preprocessing was conducted in the following stages:

- Case folding: The first step was to convert all letters in the text to lowercase. This process is crucial to avoid differences in word representation due to capitalization. For example, the words "POLITIK" and "politik" were treated as identical after case alignment. The results of this step can be seen in Table 1.

Table 1. Example of case folding applied to news headline

Before case folding	After case folding
Kapolri Imbau Jaga Kerukunan Jelang Pemilu 2024	kapolri imbau jaga kerukunan jelang pemilu 2024
Mahfud Vs TPN Prabowo-Gibran soal Program Makan Siang Gratis	mahfud vs tpn prabowo-gibran soal program makan siang gratis
Kubu AMIN Anggap Fahri Lucu-lucuan Unggah Foto Lawas Prabowo-Anies-Sandi	kubu amin anggap fahri lucu-lucuan unggah foto lawas prabowo-anies-sandi

- Data cleaning: At this stage, unnecessary characters, such as numbers, punctuation, and excessive spaces, were removed. The goal is to reduce noise in the text data, making it cleaner and more standardized. Some examples of data cleaning results can be seen in Table 2.

Table 2. Example of data cleaning applied to news headline

Before data cleaning	After case folding
ganjar: indonesia butuh 3 pabrik pupuk baru untuk penuhi kebutuhan petani	ganjar indonesia butuh pabrik pupuk baru untuk penuhi kebutuhan petani
bagaimana cara mencoblos di luar negeri? simak 3 metodenya!	bagaimana cara mencoblos di luar negeri simak metodenya
anies tak masalah 2 panelis debat pilpres dari unhan: hadapi saja	anies tak masalah panelis debat pilpres dari unhan hadapi saja

- Tokenization: This step involved breaking sentences down into smaller word components. This process is known as tokenization. Tokenization makes it easier for the system to analyze each word separately. An example of the results can be seen in Table 3.

Table 3. Example of tokenization applied to news headline

Before tokenization	After tokenization
aksi boikot produk israel ubah pola konsumsi masyarakat brand lokal	['aksi', 'boikot', 'produk', 'israel', 'ubah', 'pola', 'konsumsi', 'masyarakat', 'brand', 'lokal']
tkn prabowo gibran minta dukung sabar respons isu pemilu curang	['tkn', 'prabowo', 'gibran', 'minta', 'dukung', 'sabar', 'respons', 'isu', 'pemilu', 'curang']
kpu siap tahap pilkada mulai april	['kpu', 'siap', 'tahap', 'pilkada', 'mulai', 'april']

- Normalization: This process was performed to standardize non-standard words or those not conforming to the KBBI (Indonesian Dictionary). This stage aims to avoid variations in words that have the same meaning but are spelled differently. The results of an example of the normalization application can be seen in Table 4.

Table 4. Example of normalization applied to news headline

Before normalization	After normalization
['jokowi', 'saat', 'ditanya', 'gaya', 'debat', 'gibran', 'saya', **'nggak'**, 'mau', 'menilai', 'lagi']	['jokowi', 'saat', 'ditanya', 'gaya', 'debat', 'gibran', 'saya', **'tidak'**, 'mau', 'menilai', 'lagi']
['tpn', 'ganjar', 'prediksi', 'gibran', 'bakal', 'standar', **'aja'**, 'di', 'debat', 'malam', 'ini']	['tpn', 'ganjar', 'prediksi', 'gibran', 'bakal', 'standar', **'saja'**, 'di', 'debat', 'malam', 'ini']
['mahfud', 'soal', 'program', 'susu', 'gratis', 'prabowo', 'kita', 'yang', 'kecil', 'kecil', **'aja'**, 'impor']	['mahfud', 'soal', 'program', 'susu', 'gratis', 'prabowo', 'kita', 'yang', 'kecil', 'kecil', **'saja'**, 'impor']

- Stopword removal: This stage aimed to eliminate words deemed irrelevant to the analysis, such as "yang," "di," "ke," "dan," and other common words. Some examples of the results can be seen in Table 5.

Table 5. Example of stopword removal applied to news headline

Before stopword removal	After stopword removal
['prabowo', **'ingin'**, 'tni', 'polri', 'tetap', **'di'**, 'bawah', 'presiden']	['prabowo', 'tni', 'polri', 'tetap', 'bawah', 'presiden']
['bicara', **'ke'**, 'menkeu', 'jokowi', 'minta', 'subsidi', 'pupuk', 'ditambah', **'jadi'**, 'rp', 'triliun']	['bicara', 'menkeu', 'jokowi', 'minta', 'subsidi', 'pupuk', 'ditambah', 'rp', 'triliun']
['dinilai', 'kurang', 'keras', **'saat'**, 'debat', **'dengan'**, 'gibran', 'mahfud', 'bilang', 'begini']	['dinilai', 'kurang', 'keras', 'debat', 'gibran', 'mahfud', 'bilang', 'begini']

- Stemming: This process involved converting words with affixes to their base form. For example, the words "membangun," "pembangun," and "dibangun" will be converted to the base word "bangun." This is done to prevent the model from treating these words as separate entities when they have the same meaning. An example of the application results can be seen in Table 6.

