Implementation of Text Mining for Evaluating the Relevance Between News Headlines and Content on a Web-Based Platform

Desak Gede Inten Purnawati [1]*, Desy Purnami Singgih Putri [2]*, I Nyoman Piarsa [3]*
* Teknologi Informasi, Universitas Udayana
intenpurnawati@student.unud.ac.id [1], desysinggihputri@unud.ac.id [2], manpits@unud.ac.id [3]

Article Info

Article history:

Received 2025-06-03
Revised 2025-06-26
Accepted 2025-07-09

Keyword:

News,
Text Mining,
LSTM,
IndoBERT,
Similarity.

ABSTRACT

Technological advancements in the era of the Industrial Revolution 4.0 have significantly transformed how society accesses and consumes information, particularly through online news portals. This study aims to analyze the relevance between news headlines and article content on Indonesian online news platforms by employing text mining techniques and similarity checking methods. To enhance the accuracy of relevance assessment, this research utilizes two deep learning-based modeling algorithms: Long Short-Term Memory (LSTM) and IndoBERT. The data was collected from three leading Indonesian news portals detik.com, kompas.com, and suara.com with a total of 52,242 articles from the entertainment and national news categories, gathered between July 1 and September 30, 2024. The dataset includes attributes such as headline, category, publication date, author, article URL, and news content. The research process consists of several stages, including data collection through web scraping, data pre-processing (which involves cleaning the category, author, and content columns), content summarization, text similarity calculation, and data labeling into three classes (relevan, berlebihan, and nonrelevan). Evaluation results show that the IndoBERT model outperforms LSTM, achieving the best performance with a training accuracy of 0.9048 and a training loss of 0.2514, as well as a validation accuracy of 0.8604 and a validation loss of 0.4039. These findings demonstrate that IndoBERT is effective in assessing the coherence between news headlines and content in today's digital age.

I. INTRODUCTION

The rapid advancement of technology in the era of the Fourth Industrial Revolution has brought significant impacts on various aspects of life, including the dissemination of information and human communication. A concrete example of this development is the substantial increase in internet usage in Indonesia. According to APJII (Asosiasi Penyelenggara Jasa Internet Indonesia) in 2024, the number of internet users in Indonesia reached 221,563,479 people, nearly 80% of the total population of 278,696,200 in 2023 [1].

The internet has become an inseparable part of various human activities across different fields, from work and business to education. Its presence facilitates daily life by providing faster and more efficient access to information and enabling easier communication in the digital era [2].

One tangible manifestation of internet utilization is online news portals. These portals serve as easily accessible digital platforms stored on the internet. Online news portals provide timely, accurate, and relevant information. Through easy accessibility, people can now follow news developments across various fields without being hindered by geographical limitations.

In 2021, the Reuters Institute Digital News Report conducted research with Indonesia as one of its focal points [3]. The study highlighted that online media and social media are the most popular sources for accessing information among the public. Television and radio remain primary choices for those without internet access. The research also explored the public's trust in the news from different media channels they consume. It revealed that trust levels are influenced by the credibility of the media or portal delivering the news. One significant challenge faced by online news portals is the business strategy employed by some media, such as creating misleading headlines known as clickbait. This strategy is deemed effective as reader attention is seen as a valuable commodity that can be converted into revenue for the media [4].

In addition to clickbait, another emerging problem in the digital information landscape is the spread of fake news. Fake news refers to information that is deliberately created and has been proven to be false. Two main characteristics that distinguish fake news are authenticity and intentionality. With these two characteristics, fake news can be differentiated from other similar concepts. For instance, if the authenticity of the information has not been verified and the intent is unclear, it is referred to as a rumor. Meanwhile, if the information is false but there is no malicious intent, it is categorized as misinformation [5]. In the context of online media, the phenomena of fake news and clickbait are often interrelated and can blur public trust in digital media.

Based on this background, this research aims to analyze the relevance between headlines and content on online news portals. A similar study was previously conducted by Ahmadi & Chowanda in 2023 [6], albeit with a broad approach and lack of specificity. Moreover, the study utilized various algorithms such as Random Forest, IndoBERT, LSTM, Logistic Regression, SVM, Naive Bayes, and Decision Tree. Among these algorithms, LSTM and IndoBERT showed the best predictive values. Therefore, in contrast to previous studies that employed broader and less targeted approaches, this research offers a more focused contribution by utilizing Indonesian-language news content from selected portals and concentrating on two specific categories. This enables a deeper contextual understanding of headline-content relevance and supports the development of NLP models that are better aligned with the linguistic characteristics of local media. This study employs Long Short-Term Memory (LSTM) and IndoBERT due to their proven effectiveness in handling sequential text data and capturing semantic meaning, particularly in the context of the Indonesian language. While other models like CNN-LSTM, BiLSTM, or RoBERTa Indo exist, LSTM and IndoBERT were chosen for their balance of performance, interpretability, and suitability to the research scope without introducing unnecessary complexity.

The selection of categories for this study is driven by various considerations, particularly their relevance and prominence in contemporary news consumption. The entertainment category was chosen due to its propensity for attracting significant reader attention, often characterized by sensationalized or clickbait headlines that do not align with the actual content. This aligns with findings from the Reuters Institute Digital News Report, which indicates that entertainment and celebrity news, although still relevant to a substantial portion of the audience (56%), frequently rely on exaggerated or misleading content to maintain engagement. Such characteristics make it particularly suitable for examining headline-content relevance [7].

The national category, on the other hand, was selected for its broad scope, encompassing a diverse array of sub-categories such as finance, politics, and social issues. Given its far-reaching impact, this category serves as a crucial avenue for assessing the diversity of news content that resonates with the general public. According to the Reuters Institute Digital News Report, local news (62%) emerges as the most prominent categories that capture readers' interest, emphasizing their significance in shaping public discourse and reflecting societal concerns of wide-ranging importance [7].

Thus, this research is expected to provide insights into headline-content relevance in both categories and help mitigate the phenomenon of clickbait and fake news commonly found in online news. In addition, this study compares the performance of two deep learning-based approaches, namely Long Short-Term Memory (LSTM) and IndoBERT, in evaluating the alignment between headlines and news content through semantic similarity measurement. This approach provides a novel contribution, as it remains underexplored in local studies, particularly those using the Indonesian language in the context of semantic relevance evaluation for news articles.

II. METHODOLOGY

This study follows six stages. The initial stages include data collecting, data cleaning, data summarization, similarity check, and classification. Figure 1 will illustrate and explain the workflow of the stages undertaken in this study.

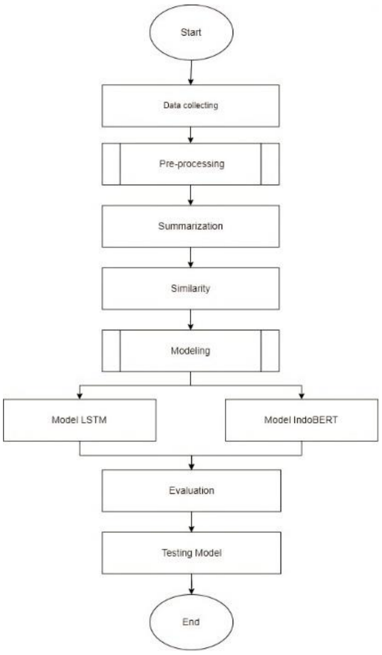


Figure 1. Research Process Flow

A. Data Collecting

Data collecting is the process of gathering and compiling data from various sources, then making it available for analysis or storage [8]. Its main goal is to collect as much reliable information and data as possible, which can then be analyzed to support important business decision-making. The data collection process in this study involves scraping news articles from four major Indonesian online news portals: detik.com, kompas.com, kapanlagi.com, and suara.com. The dataset will consist of news articles published within the time range of July 1 to September 30, 2024. These websites were selected due to their high traffic and wide coverage across various news categories, ensuring a diverse and representative dataset for analysis.

B. Data Preprocessing

The data preprocessing stage involves cleaning, standardizing, and labeling the text in both the title and content columns to improve the quality of the dataset and prepare it for classification. This process includes removing introductory phrases commonly found at the beginning of news articles, converting date formats from numerical strings into a more readable format (e.g., “17 Agustus 2024”), and eliminating promotional or advertisement content that may appear within the body of the news articles. Additionally, newline characters and other unnecessary symbols are removed to ensure the text is clean and consistent. In this stage, we also label the data according to predefined categories for classification, namely relevan, berlebihan, nonrelevan. The labeling is done manually, where each news article is carefully examined, and its title and content are classified based on the relevance and alignment of the information, ensuring that each article is properly categorized for accurate analysis and model training. Labeling was performed manually by two trained annotators who independently reviewed each news article. Annotators assessed the relevance between the title and content based on predefined guidelines, and the inter-annotator agreement, measured using Cohen’s Kappa, reached 0.82, indicating substantial agreement.

C. Data Summarization

After the data has undergone preprocessing with various treatments specific to each news portal, the next step is summarization. This process aims to condense the sentences in the content column while retaining the essential information. Text summarization is the process of condensing or simplifying a body of text while preserving the essential information from the original text [9]. In this study, extractive text summarization is employed, utilizing the IndoBERT-base-p1 model. This approach selects the most relevant sentences from the content to ensure that the summary captures the key points while reducing text length, making it more efficient for further analysis and classification. The extractive summarization method used in this study leverages sentence weighting by identifying words and phrases in the text based on their frequency [10]. Specifically, we adopted an extractive method using transformer-based sentence embeddings from IndoBERT, rather than traditional statistical techniques such as LSA or TextRank, to prioritize semantic relevance in sentence selection.

D. Similarity Check

The similarity assessment stage is conducted to evaluate the degree of alignment between the headline and the content of the collected news articles. In this study, the cosine similarity method is employed to measure the relevance and semantic similarity between these two components. Cosine similarity functions by normalizing the length of vectors and comparing sentence A and sentence B to determine their directional similarity [11]. This technique computes the cosine of the angle between the vector representations of the headline and the article content, yielding a numerical score that reflects the extent of their alignment. A higher cosine similarity value indicates a stronger correspondence between the headline and the content, whereas a lower score suggests a potential mismatch, thus facilitating a more precise analysis of their semantic relationship.

E. Modeling

The modeling process in this study is performed using Long Short-Term Memory (LSTM) and IndoBERT models to classify the relevance between the title and content of news articles. The steps in the modeling phase are as follows:

1) *Testing LSTM Model*: The first stage involves training the LSTM model, which is a type of recurrent neural network (RNN) known for capturing temporal dependencies in sequential data. LSTM is particularly effective for understanding the relationships between words in the title and content of the articles [12]. In this study, LSTM is trained on a labeled dataset where each article is classified as relevan, berlebihan, or nonrelevan based on the similarity between its title and content. This enables the model to learn from the contextual patterns and predict the relevance of unseen articles.

2) *Testing IndoBERT Model:* IndoBERT is a monolingual BERT-based language model specifically developed for the Indonesian language, trained using the Masked Language Model (MLM) approach within the Huggingface framework [13]. In the second stage of this study, IndoBERT is utilized as a pre-trained model built on the BERT architecture, designed to understand the semantic meaning and contextual nuances of the Indonesian language. The model is fine-tuned on the collected dataset to better capture the meaning and context of the news articles [14]. By leveraging IndoBERT’s ability to comprehend the intricacies of Indonesian, the model is used to evaluate the

alignment between news headlines and article content. It is fine-tuned to classify news articles into three predefined categories: relevan, berlebihan, and nonrelevan. This process aims to enhance the accuracy of relevance prediction, ensuring that the classification outcomes accurately reflect the relationship between the headline and the content in a contextual and meaningful manner.

F. Evaluation

The evaluation stage of this study is conducted to assess the performance of the LSTM and IndoBERT models in classifying the relevance between news headlines and article content. The evaluation employs key tools and metrics such as the evaluation matrix and ROC curve to ensure the accuracy of relevance classification. The evaluation matrix is an essential tool used to measure model performance in classification or prediction tasks. The choice of appropriate evaluation metrics strongly depends on the type of data and the objective of the developed model. The proposed approach can be evaluated using standard evaluation metrics. Referring to previous studies, three primary metrics are commonly used: precision, recall, and F1-score [15].

Meanwhile, the ROC curve is an analytical method illustrated in graphical form and used to evaluate the performance of binary diagnostic classification methods [16]. In this study, the ROC curve is applied to observe the balance between the true positive rate and false positive rate across various classification thresholds, with the area under the curve (AUC) serving as a global indicator of model performance. A higher AUC indicates that the model is more effective at distinguishing between relevan, berlebihan, and nonrelevan news classes [17].

G. Testing Model

The testing stage involves implementing the best-performing model from the training phase to evaluate its real-world effectiveness in classifying the relevance between the title and content of news articles. After training both the LSTM and IndoBERT models, we select the model that produces the highest accuracy and best results in the evaluation metrics, such as precision, recall, and F1-score. This chosen model is then tested on a separate dataset that it has not seen during the training phase to simulate its performance on unseen data. The model's ability to generalize to new articles is crucial for ensuring its practical applicability. By applying the trained model to this test data, we can confirm whether it maintains high accuracy and performs consistently, ensuring that it is ready for deployment and can effectively classify relevance in real-world scenarios.

III. RESULT AND DISCUSSION

The results reveal the most accurate classification outcomes from the LSTM and IndoBERT models, along with the performance metrics (accuracy, loss) generated from these models based on the analyzed news classification data.

A. Data Collecting

The data used in this study was obtained through web scraping from four news portals: detik.com, suara.com, kapanlagi, and kompas.com, focusing on two predefined categories: national news and entertainment. Data collection was conducted between July 1 and September 30, 2024

TABEL I
TOTAL DATA COLLECTED EACH NEWS PORTAL

News Portal	Total Data
Detik.com	25.218
Kompas.com	8.742
Suara.com	15.082
Kapanlagi.com	3.200

The data is distributed across four news portals: Detik.com with 25,218 entries, Kompas.com with 8,742 entries, Suara.com with 15,082 entries, and Kapanlagi.com with 3,200 entries.

B. Data Preprocessing

The preprocessing phase will clean and standardize both the content and titles, removing irrelevant information. After that, data labeling will categorize each piece into one of three labels then, preparing the data for classification model training.

1) *Preprocessing Data*: Pre-processing is carried out to prepare the data before further processing. After the data collection process, which resulted in a total of 52,242 articles, the next step is pre-processing. Each dataset obtained from the four news portals undergoes different pre-processing treatments. In general, this study focuses primarily on text cleansing, such as removing advertisements embedded within the body of the news content. After the text cleansing stage, the process continues with data labeling, which consists of three classes (relevan, berlebihan, nonrelevan)

TABEL II
EXAMPLE DATA BEFORE AND AFTER PREPROCESSING

Before Preprocessing	After Preprocessing
Polisi mengamankan seorang kakek berinisial AI (60), yang sempat diikat warga setelah kedapatan hendak mencuri di rumah warga di Kalideres, Jakarta Barat. Kakek AI kini sudah ditetapkan sebagai tersangka dan ditahan. Dari foto yang diterima detikcom, Rabu (24/7/2024), Kakek AI kini mengenakan baju tahanan berwarna oranye. ADVERTISEMENT	Polisi mengamankan seorang kakek berinisial AI (60), yang sempat diikat warga setelah kedapatan hendak mencuri di rumah warga di Kalideres, Jakarta Barat. Kakek AI kini sudah ditetapkan sebagai tersangka dan ditahan. Dari foto yang diterima detikcom, Rabu 24 Juli 2024, Kakek AI kini mengenakan baju tahanan berwarna oranye. Video AI ditangkap warga

SCROLL TO CONTINUE WITH CONTENT Video AI ditangkap warga viral di media sosial. Dalam video terlihat AI diikat di sebuah pos satpam dan sempat diamuk massa....	viral di media sosial. Dalam video terlihat AI diikat di sebuah pos satpam dan sempat diamuk massa....

The preprocessing results show the transformation of raw text data by removing irrelevant content like advertisements and extraneous words. For example, phrases such as "ADVERTISEMENT" and "SCROLL TO CONTINUE WITH CONTENT" were removed. The cleaned text is now more focused, with unnecessary terms like names and dates standardized or removed, improving the quality for further analysis or model input.

2) *Data Labeling*: The labeling process was carried out manually by two annotators the researcher and a peer by reading each news item and comparing the headline to the content. The data were categorized into three classes: relevan, berlebihan, and nonrelevan. The labeling criteria were as follows:

Relevan: The headline directly reflects the content without causing misinterpretation.