Table 6. Example of stemming applied to news headline

Before stemming	After stemming
['kpu', **'umumkan'**, 'lembaga', 'survei', **'terdaftar'**, 'pemilu']	['kpu', **'umum'**, 'lembaga', 'survei', **'daftar'**, 'pemilu']
['ridwan', 'kamil', 'ikn', 'bukan', 'ide', 'jokowi', 'implementasi', **'kewajiban'**, 'sejarah']	['ridwan', 'kamil', 'ikn', 'bukan', 'ide', 'jokowi', 'implementasi', **'wajib'**, 'sejarah']
['waka', 'mpr', 'sebut', **'peningkatan'**, 'jumlah', 'desa', 'wisata', **'berdampak'**, 'positif']	['waka', 'mpr', 'sebut', **'tingkat'**, 'jumlah', 'desa', 'wisata', **'dampak'**, 'positif']

It is important to note that stopword removal and stemming were primarily used for the lexicon-based sentiment labeling process. For the transformer-based modeling stages (IndoBERT and BERTopic), the normalized text, excluding stopword removal and stemming, was utilized to preserve contextual semantic information.

4.3. Topic Modeling Using BERTopic

As shown in Fig. 3, the topic modeling process with BERTopic began by converting the preprocessed headlines into vector representations (embeddings) using the indobenchmark/indobert-base-p1 model. These vectors are then processed by BERTopic which automatically performs clustering (HDBSCAN) and topic representation extraction (c-TF-IDF). From this process, the model managed to identify 366 topics seen in Fig. 3, with a significant finding that 13,837 documents (45.7%) were classified as noise (topic -1). The relatively high proportion of noise reflects the fragmented and event-driven structure of political news headlines. In density-based clustering methods such as HDBSCAN, documents that do not form dense semantic neighborhoods are intentionally labeled as noise to preserve cluster purity rather than being forced into artificial groupings.

	Topic	Count	Name	Representation
0	-1	13837	-1_prabowo_jokowi_soal_minta	[prabowo, jokowi, soal, minta, tak, anies, men...
1	0	713	0_demokrasi_didik_politik_bijak	[demokrasi, didik, politik, bijak, merdeka, di...
2	1	424	1_bamsoet_dorong_mkd_bangsa	[bamsoet, dorong, mkd, bangsa, mpr, tingkat, a...
3	2	329	2_pilgub_jakarta_anies_maju	[pilgub, jakarta, anies, maju, pks, dki, usung...
4	3	271	3_fransiskus_paus_misa_kunjung	[fransiskus, paus, misa, kunjung, katedral, gb...
...	...	...	...	...
362	361	10	361_giant_wall_sea_mbz	[giant, wall, sea, mbz, urbanisasi, ev, ekolog...
363	362	10	362_lalin_rekayasa_etle_situasional	[lalin, rekayasa, etle, situasional, patung, k...
364	363	10	363_podomoro_tenjo_angsur_jtan	[podomoro, tenjo, angsur, jtan, millennials, j...
365	364	10	364_miss_tuyul_singapura_skandal	[miss, tuyul, singapura, skandal, ukraina, lan...
366	365	10	365_pecat_bobby_nasution_pdip	[pecat, bobby, nasution, pdip, rakabuming, cam...

Fig.3. Topic identification results using BERTopic

Since the dataset consists of short news headlines, BERTopic's class-based TF-IDF (c-TF-IDF) mechanism is particularly suitable for this scenario. Unlike traditional document-level TF-IDF, c-TF-IDF emphasizes discriminative terms within each topic cluster, which helps mitigate the sparsity problem commonly found in short-text data such as headlines.

The BERTopic modeling process initially produced 366 topics from 30,329 news headlines. On average, this corresponds to approximately 82 documents per topic, which remains within a reasonable range for density-based clustering applied to large, event-driven corpora. This relatively large number of topics reflects the event-driven nature of political news, where many headlines correspond to specific incidents, short-term issues, or individual political statements. As a result, the model tends to generate fine-grained clusters that capture micro-level events rather than broad thematic categories.

However, since the objective of this study is to analyze dominant discourse patterns rather than isolated event fragments, the analysis focuses exclusively on the 10 most dominant topics ranked by document frequency. These dominant topics represent the most recurring and substantively significant political issues during the observation period, while lower-frequency clusters are treated as event-specific or peripheral semantic variations. These 10 dominant topics were subsequently integrated with sentiment classification results to examine how thematic prominence corresponds to sentiment distribution, forming the core analytical contribution of this study.