Berlebihan: The headline is related to the content but contains dramatization or hyperbole.

Nonrelevan: The headline does not reflect the content and may mislead the reader.

To ensure consistency, the two annotators discussed any items that raised potential disagreement before assigning a final label. To measure the reliability of the labeling, inter-annotator agreement was calculated using Cohen’s Kappa, which yielded a score of 0.82, indicating substantial agreement. This labeling step was completed prior to the training and testing of the classification models.

TABEL III
LABELING DATA EXAMPLE

Title	Content	Class
Dijamin JKN, Pensiunan ASN di Toraja Utara Jalani Rawat Jalan Gratis	Pemerintah melalui program JKN menghadirkan layanan kesehatan komprehensif bagi masyarakat, termasuk pensiunan Aparatur Sipil Negara (ASN). Manfaat program jaminan kesehatan ini telah dirasakan, salah satunya oleh PD Sarungu (76). Sehingga ia pun harus melakukan kontrol rutin ke rumah sakit. " Awal mula saya terkena penyakit ini pada beberapa tahun lalu, saat itu saya merasa nyeri di kepala saya terus setelah saya periksa ke rumah sakit ternyata saya terkenan tekanan darah tinggi yang membuat saya harus menjalani rawat inap selama 6 hari di rumah sakit," jelas Sarungu. Secara pribadi dirinya tidak mengalami kendala yang berarti dalam hal layanan kesehatan yang didapat. " Sarungu menambahkan selama perawatan tidak berhenti mengonsumsi obat atas saran dokter. Menurutnya program JKN tidak hanya bisa dimanfaatkan oleh segmen pensiunan. Selama menjalani pengobatan di rumah sakit, dirinya banyak melihat pasien yang memanfaatkan program JKN untuk pengobatan mereka. Sarungu menilai menjadi peserta aktif JKN adalah pilihan yang bijak.	berlebihan
Prabowo Sarankan Otorita Prioritaskan Bangun Gedung Lembaga Negara di IKN	Menhan sekaligus Presiden terpilih Prabowo Subianto bicara terkait prioritas pembangunan di Ibu Kota Negara (IKN) Nusantara. Kalau menurut saya ibu kota negara intinya adalah pusat pemerintahan, berarti yudiaktif, eksekutif dan legislatif, berarti gedung MPR/DPR menjadi prioritas dengan perumahan anggota DPR dan MPR dan ruangan kantornya dan juga MA, MK juga sangat mendesak menurut saya," kata Prabowo saat memberikan keterangan di sidang kabinet perdana di Istana Garuda, Senin 12 Agustus 2024. Prabowo juga berharap agar Otorita IKN segera membuat sayembara desain untuk pembangunan gedung-gedung prioritas tersebut. Komentar saya tentunya kepada otorita, terutama terima kasih atas jeri payah dan prestasi yang sudah dicapai, saya sebagai anak bangsa melihat, saya juga cukup bangga nuansa budaya bangsa kita sangat kuat dan ini juga membesarkan hati ini juga membuat saya ingin cepat beroperasi di sini Pak," kata Prabowo yang disambut tepuk tangan jajaran kabinet. Prabowo lantas permisi di depan Jokowi karena akan menjadi yang pertama menempati IKN nantinya.	relevan
Hakim MK Gelar Rapat Bahas Tindak Lanjut Putusan	Mahkamah Konstitusi (MK) akan melaksanakan rapat permusyawaratan hakim (RPH) hari ini. Sebelumnya, Pengadilan Tata Usaha Negara (PTUN) mengabulkan sebagian gugatan	non relevan

PTUN soal Anwar Usman	hakim konstitusi Anwar Usman kepada Ketua Mahkamah Konstitusi Suhartoyo. "Menyatakan batal atau tidak sah Keputusan Mahkamah Konstitusi Republik Indonesia Nomor: 17 Tahun 2023, tanggal 9 November 2023 tentang Pengangkatan Dr. Suhartoyo, S.H, M.H. sebagai Ketua Mahkamah Konstitusi Masa Jabatan 2023-2028," petikan bunyi putusan seperti dikutip, Selasa 13 Agustus. Gugatan dari Anwar Usman itu teregistrasi dengan nomor perkara 604/G/2023/PTUN.JKT. Anwar Usman sebagai penggugat dan Suhartoyo sebagai pihak tergugat.	

Manual data labeling is the process of assigning predefined categories to each data point by human annotators. In this approach, the text data is reviewed one by one and labeled based on specific criteria, ensuring accuracy and relevance according to the research objectives. For this study, each piece of news content was manually categorized into one of three classes relevan, berlebihan, nonrelevan on its alignment with the main topic. This careful labeling step is crucial for building a reliable training dataset for the classification model.

TABEL IV
LABEL DISTRIBUTION

Label Category	Number of Samples
Berlebihan	4890
Nonrelevan	4500
Revelan	4926

Berlebihan label category has the highest number of samples, with 4890 samples. The relevan label category follows with 4926 samples, while the nonrelevan label category has the fewest samples, totaling 4500. This table provides a clear overview of how the data is distributed across the three categories, which is important for further analysis and classification processes.

C. *Data Summarization*

Summarization is performed to condense the content of the data that has undergone the preprocessing stage. After preprocessing, the dataset was reduced to 40,577 entries, which were then ready to proceed to the summarization process. This stage utilizes the initialization and configuration of the IndoBERT base-p1 model. The model configuration is managed using AutoConfig, while the tokenizer and model are loaded using AutoTokenizer and AutoModel. These custom tokenizer and model components are then used to create a summarizer object that performs the text summarization. The summarization is designed to generate a summary that is approximately 50% the length of the original text by using the parameter ratio=0.5.

TABEL V
EXAMPLE DATA BEFORE AND AFTER SUMMARIZATION

Before Summarize	After Summarize
Selain dengan pembalap MotoGP, interaksi Nikita Mirzani dengan penonton di Sirkuit Mandalika pada Minggu 29 September 2024 kemarin juga tak kalah menarik untuk dibicarakan. Dari unggahan yang beredar, Nikita Mirzani tampak mononton balapan MotoGP di bagian dalam Sirkuit Mandalika. Dia terpantau membuat konten bersama seorang videografer. Lagi asyik bikin konten, Nikita Mirzani mendadak disoraki oleh para penonton yang berada di tribun belakang. Dalam video tersebut, para penonton menyoraki mantan istri Antonio Dedola tersebut dengan nama anak sulungnya, Laura Meizani alias Lolly."Lolly, Lolly, Lolly," kata para penonton berteriak. Di lain pihak, Nikita Mirzani memberikan reaksi terbuka alih-alih defensif. Dia tampak memberikan tanda hati ke arah para penonton yang meneriakinya. Nikita Mirzani juga melambaikan tangannya, menunjukan isyarat kiss by, serta membungkukan badannya ke tribun penonton. "Potret seru Nikita Mirzani nonton MotoGP Mandalika 2024," bunyi caption yang disertakan. Cuplikan unggahan video detik-detik Nikita Mirzani disoraki nama Lolly oleh penonton MotoGP di Sirkuit Mandalika ini viral di media sosial Instagram dengan atensi sebanyak 145 ribu jumlah tayangan."NM seasyik ini sekarang ya bebs," tulis akun @gossipnesia, dikutip pada Senin 30 September 2024. Perihal itu, sejumlah netizen turut memberikan respons dan komentar yang beragam. "Dia sekarang makin tenang, lebih bahagia. Mungkin udah nggak ada beban, setidaknya anaknya udah sama dia," tulis seorangnetizen.	Selain dengan pembalap MotoGP, interaksi Nikita Mirzani dengan penonton di Sirkuit Mandalika pada Minggu 29 September 2024 kemarin juga tak kalah menarik untuk dibicarakan. "Lolly, Lolly, Lolly," kata para penonton berteriak. Nikita Mirzani juga melambaikan tangannya, menunjukan isyarat kiss by, serta membungkukan badannya ke tribun penonton. " Cuplikan unggahan video detik-detik Nikita Mirzani disoraki nama Lolly oleh penonton MotoGP di Sirkuit Mandalika ini viral di media sosial Instagram dengan atensi sebanyak 145 ribu jumlah tayangan. Perihal itu, sejumlah netizen turut memberikan respons dan komentar yang beragam. " Dia sekarang makin tenang, lebih bahagia. Akhirnya Nikmir bisa tertawa bahagia," ujar netizen yang lainnya.