The quality of topics was evaluated using the coherence score, which ranged from 0.3227 to 0.7095, with an average of 0.5145 for the top 10 topics. This score indicates a moderate level of topic quality. Although the value is not considered high in absolute terms, it reflects acceptable semantic consistency given the short-text and event-driven nature of news

headlines. For comparison, an LDA baseline on the same dataset produced a coherence score of 0.3252 for the top 10 dominant topics. This result shows that BERTopic provides a substantially better semantic grouping than the traditional LDA approach in this short-text news scenario. Table 7 presents the 10 most dominant topics identified by the BERTopic model.

Table 7. Distribution of the ten most dominant topics identified by BERTopic

Topic	Number of Headline	Main Topic
1	713	*Pendidikan politik dan demokrasi*
2	424	*Peran lembaga legislatif dan bamsoet*
3	329	*Kontestasi pilkada jakarta dan tokoh politik*
4	271	*Kunjungan paus fransiskus*
5	250	*Polemik capres dan capres alternatif*
6	241	*Politik internasional dan tokoh global*
7	234	*Inisiatif MPR dalam peningkatan sektor strategis*
8	232	*Isu pekerja migran dan imigrasi ilegal*
9	224	*Kerja sama bilateral dan pertemuan internasional*
10	222	*Seruan pemilu damai*

4.4. Sentiment Analysis Using IndoBERT

Sentiment analysis in this study was conducted using the IndoBERT model, a transformer-based language model specifically pre-trained on Indonesian text. This process aims to classify news headlines into three sentiment categories: positive, negative, and neutral. To achieve this, a series of preprocessing and modeling steps was carried out, including automatic sentiment labeling, handling of data imbalance, and performance evaluation. The following subsections describe these stages in detail, starting with sentiment labeling of the dataset, followed by the application of data balancing techniques, and concluding with the evaluation of the IndoBERT model.

- Data labeling using lexicon based: Before training the supervised learning model, each headline in the dataset is automatically labeled sentiment. This process uses a lexicon-based approach with the InSet Lexicon sentiment dictionary, which is specially designed for Indonesian. Each headline was assigned a sentiment score based on predefined dictionaries, with the dominant polarity serving as the ground-truth label. This method was selected to facilitate the large-scale labeling of 30,329 headlines without manual annotation. The resulting labels were subsequently used to train and evaluate the IndoBERT sentiment classification model. As shown in Table 8, the labeling results reveal a significant class imbalance, consisting of 19,262 positive headlines (63.5%), 8,629 negative headlines (28.5%), and 2,438 neutral headlines (8.0%).

Table 8. Number of headlines per sentiment class

Sentiment	**Amount**
Positive	19262
Negative	8629
Neutral	2438

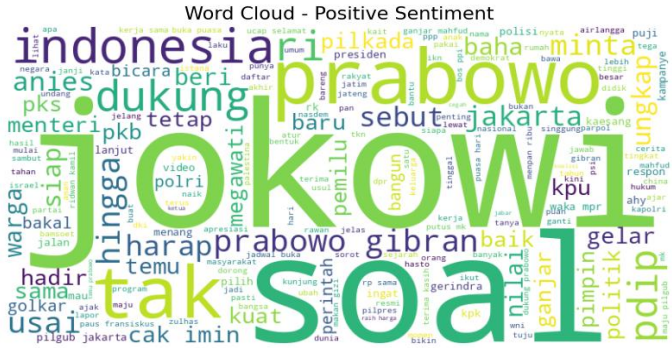

Fig.4. Positive sentiment word cloud visualization

Sentiment labels were generated automatically using a lexicon-based scoring approach. Therefore, the reported performance reflects the IndoBERT model's ability to learn and generalize from lexicon-supervised labels rather than manually annotated ground truth. To better capture the dominant lexical characteristics of each sentiment category, a word cloud visualization was employed. This technique visually represents the most frequent words in the dataset, with

separate visualizations provided for the positive, negative, and neutral sentiment classes. In Fig. 4 which illustrates positive sentiment, words such "dukung", "ganjar", and "temu" emerge as dominant elements. These words indicate support, active participation, and a positive tone within the discussed topics, reflecting a tendency toward opinions that are approving or appreciative.

Furthermore, Fig. 5 illustrates the distribution of words in the negative sentiment category. Words such as "anak", "apa", and "korupsi" dominate, pointing to issues that are often critical or problematic. This indicates that the narratives in this category predominantly contain criticism, skeptical inquiries, or discussions of issues that evoke negative perceptions.

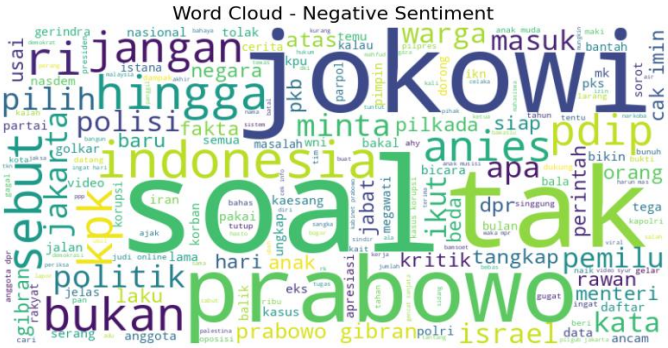

Fig.5. Negative sentiment word cloud visualization

Meanwhile, Fig. 6 presents the word cloud for the neutral sentiment category. In this category, words such as "politik", and "soal" frequently appear, reflecting discussions that are primarily informative, descriptive, or emotionally neutral. Headlines in this category tend not to convey strong emotional affiliation, either positively or negatively.

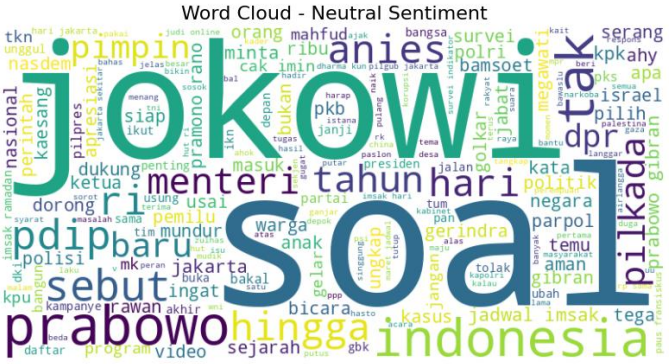

Fig.6. Neutral sentiment word cloud visualization

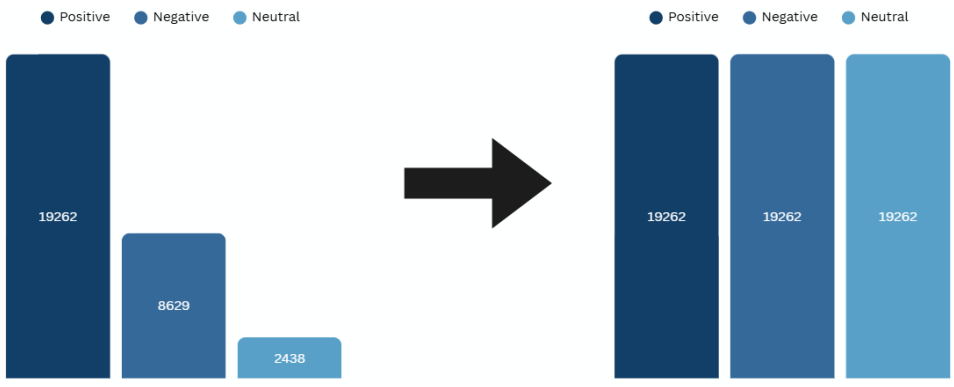

Fig.7. Random oversampling applied to sentiment class distribution

- Data balancing using random oversampling: To overcome the problem of class imbalance, the random oversampling (ROS) technique is applied. ROS works by randomly duplicating samples from minority classes (negative and neutral) until the number is equal to the number of samples in the majority (positive) class as shown in Fig. 7. The goal of this stage is to create a balanced training dataset, thus ensuring that the IndoBERT model does not tend to predict the majority class and can recognize the patterns of each sentiment class fairly. In addition to accuracy, macro-average F1-score and confusion matrix analysis were used to evaluate

classification performance in the presence of class imbalance.

- IndoBERT model classification performance analysis: The model was evaluated using four different training, validation, and test data sharing ratio scenarios to analyze the impact of the proportion of training data on performance. The results of the evaluation show a consistent trend of performance improvement along with the increase in the amount of training data, as summarized in Table 9.

Table 9. Performance comparison of the IndoBERT model at different data partition ratios

Ratio	IndoBERT			
	Accuracy	Precision	Recall	F1-Score
60:20:20	91,54%	92,10%	91,53%	91,58%
70:15:15	92,90%	93,32%	92,95%	92,99%
80:10:10	94,64%	94,78%	94,64%	94,63%
90:05:05	94,78%	95,04%	94,79%	94,81%

The highest performance was achieved at a 90:5:5 ratio, which indicates that a larger proportion of trained data has a positive impact on the model's generalization capabilities. In this configuration, the model shows excellent performance with an accuracy of 94.78% and an f1-score of 94.81%. To ensure the reliability, stability, and generalization of results, the model was validated using a stratified 5-fold cross-validation technique. This procedure partitions the dataset into five subsets while preserving class distribution in each fold. In each iteration, four folds were used for training and one-fold for validation. Oversampling was applied exclusively within the training portion of each fold to address class imbalance, while the validation fold retained the original data distribution. This design prevents data leakage and ensures that no synthetic or duplicated samples appear in validation partitions. The averaged results across folds, shown in Table 10, demonstrate consistent performance, indicating that the reported accuracy does not depend on a single data split.