D. *Similarity Check*

The similarity process is carried out to measure the degree of alignment between the news titles and their corresponding content. The initial step involves word embedding, which aims to convert the words in the titles and content into numerical vector representations, allowing them to be processed by machines and enabling the calculation of cosine similarity. The model indobenchmark/indobert-base-p1 is used in this study to generate word embeddings. Afterward, the similarity between the texts is computed using the cosine similarity method, which measures the cosine of the angle between two vector representations to assess the level of semantic similarity between the title and the summary. The output of this process is a twodimensional matrix that indicates the cosine similarity score of each title-summary pair.

TABEL VI
EXAMPLE DATA OF SIMILARITY CHECK

Title	Content	Cosine Similarity
Abdel Ingat Hanya Temon yang Terima Dirinya Usai Terjerat Narkoba	Pertemanan Abdel dan Temon memang sudah tidak bisa diragukan lagi. Meski pernah bertengkar selama dua tahun sampai tidak bertegur sapa, tapi hubungan keduanya kini sudah baik-baik saja. Abdel mengungkapkan ada satu momen yang hanya Temon menerima dirinya usai terjerat narkoba. Beberapa tahun lalu, Abdel memang mengaku pernah terjerat narkoba. Baginya setiap orang pasti pernah melakukan kesalahan. " Abdel mengatakan sangat bersyukur karena Temon bisa menerima dirinya setelah badai dalam kehidupannya bisa dilalui. " Setelah keluar dari rehab itu kan memang saya nggak boleh keluar dulu dan hanya Temon yang bisa menerima gue di stasiun radionya waktu itu," tutur Abdel lagi.	0.8934405

E. Modeling

Model testing is a crucial stage in this study, conducted by applying various conditions to evaluate the performance of each model in classifying the relevance between news headlines and their content. Two main approaches were used: the LSTM algorithm and the pre-trained IndoBERT model. The selection of these models was based on their effectiveness in understanding textual context and their suitability for Indonesian-language data. LSTM (Long Short-Term Memory) was chosen for its ability to capture long-term dependencies in sequential data, such as news texts, which is essential for assessing the semantic relationship between headlines and content. Compared to more complex architectures like CNN-LSTM or GRU, LSTM provides a good balance between accuracy and computational efficiency. To enhance contextual understanding, the Bi-LSTM variant was also tested, which processes text bidirectionally—both forward and backward. The evaluation results showed that Bi-LSTM in Scenario I delivered the best performance among all LSTM-based models.

Meanwhile, IndoBERT was selected as it is a pre-trained language model specifically trained on an Indonesian corpus, making it superior in understanding local sentence structure and context. Compared to similar models like IndoRoBERTa, IndoBERT is more lightweight yet still highly effective in text classification tasks. In this study, IndoBERT in Scenario V emerged as the best-performing model overall in assessing news relevance. Therefore, the choice and testing of LSTM and IndoBERT were made through careful theoretical and empirical consideration to determine the optimal model.

1) *Testing LSTM and Bi-LSTM Model*: In the performance testing of the LSTM and Bi-LSTM model, several conditions or variations were applied to determine the best performance of the LSTM model. In this study, six deep learning models were developed with different architectures and input treatments to evaluate their performance in classifying news relevance. All models used the Keras Tokenizer (keras.preprocessing.text.Tokenizer) to convert text into sequences of integers, which were then padded to a uniform maximum length (maxlen) of 3026, representing the longest combined sequence of headline and content in the training data. The following are the various architecture and input used for the LSTM model.

TABEL VII
ARCHITECTURE OF LSTM AND BI-LSTM MODEL

Model	Sc.	Architecture Description	Input
LSTM	I	Embedding(58700, 8, maxlen) → Dropout(0.5) → LSTM(128, return_seq=True) → Dropout(0.5) → BatchNorm → GlobalAvgPool1D → Dense(64, ReLU) → Dropout(0.5) → Dense(3, Softmax)	Sum. Text
	II	Embedding(58700, 8, maxlen) → Dropout(0.5) → LSTM(128, seq) → Dropout(0.5) → BatchNorm → GlobalAvgPool1D → Dense(64, ReLU) → Dropout(0.5) → Dense(3, Softmax)	Sum. Text
Bi-LSTM	I	Embedding(58700, 64, maxlen) → SpatialDropout1D(0.3) → BiLSTM(64, seq) → GlobalMaxPool1D → Dense(64, ReLU) → Dropout(0.5) → Dense(3, Softmax)	Sum. Text
	II	Embedding(58700, 64, maxlen) → SpatialDropout1D(0.3) → BiLSTM(64, seq) → GlobalMaxPool1D → Dense(64, ReLU) → Dropout(0.5) → Dense(3, Softmax)	Sum. Text
LSTM non-summary	I	Embedding(58700, 8, maxlen) → Dropout(0.5) → LSTM(128, seq) → Dropout(0.5) → BatchNorm → GlobalAvgPool1D → Dense(64, ReLU) → Dropout(0.5) → Dense(3, Softmax)	Full Text
	II	Embedding(58700, 64, maxlen) → SpatialDropout1D(0.3) → BiLSTM(64, seq) → GlobalMaxPool1D → Dense(64, ReLU) → Dropout(0.5) → Dense(3, Softmax)	Full Text

The table presents three neural network architectures LSTM, Bi-LSTM, and LSTM non-summary, each with two scenarios that differ in embedding dimensions, pooling strategies, and sequence processing layers. While all models aim to classify text inputs using a combination of dropout, pooling, and dense layers, the key distinctions lie in the input type (summary vs. full text) and whether the model uses unidirectional or bidirectional LSTM layers for sequence learning. The following are the various optimizer, learning rate, epochs used for the LSTM model.

TABEL VIII
PARAMETER ARCHITECTURE OF LSTM AND BI-LSTM MODEL

Model	Sc.	Parameter
LSTM	I	Optimizer (Adam), learning rate (0.0002), loss function (categorical cross-entropy), dan epochs (40).
	II	Optimizer (Adam), learning rate (0.0005), loss function (categorical cross-entropy), dan epochs (40).
Bi-LSTM	I	Optimizer (Adam), learning rate (0.0002), loss function (categorical cross-entropy), dan epochs (10).
	II	Optimizer (Adam), learning rate (0.0005), loss function (categorical cross-entropy), dan epochs (10).
LSTM non-summary	I	Optimizer (Adam), learning rate (0.0005), loss function (categorical cross-entropy), dan epochs (10).
	II	Optimizer (Adam), learning rate (0.0002), loss function (categorical cross-entropy), dan epochs (10).

The table outlines the training parameters used for each model scenario, including optimizer (Adam), loss function (categorical cross-entropy), learning rate, and number of epochs. While LSTM scenarios were trained for 40 epochs, all Bi-LSTM and LSTM non-summary scenarios were trained for only 10 epochs, with varying learning rates between 0.0002 and 0.0005. The following are the result of various test condition variations used for the LSTM model.

TABEL IX
RESULT OF LSTM AND BI-LSTM MODEL PERFORMANCE EVALUATION

Model	Sc.	Training		Validation	
		Acc	Loss	Acc	Loss
LSTM	I	0.9893	0.0327	0.7916	1.7654
	II	0.9875	0.0414	0.7769	1.0118
Bi-LSTM	**I**	**0.9207**	**0.2351**	**0.8439**	**0.4621**
	II	0.9513	0.3587	0.8264	0.5415
LSTM non-summary	I	0.9640	0.1079	0.4001	4.6097
	II	0.3394	1.0986	0.3523	1.0979

The results in the table show that the LSTM model in scenario I achieved the highest training accuracy of 0.9893, but its validation accuracy was only 0.7916 with a relatively high validation loss of 1.7654, indicating a potential overfitting issue. In contrast, the Bi-LSTM model in scenario I demonstrated the best validation performance with an accuracy of 0.8439 and the lowest validation loss of 0.4621, reflecting a good balance between training accuracy and generalization. Meanwhile, the non-summary LSTM model, especially in scenario I, showed the poorest validation performance with an accuracy of only 0.4001 and the highest validation loss of 4.6097, indicating very weak generalization capability. Overall, based on the combination of accuracy and loss values, the Bi-LSTM in scenario I can be considered the most optimal performing model.

TABEL X
COMPARISON OF LSTM AND BI-LSTM MODEL TESTING ACCURACY

Model	Skenario	Testing Accuracy
LSTM	I	79.16%
	II	77.69%
Bi-LSTM	**I**	**84.39%**
	II	83.24%
LSTM non-summary	I	40.01%
	II	35.23%

From the two experiment tables above, it can be seen that the Bi-LSTM model achieves the best test accuracy with 84.39% in scenario I, followed closely by scenario II with 83.24%. This confirms that the Bi-LSTM consistently outperforms the other models, including the LSTM and non-summary LSTM variants, which have significantly lower accuracies. Therefore, the Bi-LSTM model in scenario I can be concluded as the best-performing model overall, demonstrating the highest classification capability during the testing phase among all evaluated configurations.

2) *Testing IndoBERT Model*: In this implementation, the model uses the tokenizer indobenchmark/indobert-base-p1 to efficiently preprocess text data into tokenized input suitable for the model. The training process is conducted with a batch size of 15, which balances memory usage and gradient updates for stable learning during fine-tuning on downstream classification tasks. The following are the various architecture and input used for the IndoBERT model.

TABEL XI
ARCHITECTURE OF INDOBERT MODEL

Model	Sc.	Architecture Description	Input
IndoBERT	I	IndoBERT-base-p1 (pretrained transformer) → Classification Head (Dropout 0.3 → Linear → Softmax)	Sum. Text
	II	IndoBERT-base-p2 (pretrained transformer) → Classification Head (Dropout 0.2 → Linear → Softmax)	Sum. Text
	III	IndoBERT-base-p2 (pretrained transformer) → Classification Head (Dropout 0.2 → Linear → Softmax)	Sum. Text
	IV	IndoBERT-base-p2 (pretrained transformer) → Classification Head (Dropout 0.1 → Linear → Softmax)	Sum. Text
	V	IndoBERT-base-p2 (pretrained transformer) → Classification Head (Dropout 0.3 → Linear → Softmax)	Sum. Text
IndoBERT non-summary	I	IndoBERT-base-p1 (pretrained transformer) → Classification Head (Dropout 0.3 → Linear → Softmax)	Full Text
	II	IndoBERT-base-p2 (pretrained transformer) → Classification Head (Dropout 0.3 → Linear → Softmax)	Full Text

The table shows several configurations of the IndoBERT model using either the IndoBERT-base-p1 or p2 pretrained transformers, each followed by a classification head with dropout, linear, and softmax layers. These configurations vary in dropout rates (0.1–0.3) and input type (summary or full text) to evaluate the impact of different settings on classification performance. The following are the various optimizer, learning rate, epochs used for the IndoBERT model.

TABEL XII
PARAMETER ARCHITECTURE OF INDOBERT MODEL

Model	Sc.	Parameter
IndoBERT	I	model IndoBERT-base-p1 dengan optimizer (AdamW), learning rate (2e-5), loss function (CrossEntropyLoss), dropout (0.3), dan epochs (5).
	II	model IndoBERT-base-p2 dengan optimizer (AdamW), learning rate (1e-5), loss function (CrossEntropyLoss), dropout (0.2), dan epochs (10).
	III	model IndoBERT-base-p2 dengan optimizer (AdamW), learning rate (2e-5), loss function (CrossEntropyLoss), dropout (0.2), dan epochs (10).
	IV	model IndoBERT-base-p2 dengan optimizer (AdamW), learning rate (1e-5), loss function (CrossEntropyLoss), dropout (0.1), dan epochs (10).
	V	IndoBERT-base-p2 dengan optimizer (AdamW), learning rate (2e-5), loss function (CrossEntropyLoss), dropout (0.3), dan epochs (3).
IndoBERT non-summary	I	model IndoBERT-base-p1 dengan optimizer (AdamW), learning rate (2e-5), loss function (CrossEntropyLoss), dropout (0.3), dan epochs (5).
	II	model IndoBERT-base-p2 dengan optimizer (AdamW), learning rate (2e-6), loss function (CrossEntropyLoss), dropout (0.3), dan epochs (5).

The table describes the training parameters for each IndoBERT model configuration, which include the use of the AdamW optimizer, CrossEntropyLoss as the loss function, varying learning rates (ranging from 2e-5 to 2e-6), and different dropout rates. All models were trained for a relatively short number of epochs (3 to 5), aiming to fine-tune the pretrained transformer efficiently while preventing overfitting. The following are the result of various test condition variations used for the IndoBERT model.

TABEL XIII
RESULT OF INDOBERT MODEL PERFORMANCE EVALUATION

Model	Sc.	Training		Validation	
		Acc	Loss	Acc	Loss
IndoBERT	I	0.8113	0.4726	0.7862	0.5438
	II	0.9537	0.1476	0.8656	0.6311
	III	0.8303	0.2706	0.8509	0.5368
	IV	0.9669	0.2169	0.8639	0.6153
	V	**0.8984**	**0.2749**	**0.8560**	**0.4140**
IndoBERT non-summary	I	0.3478	7.9465	0.3706	1.6865
	II	0.8731	0.3090	0.8404	0.4853

The IndoBERT model was evaluated under five different scenarios (I–V). Among these, scenario IV achieved the highest training accuracy at 0.9669, while scenario II yielded the highest validation accuracy at 0.8656. The lowest training loss was also observed in scenario II (0.1476), suggesting effective learning during training. Notably, scenario V attained the lowest validation loss of 0.4140, indicating better generalization capability. In comparison, the IndoBERT non-summary model performed significantly worse in scenario I, with a training accuracy of only 0.3478 and a high training loss of 7.9465. Its performance improved considerably in scenario II, achieving a training accuracy of 0.8731 and a validation accuracy of 0.8404. Overall, IndoBERT scenario V is considered the optimal configuration, as it offers a strong balance between validation accuracy (0.8560) and the lowest validation loss, thereby demonstrating superior generalization performance.

TABEL XIV
COMPARISON OF INDOBERT MODEL TESTING ACCURACY

Model	Skenario	Testing Accuracy
IndoBERT	I	79.54%
	II	89.18%
	III	87.47%
	IV	88.16%
	V	**88.86%**
IndoBERT non-summary	I	37.05%
	II	87.08%

The best testing accuracy was found in IndoBERT scenario V, which achieved a testing accuracy of 88.86%, indicating the most optimal model performance in classifying data during the testing phase. Although scenario II had the highest accuracy, as seen in Tabel IX, the loss for scenario II was much higher (0.6311), indicating instability in predictions. On the other hand, scenario V managed to maintain high accuracy while keeping the loss low, demonstrating a more stable model and better generalization to the validation data.

F. Evaluation

In this study, the evaluation methods include the confusion matrix and ROC-Curve. Additionally, evaluation metrics such as precision, recall, and F1-score are used to provide a more comprehensive assessment of model performance. To accommodate the multi-class classification task, both macro average and weighted average scores for each metric are calculated. The evaluation is conducted on the two best-performing models, LSTM and IndoBERT, based on the highest testing accuracy.

1) *Evaluation Bi-LSTM Model*: The evaluation results conducted on this LSTM model use the results from the modeling that has the highest accuracy. As seen in Table below, the Bi-LSTM model with Scenario I has the highest testing accuracy at 84.39%.