Table 10. Performance analysis of IndoBERT with 5-fold cross-validation

Fold	Accuracy	Precision	Recall	F1-Score
1	0,9431	0,9456	0,9431	0,9432
2	0,9324	0,9385	0,9324	0,9329
3	0,9388	0,9425	0,9388	0,9387
4	0,9460	0,9486	0,9460	0,9460
5	0,9372	0,9420	0,9372	0,9375

4.5. Topic–Sentiment Intersection Analysis

To address the relationship between dominant topics and public sentiment, we merged the BERTopic results with the IndoBERT sentiment predictions at the headline level. This integration allows us to examine the emotional distribution within each dominant topic. As shown in Table 11, most dominant topics are associated primarily with neutral sentiment. For instance, Topic 1 exhibits 94.39% neutral sentiment, while Topic 4 shows 96.00% neutral sentiment. This indicates that political news headlines tend to maintain an informative and objective tone rather than expressing explicit evaluative polarity.

Table 11. Distribution of the ten most dominant topics identified by BERTopic

Topic	**Main Topic**	**Negative (%)**	**Neutral (%)**	**Positive (%)**
1	*Pendidikan politik dan demokrasi*	10.30	71.30	18.40
2	*Peran lembaga legislatif dan bamsoet*	3.04	94.39	2.57
3	*Kontestasi pilkada jakarta dan tokoh politik*	4.28	86.65	9.07
4	*Kunjungan paus fransiskus*	7.75	91.09	1.16
5	*Polemik capres dan capres alternatif*	1.20	96.00	2.80
6	*Politik internasional dan tokoh global*	2.07	95.85	2.07
7	*Inisiatif MPR dalam peningkatan sektor strategis*	0.00	60.68	39.32
8	*Isu pekerja migran dan imigrasi ilegal*	0.00	70.26	29.74
9	*Kerja sama bilateral dan pertemuan internasional*	1.79	87.50	10.71
10	*Seruan pemilu damai*	9.01	89.64	1.35

4.6. Discussion

This study demonstrates that the integration of BERTopic and IndoBERT provides an effective framework for discourse analysis of Indonesian political news headlines. BERTopic achieved a coherence score of 0.5145, indicating a

strong ability to cluster short and formal texts into semantically meaningful topics. This result substantially exceeds the coherence score reported in [27], which obtained 0.09 on election headline data, and is competitive with [28], which achieved 0.53 using informal social media texts. These findings suggest that BERTopic performs particularly well in structured news contexts, where semantic consistency across headlines supports more stable topic formation. The use of contextual embeddings further enhances the model's capacity to preserve semantic relationships even in short textual units such as headlines.

In terms of sentiment classification, IndoBERT achieved an accuracy of 94.78% and an F1-score of 94.81%, outperforming previous studies such as [29], which reported 90% accuracy in hoax detection, and [30], which achieved 80% accuracy in sentiment analysis of presidential candidates on Twitter. These results confirm the robustness of IndoBERT for formal Indonesian texts, particularly when combined with data balancing techniques. The high performance indicates that transformer-based architectures adapted to local languages are highly reliable for sentiment classification tasks in political news domains.

Beyond individual model performance, the intersection analysis between dominant topics and sentiment provides deeper insight into political news discourse. The analysis reveals that neutral sentiment predominates across most dominant topics. This pattern suggests that Indonesian political news headlines tend to adopt an informative and objective framing style rather than explicitly evaluative or emotionally charged language. While certain topics display relatively higher proportions of positive or negative sentiment, neutrality remains the prevailing tone across the majority of thematic clusters. This finding aligns with journalistic conventions that emphasize informational delivery over opinion expression in headline construction.

Interestingly, the initial lexicon-based labeling approach yielded a substantially higher proportion of positive sentiment compared to the IndoBERT model. However, after applying the IndoBERT model, the overall sentiment distribution shifted toward neutrality. This discrepancy highlights fundamental methodological differences between dictionary-based and transformer-based approaches. Lexicon-based methods assign polarity at the word level, which may overestimate sentiment intensity by ignoring contextual nuance. In contrast, transformer-based models interpret sentiment through contextual semantic representation, allowing for more nuanced and conservative classification. Consequently, IndoBERT appears to provide a more context-aware interpretation of sentiment, particularly suitable for short and formal texts such as political news headlines.

From a broader perspective, this study addresses several research gaps. Regarding the data scale, the analysis covers 30,329 headlines over a full year, offering broader and more representative coverage than prior studies that relied on smaller datasets or different domains. Regarding the methodological gap, this research provides strong evidence that the integration of BERTopic and IndoBERT is effective for analyzing short and formal political news headlines. Addressing the knowledge gap, the mapping of dominant topics and their associated sentiment distributions contributes new insights into thematic patterns and sentiment framing within Indonesian political news discourse.