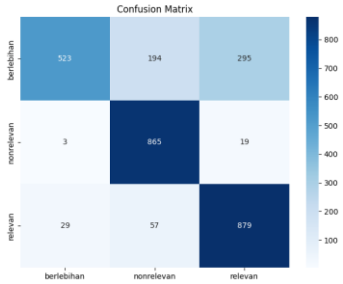


Figure 2. Confusion Matrix for the LSTM Model Scenario I

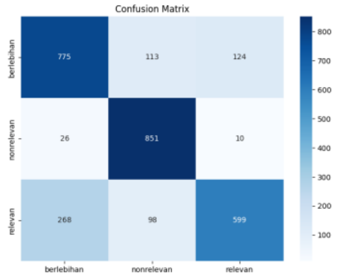


Figure 3. Confusion Matrix for the LSTM Model Scenario II

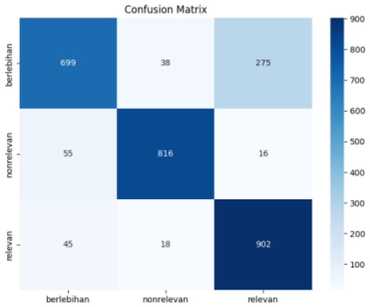


Figure 4. Confusion Matrix for the Bi-LSTM Model Scenario I

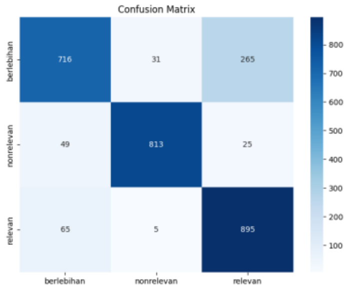


Figure 5. Confusion Matrix for the Bi-LSTM Model Scenario II

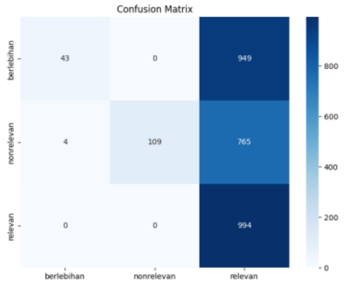


Figure 6. Confusion Matrix for the LSTM-NonSummary Model Scenario I

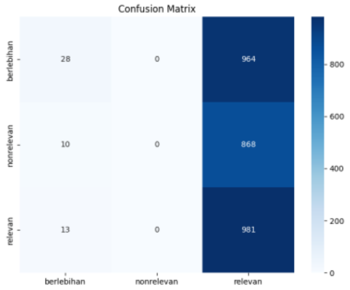


Figure 7. Confusion Matrix for the LSTM-NonSummary Model Scenario II

The evaluation of the LSTM model is based on the best-performing model, the Bi-LSTM from Scenario I, which uses three classes: berlebihan (excessive), nonrelevan (irrelevant), and relevan (relevant), assessed using the confusion matrix. The model successfully classified the data, especially for the nonrelevan class (816 correct predictions) and the relevan class (902 correct predictions). However, it frequently misclassified berlebihan instances as relevan, with 275 misclassifications, suggesting that these two classes may have overlapping features that confuse the model. In total, only 699 berlebihan instances were correctly classified. Overall, the model shows strong performance, but improvements are needed in distinguishing between the berlebihan and relevan categories.

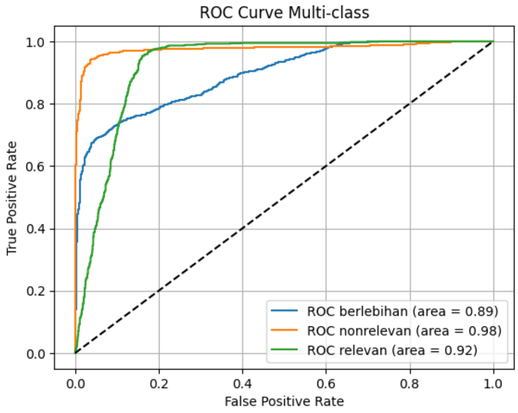


Figure 8. ROC Curve for the Bi-LSTM Model Scenario I

The multi-class ROC curve of the LSTM model demonstrates strong overall performance, with AUC scores of 0.98 for nonrelevan, 0.92 for relevan, and 0.89 for berlebihan. The model performs best at detecting nonrelevan content, indicating a strong ability to distinguish clearly irrelevant headlines and content. However, the model shows relatively lower performance in identifying berlebihan cases, which suggests some difficulty in distinguishing exaggerated news from other categories. This may be due to semantic similarities between berlebihan and relevan, where exaggerated headlines often contain relevant content but are presented with hyperbolic expressions, posing challenges for the LSTM model's semantic sensitivity.

TABEL XV
OVERVIEW OF CLASSIFICATION METRICS BI-LSTM MODEL

Class	Presisi	Recall	F1-Score	Support
Berlebihan	0.87	0.69	0.77	1012
Nonrelevan	0.94	0.92	0.93	887
Relevan	0.76	0.93	0.84	965

The model performs reasonably well with an overall accuracy of 84%. For the berlebihan class, the precision is relatively high at 0.87, but the recall is notably lower at 0.69, indicating that many actual berlebihan instances are missed. This suggests room for improvement in identifying this class. For the nonrelevan class, the model achieves excellent results, with high precision (0.94) and recall (0.92), resulting in a strong F1-score of 0.93. The relevan class shows high recall (0.93) but lower precision (0.76), which implies more false positives. The resulting F1-score of 0.84 reflects solid performance, though increasing the precision for relevan and recall for berlebihan would enhance overall effectiveness.

The macro average scores are 0.86 for precision, 0.85 for recall, and 0.85 for F1-score, indicating that performance is relatively balanced across classes. The weighted average, which takes class support into account, results in slightly lower values (0.85 precision, 0.84 recall, and 0.84 F1-score), suggesting that misclassifications in more frequent classes like berlebihan slightly affect the overall balance. These averages reinforce that the model is consistent but still has room for targeted improvements.

2) *Evaluation IndoBERT Model*: The evaluation results conducted on the IndoBERT model are based on the modeling results with the best accuracy. As seen in Tabel X, the IndoBERT model in Scenario II has the best testing accuracy at 88.86%.