Practically, the proposed framework can assist media practitioners, political analysts, and policymakers in monitoring issue trends and sentiment framing efficiently. Theoretically, this study reinforces the growing body of evidence that transformer-based language models tailored to local languages, when combined with advanced topic modeling techniques such as BERTopic, constitute a reliable and scalable approach for in-depth media discourse analysis.

5. Conclusions

This study successfully implemented and evaluated the combination of BERTopic and IndoBERT for topic modeling and sentiment analysis on Indonesian political news headlines. BERTopic achieved coherence scores ranging from 0.3227 to 0.7095 across different topics, with an average of 0.5145 for the top dominant topics. Compared to the LDA baseline, which produced a coherence score of 0.3252 on the same dataset, BERTopic demonstrated superior semantic clustering performance. Furthermore, the IndoBERT model achieved an accuracy of 94.78% with a macro-F1 score of 94.81%, indicating strong alignment with lexicon-generated sentiment labels. The evaluation was conducted using a stratified train–validation–test split to reduce potential data imbalance effects.

In terms of substantive findings, the integration analysis reveals that Indonesian political news discourse in 2024 was predominantly neutral across dominant thematic clusters. Topics related to legislative roles, political contestation, and international affairs exhibited overwhelmingly neutral sentiment, suggesting an informative rather than overtly polarized reporting style. Positive sentiment was more pronounced in governance-oriented topics, such as strategic sector initiatives and migrant worker issues, indicating supportive framing in policy-driven narratives. Notably, none of the dominant topics were characterized by strong negative polarity, implying that headline-level political coverage during 2024 remained largely balanced and event-contextual rather than emotionally charged.

This study analyzes the dataset as a static corpus and does not incorporate temporal topic evolution. Considering that news data is inherently dynamic, future research may explore Dynamic BERTopic or time-aware topic modeling approaches to capture emerging themes and topic transitions. Although the dataset spans a full year, the analysis was conducted on the entire corpus to provide a structural mapping of dominant issues rather than examining monthly or event-driven fluctuations. Future studies may segment the dataset temporally to analyze sentiment trends, event-driven spikes, and discourse evolution across specific political periods. In addition, this study relies solely on data from detik.com, which has a distinctive editorial style characterized by concise and fast-paced headlines. Therefore, the findings may not

fully represent other Indonesian news portals with different editorial approaches. Moreover, this study primarily relies on automated evaluation metrics, including the coherence score for topic modeling and classification metrics for sentiment analysis. We did not conduct human evaluation of topic interpretability or expert validation of sentiment labels. Furthermore, the sentiment classification performance was evaluated against lexicon-generated labels rather than manually annotated ground-truth data, which may limit the reliability and validity of the reported accuracy. Given that news headlines may contain nuanced, neutral, or context-dependent expressions, future research should incorporate expert-based validation or inter-annotator agreement analysis to further strengthen the reliability and interpretability of the findings.

Despite these limitations, the findings indicate that the integration of transformer-based topic modeling and sentiment classification provides a structured and scalable framework for analyzing large-scale Indonesian news discourse. Another limitation is that the use of pretrained transformer models may introduce implicit biases originating from the large-scale web corpora used during pretraining. IndoBERT, trained on general Indonesian web text, may reflect political, cultural, or regional biases embedded in its source data. Therefore, the sentiment classifications should be interpreted as algorithmic predictions rather than definitive judgments of media stance. Future research may incorporate bias auditing or cross-model comparison to assess potential algorithmic bias.

All the Declarations and Statements

Author Contributions Statement

Bagas Yana Prayoga – Conceptualization, Methodology, Data Curation, Software Implementation, Model Training, Validation, Performance Evaluation, and Writing: Conducted the main research, collected and preprocessed the data, implemented the BERTopic and IndoBERT models, performed experiments, and wrote the manuscript.
Qurrotul Aini – Supervision, Methodology Review, and Writing Review & Editing: Provided research guidance, reviewed the methodology, and contributed to manuscript revision.
Fitroh Fitroh – Supervision, Validation, and Writing Review & Editing: Supervised the research process, validated the results, and contributed to manuscript revision.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research received no external funding.

Data Availability Statement

This study analyzed publicly available data collected from the detik.com news portal. The dataset used in this study is available from the corresponding author upon reasonable request.

Ethical Declarations

This study did not involve human participants, animals, or any confidential data. All data used in this research were publicly available news headlines collected from online media sources.

Acknowledgments

The authors thank the supervisors and the Department of Information Systems, UIN Syarif Hidayatullah Jakarta, for their guidance and support during the research process.

Declaration of Generative AI in Scholarly Writing

The authors used generative AI tools to assist with language editing and readability improvement during the manuscript preparation. All content was reviewed and revised by the authors, who take full responsibility for the final version of the manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
NLP - Natural Language Processing
DL - Deep Learning
BERT - Bidirectional Encoder Representations from Transformers

BERTopic - Bidirectional Encoder Representations from Transformers for Topic Modeling
IndoBERT - Indonesian Bidirectional Encoder Representations from Transformers
ROS - Random Oversampling

References

[1] A. D. Riyanto, "Hootsuite (We Are Social): Data Report Digital Indonesia 2024." Accessed: Apr. 24, 2025. [Online]. Available: https://andi.link/hootsuite-we-are-social-data-digital-indonesia-2024/

[2] A. Yaman, B. Sartono, and A. M. Soleh, "Topic modeling in fertilizer-related patent documents in Indonesia based on latent Dirichlet allocation," (in Indonesian), *Berk. Ilmu Perpust. dan Inf.*, vol. 17, no. 2, pp. 168–180, 2021, doi: 10.22146/bip.v17i2.2147.

[3] Z. Alhaq, A. Mustopa, S. Mulyatun, and J. D. Santoso, "Application of support vector machine method for Twitter user sentiment analysis," (in Indonesian), *J. Inf. Syst. Manag.*, vol. 3, no. 1, pp. 16–21, 2021, doi: 10.24076/joism.2021v3i2.558.

[4] R. Egger and J. Yu, "A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts," Front. Sociol., vol. 7, p. 886498, 2022, doi: 10.3389/fsoc.2022.886498.

[5] D. Hendry *et al.*, "Topic modeling for customer service chats," in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, 2021, pp. 1–6, doi: 10.1109/ICACSIS53237.2021.9631322.*

[6] Y. Liu, "Comparison of LDA and BERTopic in news topic modeling: A case study of the New York Times' reports on China," *Pacific Int. J.*, vol. 7, no. 3, pp. 47–51, 2024, doi: 10.55014/pij.v7i3.616.

[7] B. Wilie *et al.*, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," arXiv preprint arXiv:2009.05387, 2020, [Online]. Available: http://arxiv.org/abs/2009.05387

[8] Fransiscus and A. S. Girsang, "Sentiment analysis of COVID-19 public activity restriction (PPKM) impact using BERT method," *Int. J. Eng. Trends Technol.*, vol. 70, no. 12, pp. 281–288, 2022, doi: 10.14445/22315381/IJETT-V70I12P226.

[9] W. Nurfitri and A. Chowanda, "Sentiment analysis on COVID-19 positive cases based on media coverage in Indonesia using IndoBERT," (in Indonesian), *Progresif J. Ilm. Komput.*, vol. 20, no. 1, pp. 580–593, 2024, doi: 10.35889/progresif.v20i1.1897.

[10] N. Mahfudiyah and A. Alamsyah, "Understanding user perception of ride-hailing services sentiment analysis and topic modelling using IndoBERT and BERTopic," in 2023 International Conference on Digital Business and Technology Management (ICONDBTM), IEEE, 2023, pp. 1–6, doi: 10.1109/ICONDBTM59210.2023.10327320.

[11] E. Effendi, I. Sartika, N. L. T. Purba, and S. Ritonga, "Writing headlines and leads for news and features," (in Indonesian), *J. Pendidik. dan Konseling*, vol. 5, no. 2, pp. 4680–4683, 2023, doi: 10.31004/jpdk.v5i2.14206.

[12] E. C. D. A. Nanda, "Apa Konten Berita yang Paling Banyak Diakses Warganet 2024?" (in Indonesian), Goodstats, 2024. [Online]. Available: https://goodstats.id/article/apa-konten-berita-yang-paling-banyak-diakses-warganet-pada-2024-B5hDY. Accessed: Jan. 16, 2025.

[13] N. Newman, R. Fletcher, C. T. Robertson, A. R. Arguedas, and R. K. Nielsen, "Reuters Institute Digital News Report 2024," 2024. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024

[14] M. Han, I. Canli, J. Shah, X. Zhang, I. G. Dino, and S. Kalkan, "Perspectives of machine learning and natural language processing on characterizing positive energy districts," *Buildings*, vol. 14, no. 2, pp. 1–28, 2024, doi: 10.3390/buildings14020371.

[15] G. S. Mahendra *et al.*, *Tren Teknologi AI: Pengantar, Teori, dan Contoh Penerapan Artificial Intelligence di Berbagai Bidang (in Indonesian). PT. Sonpedia Publishing Indonesia, 2024.*

[16] D. L. C. Pardede and M. A. I. Waskita, "Topic modeling analysis for reviews about PeduliLindungi," (in Indonesian), *J. Ilm. Inform. Komput.*, vol. 28, no. 1, pp. 17–26, 2023, doi: 10.35760/ik.2023.v28i1.7925.