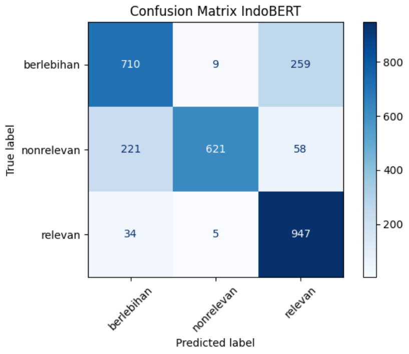


Figure 9. Confusion Matrix for the IndoBERT Model Scenario I

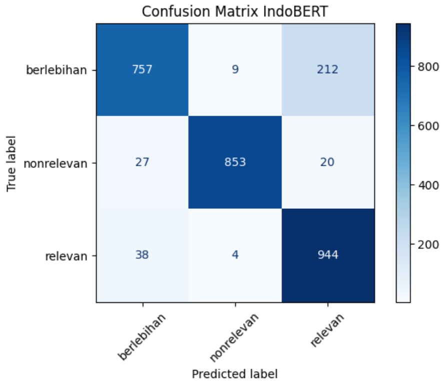


Figure 10. Confusion Matrix for the IndoBERT Model Scenario II

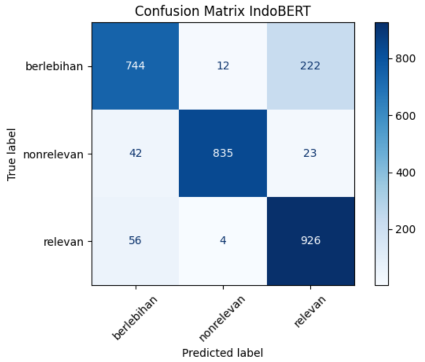


Figure 11. Confusion Matrix for the IndoBERT Model Scenario III

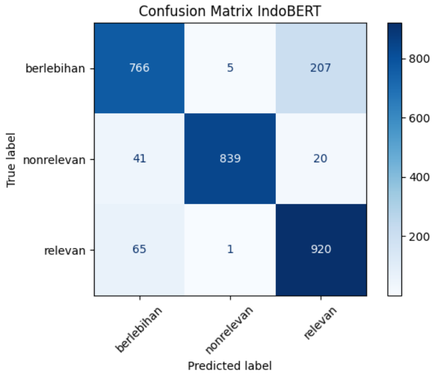


Figure 12. Confusion Matrix for the IndoBERT Model Scenario IV

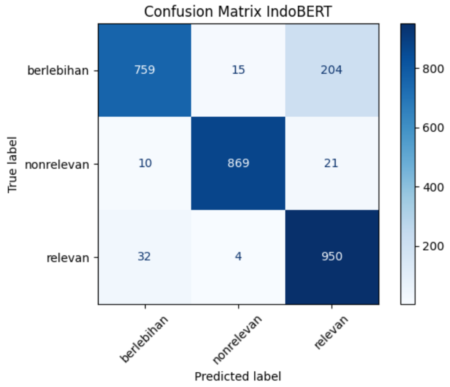


Figure 13. Confusion Matrix for the IndoBERT Model Scenario V

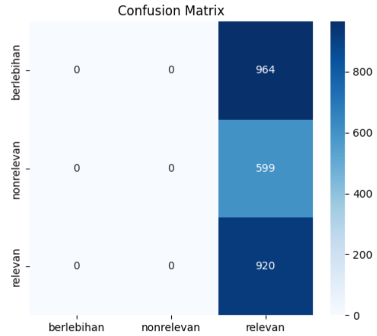


Figure 14. Confusion Matrix for the IndoBERT Model Scenario VI

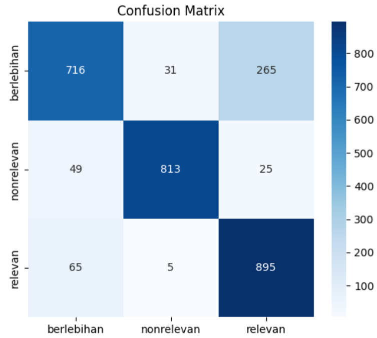


Figure 15. Confusion Matrix for the IndoBERT Model Scenario VII

The confusion matrix for the IndoBERT model in Scenario V, which is the best-performing model overall, demonstrates excellent classification performance across all three classes. The nonrelevan class has the highest prediction accuracy with 869 data points correctly classified, while the relevan class also shows strong performance with 950 correct predictions. For the berlebihan class, 759 instances were correctly classified, though there is still some misclassification, with 204 instances predicted as relevan and 15 as nonrelevan. This indicates that the model occasionally confuses berlebihan content with relevan, although the overall model performance remains robust.

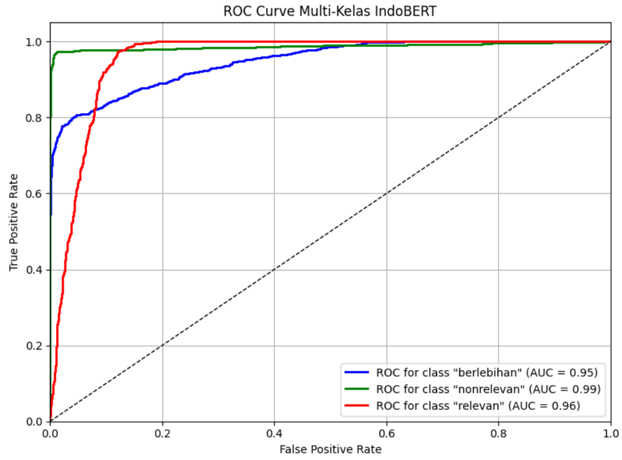


Figure 16. ROC Curve for the IndoBERT Model Scenario V

Based on the multi-class ROC curve of the IndoBERT model, classification performance across the three classes *berlebihan*, *nonrelevan*, and *relevan* is excellent, as indicated by the high AUC (Area Under Curve) values: 0.95 for *berlebihan*, 0.99 for *nonrelevan*, and 0.96 for *relevan*. These values demonstrate the model's strong ability to distinguish between classes, especially in identifying *nonrelevan* content. The *berlebihan* class shows slightly lower performance but still with a very good AUC of 0.95. The curves indicate the model effectively differentiates all classes with high confidence. This suggests the model can reliably classify headlines and content, although subtle challenges may remain in distinguishing *berlebihan* from *relevan* content due to semantic similarities.

TABEL XVI
OVERVIEW OF CLASSIFICATION METRICS INDOBERT MODEL

Class	Precision	Recall	F1-Score	Support
Berlebihan	0.92	0.77	0.84	978
Nonrelevan	0.97	0.94	0.95	900
Relevan	0.80	0.96	0.88	986

The model performs strongly with an overall accuracy of 89%. For class berlebihan, precision is high at 0.92, but recall is lower at 0.77, indicating that while the predictions made for this class are often correct, the model fails to identify a substantial portion of actual class berlebihan instances. This suggests potential for improvement in capturing more true positives. For class nonrelevan, the model performs exceptionally well with a precision of 0.97 and recall of 0.94, leading to a high F1-score of 0.95, indicating strong reliability in both detecting and correctly predicting class nonrelevan. For class relevan, recall is very high at 0.96, but the precision is lower at 0.80, meaning that while most actual class relevan instances are found, a significant number of incorrect predictions are also made as class relevan.

The macro average scores (0.90 precision, 0.89 recall, and 0.89 F1-score) show balanced performance across classes, treating each class equally regardless of size. The weighted average, which considers class support, also yields consistent values (0.90, 0.89, and 0.89 respectively), reflecting stable model behavior even in the presence of class distribution differences. Overall, while the model shows robust classification ability, enhancements in recall for class berlebihan and precision for class relevan could further boost performance.

3) *Cross Validation*: The model evaluation was conducted using IndoBERT and Bi-LSTM, which achieved the highest testing accuracy through 5-fold cross-validation.

TABEL XVII
INDOBERT MODEL EVALUATION RESULTS USING CROSS VALIDATION

IndoBERT Model					
Fold	**Train Acc**	**Val Acc**	**Precision**	**Recall**	**F1 Score**
1	0.88840	0.87600	0.88800	0.87600	0.87480
2	0.88690	0.87220	0.87610	0.87220	0.87170
3	0.88700	0.87220	0.87900	0.87220	0.87110
4	0.89080	0.86900	0.88360	0.86900	0.86610
5	0.88710	0.87640	0.88300	0.87640	0.87560
Mean	**0.88804**	**0.87316**	**0.88194**	**0.87316**	**0.87186**

Model evaluation was performed using 5-fold cross-validation to assess the stability and generalization capability of the IndoBERT and Bi-LSTM models. IndoBERT demonstrated the best performance, achieving an average validation accuracy of 87.32% and an F1-score of 87.19%. Additionally, the small gap between training and validation accuracy (1.48%) indicates that no significant overfitting occurred.

TABEL XVIII
BI-LSTM MODEL EVALUATION RESULTS USING CROSS VALIDATION

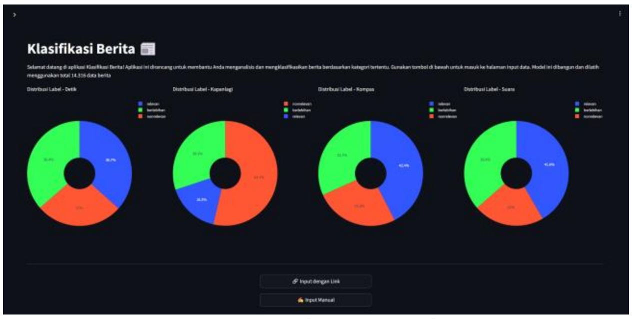
Bi-LSTM Model					
Fold	**Train Acc**	**Val Acc**	**Precision**	**Recall**	**F1 Score**
Fold 1	0.94090	0.8127	0.82300	0.8127	0.81530
Fold 2	0.95820	0.8394	0.84550	0.8394	0.83860
Fold 3	0.92940	0.8511	0.85800	0.8511	0.84820
Fold 4	0.94330	0.8411	0.84610	0.8411	0.83890
Fold 5	0.94110	0.8502	0.85460	0.8502	0.84930
Mean	**0.94258**	**0.8389**	**0.84544**	**0.8389**	**0.83806**

The Bi-LSTM model achieved an average validation accuracy of 83.89% and an F1-score of 83.81%. However, a large gap between training and validation accuracy (10.37%) indicates a potential overfitting issue.

Standard deviation analysis further revealed that IndoBERT exhibited more stable performance than Bi-LSTM across all evaluation metrics. Moreover, paired t-tests yielded p-values less than 0.05 for all metrics, indicating that the performance differences between the models are statistically significant. IndoBERT consistently outperformed Bi-LSTM in terms of validation accuracy, precision, recall, and F1-score, and is therefore recommended as the optimal model for classifying the relevance between news headlines and content.

G. Testing Model

The model testing in this study involves implementing the best-performing model, obtained after testing various scenarios, into a web-based application.

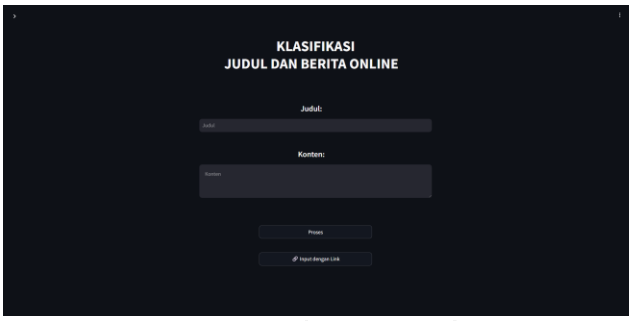


Klasifikasi Berita

Selamat datang di aplikasi Klasifikasi Berita! Aplikasi ini dirancang untuk membantu Anda menganalisis dan mengklasifikasikan berita berdasarkan kategori tertentu. Gunakan tombol di bawah untuk masuk ke halaman input data. Model ini dibangun dan dilatih menggunakan total 14.316 data berita

Figure 17. Landing Page Display

The Landing Page displays the label distribution of the data per portal in the form of a pie chart. Additionally, the landing page features two button options: “Input via link” and “Manual input.”



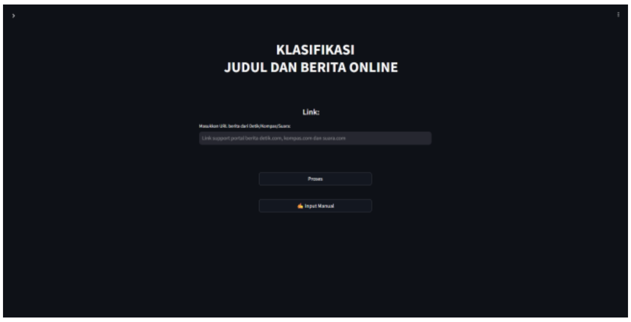
KLASIFIKASI
JUDUL DAN BERITA ONLINE

Judul:

Konten:

Figure 18. Manual Input Page Display

The manual input page displays a web interface for manually entering the news title and content. To perform classification using manual input, the user must copy the data (title and content) from the website they wish to classify, then click the “Process” button to get the result.



KLASIFIKASI
JUDUL DAN BERITA ONLINE

Link:

Masukkan URL berita dari Detik/Kompas/Suara:

Link support portal berita detik.com, kompas.com dan suara.com
Proses
Input Manual

Figure 19. Input Page Display Using a Link

The input via link page displays a web interface containing a single input field for entering the data link to be processed from a portal (detik.com, kompas.com, suara.com) for classification.

IV. CONCLUSION

Based on the research that has been conducted, several conclusions can be drawn regarding the process of evaluating the relevance between headlines and content of online news articles using LSTM and IndoBERT. The following are the conclusions obtained:

1) Based on cross-validation results, the IndoBERT model demonstrated better and more balanced performance compared to the Bi-LSTM model, with an average validation accuracy of 0.87316 and F1-score of 0.87186. In contrast, Bi-LSTM achieved lower validation accuracy and F1-score, at 0.83890 and 0.83806 respectively, and showed signs of overfitting due to a significant gap between training and validation accuracy. Therefore, IndoBERT is selected as the best-performing model overall.

2) The relevance level between headlines and article content on Indonesian news portals, as observed from the average class distribution per portal, showed that kompas.com had an average of 0.478460384, detik.com at 0.457657861, suara.com at 0.45163562, and kapanlagi.com at 0.2329230. Therefore, it can be concluded that kapanlagi.com has the lowest average relevance level among the evaluated portals.

3) In this study, the model’s ability to assess relevance has not been directly compared to human judgment (*human baseline*), which is a limitation worth acknowledging. Ideally, the model's predictions should be compared with annotations made by human evaluators (e.g., expert annotators or typical news readers) to determine how closely the model aligns with human perception of relevance between headlines and article content. Nevertheless, the models used particularly Bi-LSTM in Scenario I and IndoBERT in Scenario V demonstrated strong performance based on evaluation metrics such as accuracy, F1-score, and consistent confusion matrix results. Although an explicit comparison with human performance has not been conducted, the high scores suggest that the models are capable of identifying logical patterns of relevance. Future studies are encouraged to conduct a human agreement study or compare model predictions with human annotations to further validate model performance.

REFERENCES

[1] APJII, “Survei APJII: Pengguna Internet Indonesia Tembus 221 Juta Orang ,” https://www.cnnindonesia.com/teknologi/20240131152906-213-1056781/survei-apjii-pengguna-internet-indonesia-tembus-221-juta-orang.

[2] E. Juliyana and C. A. Nuraflah, “Peranan Internet Dalam Meningkatkan Citra Sma Swasta Budi Agung Medan,” *Jurnal Network Media*, vol. 3, Feb. 2020.

[3] BPTI, “Badan Pengembangan Teknologi dan Informasi,” https://bpti.uhamka.ac.id/sharing/mengenal-python-penjelasan-dan-penggunaannya/.

[4] N. Rahmatika, G. F. Prisanto, S. Tinggi Ilmu Komunikasi InterStudi, J. I. Wijaya No, and J. Selatan, “Pengaruh Berita Clickbait Terhadap Kepercayaan pada Media di Era Attention Economy,” *Avant Garde: Jurnal Ilmu Komunikasi*, vol. 10, no. 02, pp. 190–200, 2022.

[5] B. Hu, Z. Mao, and Y. Zhang, “An overview of fake news detection: From a new perspective,” *Fundamental Research*, vol. 5, no. 1, pp. 332–346, Jan. 2025, doi: 10.1016/j.fmre.2024.01.017.

[6] H. A. Ahmadi and A. Chowanda, “Clickbait Classification Model on Online News with Semantic Similarity Calculation Between News Title and Content,” *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 4, Mar. 2023, doi: 10.47065/bits.v4i4.3030.

[7] N. Newman, R. Fletcher, C. T. Robertson, A. R. Arguedas, and R. K. Nielsen, “Reuters Institute Digital News Report 2024,” 2024. doi: 10.60625/risj-vy6n-4v57.

[8] N. Tendikov *et al.*, “Security Information Event Management data acquisition and analysis methods with machine learning principles,” *Results in Engineering*, vol. 22, p. 102254, Jun. 2024, doi: 10.1016/j.rineng.2024.102254.

[9] M. A. Zamzam, “Sistem Automatic Text Summarization Menggunakan Algoritma Textrank,” *MATICS*, vol. 12, no. 2, pp. 111–116, Sep. 2020, doi: 10.18860/mat.v12i2.8372.

[10] M. R. Hadwirianto, F. Hamami, and O. N. Pratiwi, “Extractive Text Summarization Terhadap Artikel Berita Indonesia Berbasis Machine Learning,” *e-Proceeding of Engineering* , vol. 11, 2024.

[11] A. Arsad, M. Hamid, and M. Santosa, “Penerapan Teks Mining Dan Cosine Similarity Untuk Menentukan Kesamaan Dokumen Skripsi,” *IJIS - Indonesian Journal On Information System*, vol. 9, no. 1, p. 99, Apr. 2024, doi: 10.36549/ijis.v9i1.314.

[12] A. Sagheer and M. Kotb, “Time series forecasting of petroleum production using deep LSTM recurrent networks,” *Neurocomputing*, vol. 323, pp. 203–213, Jan. 2019, doi: 10.1016/j.neucom.2018.09.082.

[13] R. Pramana, M. Jonathan, H. S. Yani, and R. Sutoyo, “A Comparison of BiLSTM, BERT, and Ensemble Method for Emotion Recognition on Indonesian Product Reviews,” *Procedia Comput Sci*, vol. 245, pp. 399–408, 2024, doi: 10.1016/j.procs.2024.10.266.

[14] S. Alaparthi and M. Mishra, “Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey,” Jul. 2020.

[15] K. Kaur and P. Kaur, “BERT-CNN: Improving BERT for Requirements Classification using CNN,” *Procedia Comput Sci*, vol. 218, pp. 2604–2611, 2023, doi: 10.1016/j.procs.2023.01.234.

[16] F. S. Nahm, “Receiver operating characteristic curve: overview and practical use for clinicians,” *Korean J Anesthesiol*, vol. 75, no. 1, pp. 25–36, Feb. 2022, doi: 10.4097/kja.21209.

[17] M. P. Behera, A. Sarangi, D. Mishra, and S. K. Sarangi, “A Hybrid Machine Learning algorithm for Heart and Liver Disease Prediction Using Modified Particle Swarm Optimization with Support Vector Machine,” *Procedia Comput Sci*, vol. 218, pp. 818–827, 2023, doi: 10.1016/j.procs.2023.01.062.