[17] S. Alaparthi and M. Mishra, "Bidirectional encoder representations from transformers (BERT): A sentiment analysis odyssey," *arXiv preprint arXiv:2007.01127, 2020*, [Online]. Available: https://arxiv.org/abs/2007.01127

[18] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794, 2022*, [Online]. Available: https://arxiv.org/abs/2203.05794

[19] B. Liu, *Sentiment Analysis and Opinion Mining*. Chicago: Morgan & Claypool Publisher, 2020. doi: 10.1017/9781108639286.

[20] S. B. Panuntun, D. Krismawati, S. Pramana, and E. T. Astuti, "Analysis of telemedicine news coverage in Indonesia: Sentiment, NER, topic modeling, and social network approaches to understanding issues and perceptions," (in Indonesian), *Indones. Heal. Inf. Manag. J.*, vol. 11, no. 1, pp. 56–67, 2023, doi: 10.47007/inohim.v11i1.500.

[21] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020), 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.

[22] A. Simanjuntak *et al.*, "Study and analysis of IndoBERT hyperparameter tuning in fake news detection," (in Indonesian), *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 13, no. 1, pp. 60–67, 2024, doi: 10.22146/jnteti.v13i1.8532.

[23] S. Sazzed and S. Jayarathna, "SSentiA: A self-supervised sentiment analyzer for classification from unlabeled data," *Mach. Learn. with Appl.*, vol. 4, pp. 1–14, 2021, doi: 10.1016/j.mlwa.2021.100026.

[24] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in 2020 11th International Conference on Information and Communication Systems (ICICS), IEEE, 2020, pp. 243–248, doi: 10.1109/ICICS49469.2020.239556.

[25] Y. Yunitasari and A. R. Putera, "Analysis of public sentiment on Twitter regarding the COVID-19 pandemic," (in Indonesian), *Smatika J.*, vol. 11, no. 01, pp. 22–26, 2021, doi: 10.32664/smatika.v11i01.520.

[26] B. P. E. Wijaya, D. P. Hostiadi, and P. D. W. Ayu, "Sentiment analysis of Instagram comments for cyberbullying detection using gradient boosting models," (in Indonesian) in *Seminar Hasil Penelitian Informatika dan Komputer (SPINTER), Institut Teknologi dan Bisnis STIKOM Bali*, 2025, pp. 1093–1098.

[27] D. Aryani, I. Lucia Kharisma, A. Sujjada, and K. Kamdan, "Topic modeling of the 2024 election using the BERTopic method on detik.com news articles," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 9, no. 2, pp. 171–180, 2024, doi: 10.25139/inform.v9i2.8429.

[28] I. N. Nugraha and E. Utami, "Evaluation of creative economy and tourism industry trends based on LDA analysis with BERTopic," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 15, no. 2, pp. 182–195, 2024, doi: 10.31849/digitalzone.v15i2.23796.
[29] A. Rahmawati, A. Alamsyah, and A. Romadhony, "Hoax news detection analysis using IndoBERT deep learning methodology," in *2022 10th International Conference on Information and Communication Technology (ICoICT)*, 2022, pp. 368–373. doi: 10.1109/ICoICT55009.2022.9914902.
[30] P. Sayarizki, Hasmawati, and H. Nurrahmi, "Implementation of IndoBERT for sentiment analysis of indonesian presidential candidates," *J. Comput.*, vol. 9, no. 2, pp. 61–72, 2024, doi: 10.34818/indojc.2024.9.2.934.

Authors' Profiles

Bagas Yana Prayoga was born in South Jakarta, Indonesia. He is currently pursuing the Bachelor of Science degree in Information Systems at UIN Syarif Hidayatullah Jakarta, Indonesia. His areas of interest include software quality assurance, user interface and user experience (UI/UX) principles, usability testing, and the application of data science and natural language processing (NLP) for text analytics.

Qurrotul Aini is currently working as an Assistant Professor and Head of Department of Information Systems, Faculty Science and Technology UIN Syarif Hidayatullah Jakarta Indonesia. She received her Doctor of Electrical Engineering degree from Institut Teknologi Sepuluh Nopember Surabaya Indonesia. She is a Member IEEE since 2013. Her research interests include text mining, intelligence computation, and social computing.

Fitroh holds a Bachelor's degree in Computer Science and a Master's degree in Information Systems Management from Budi Luhur University, Indonesia. She is currently pursuing a doctotal degree in the Computer Science Program at IPB University, Indonesia. She is also a lecturer in the Information Systems Program, Faculty of Science and Technology, UIN Jakarta. Her courses include Audit Information Systems, Enterprise Architecture, and Management Information Systems.

How to cite this paper

Bagas Yana Prayoga, Qurrotul Aini, Fitroh Fitroh, "Topic Modeling and Sentiment Analysis on News Headlines Using BERTopic and IndoBERT Models", International Journal of Intelligent Systems and Applications(IJISA), Vol.18, No.4, pp.226-238, 2026. DOI:10.5815/ijisa.2026.04.12