Journal of Deep Learning, Computer Vision and Digital Image Processing
E-ISSN: 2986-8939 P-ISSN: 2986-8920
Vol. 4, No. 2, June 2026
DOI. https://doi.org/10.61255/decoding.v4i2.1528

Integration of named entity recognition and latent Dirichlet allocation for extracting cyberbullying issues on X

Juanda Pratama✉
(Universitas Malikussaleh, Aceh, Indonesia)

Defry Hamdhana
(Universitas Malikussaleh, Aceh, Indonesia)

Zara Yunizar
(Universitas Malikussaleh, Aceh, Indonesia)

OPEN ACCESS

ARTICLE HISTORY
Received: May 7, 2026
Revised: June 21, 2026
Accepted: June 30, 2026

KEYWORDS
Cyberbullying;
Latent dirichlet allocation;
Named entity recognition;
X Platform;
TF-IDF

ABSTRACT

Purpose – The rapid growth of social media Platform X has increased the risk of cyberbullying, which is difficult to detect due to the unstructured nature of textual data. This study proposes an integration of Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) to support the extraction of cyberbullying-related social issues.

Methods – A total of 2,000 tweets were processed through preprocessing, spaCy-based entity extraction, TF-IDF weighting, and LDA topic modeling. The latent topics generated by LDA were manually mapped into four predefined categories (Bodyshaming, Racism, Gender, and Neutral) and evaluated against researcher-annotated ground truth labels.

Findings – Experimental results achieved an overall accuracy of 80%, with F1-scores of 94% for Racism, 93% for Gender, 70% for Bodyshaming, and 63% for Neutral.

Research implications – The proposed framework provides practical support for monitoring cyberbullying patterns and assisting policymakers in understanding online social issues.

Originality – The originality of this research lies in the sequential integration of NER as an entity-filtering stage prior to LDA, enabling a more comprehensive analysis of cyberbullying discussions than the isolated application of either method.

Correspondence Author: ✉defryhamdhana@unimal.ac.id
To cite this article : Pratama, J., Hamdhana, D., & Yunizar, Z. (2026). Integration of named entity recognition and latent Dirichlet allocation for extracting cyberbullying issues on X. Journal of Deep Learning, Computer Vision and Digital Image Processing, 4(2), 277-296. https://doi.org/10.61255/decoding.v4i2.1528

INTRODUCTION

The development of information technology has made social media the primary space for public interaction and communication. Social media is an online platform that provides opportunities for individuals or groups to interact, exchange information, communicate, and build connections in the virtual world using text, photos, videos, or audio. There is a wide and diverse range of social media platforms, with one of the most well-known and widely used by the public being the X application. Platform X is currently a popular social media platform among the Indonesian public and even globally, as it allows for a wealth of news and information to be responded to quickly and accurately from various perspectives. This phenomenon brings various impacts of social media use, both positive and negative, a small example of which is the bullying phenomenon.

Bullying is a variety of aggressive actions carried out repeatedly, characterized by an imbalance of power between the perpetrator and the victim, making it difficult for the victim to protect themselves [1]. Bullying has several types, such as physical and non-physical bullying. Physical bullying is the most visible type of bullying, which includes physically aggressive actions such as hitting, kicking, or ganging up on someone [2]. Meanwhile, non-physical bullying involves indirect aggressive behavior and does not involve physical interaction. This type of bullying consists of verbal insults, mockery, exclusion, and other psychological behaviors, such as emotional manipulation. Non-physical bullying can have a detrimental effect on the development of adolescents' self-concept, including mental health issues such as depression and anxiety [3]. Bullying can occur anywhere, and anyone can be a perpetrator or victim. A frequently occurring case recently is bullying conducted through social media, better known as cyberbullying.

Cyberbullying is an act carried out by an individual or group repeatedly with the aim of frightening, degrading, and threatening others, which can cause psychological, social, and emotional impacts on the victim. In this study, this behavior is manifested through the social media Platform X [4]. There is no minimum or maximum age limit for cyberbullying perpetrators or victims. Children and adolescents, especially those active on the internet across various age ranges, are easily influenced to become perpetrators or even targets of cyberbullying on social media platforms because they spend a lot of time in the virtual world [5]. In Law Number 11 of 2008 concerning Electronic Information and Transactions (UU ITE), which was updated by Law No. 19 of 2016, particularly Article 27 paragraph (3), it emphasizes the prohibition of disseminating electronic information containing insults or defamation. This provision indicates that acts of cyberbullying on social media are not just a social phenomenon but also a legal issue carrying criminal consequences. In February 2020, UNICEF released a report on the issue of bullying in Indonesia. The report revealed that 45% of 2,777 youths in Indonesia aged 14 to 24 had been victims of cyberbullying, based on a survey conducted on U-Report. In terms of reporting, 49% came from boys and 41% from girls (UNICEF, 2020). The resulting impacts can be severe, ranging from decreased self-confidence and excessive anxiety to mental issues such as depression and a tendency to isolate oneself [6].

A study conducted by [7], published in a journal titled “Analisis Teks Pemberitaan Telemedicine di Indonesia: Pendekatan Sentimen, NER, Topic Modeling, dan Social Network dalam Memahami Isu dan Persepsi”, showed that out of 802 online news articles about telemedicine, 49% had a positive sentiment, 40.4% neutral, and 10.6% negative. Using the NER (Named Entity Recognition) method, the research found that the most frequently appearing entities were government officials such as President Joko Widodo and the Minister of Health. Furthermore, research conducted by [8]stated that combining the NER and LDA methods to improve the accuracy of generated addresses showed clearly better results, as the integration of both reached an accuracy rate of 71.97%, exceeding the use of each method separately. The relatively low performance of NER, with an accuracy of only 29.78%, indicates that this method has limitations in providing complete addresses as it only focuses on known entities. On the other hand, the LDA method achieved a better accuracy of 67.80%, thanks to its ability to identify hidden topics, but it remains less accurate than the combination of both methods. This demonstrates that combining NER and LDA produces more accurate and comprehensive addresses compared to using just one of the methods.

Given the rise of cyberbullying on Platform X, this study proposes using an integration of Named Entity Recognition (NER) and Topic Modeling Latent Dirichlet Allocation methods to extract and analyze social issues related to cyberbullying on X. Named Entity Recognition (NER) plays a role in recognizing and categorizing named entities in writing, such as names of people, institutions, locations, and so on. The use of NER in the keyword filtering extraction process provides a deeper understanding of the content and improves the proximity and accuracy of the resulting data [9]. Meanwhile, the use of Topic Modeling (LDA) can be employed to analyze information from social media to understand writing patterns that indicate bullying behavior through digital platforms [10].

Unlike previous studies that only focused on examining formal health issues in media reporting and used four standalone methods without any integration, this study offers a solution approach to understand the increasingly prevalent phenomenon of cyberbullying in the digital world of X. By using the Named Entity Recognition method, this study can recognize related entities, such as user accounts, organizations, and specific groups frequently discussed in that context. Latent Dirichlet Allocation topic modeling is used to identify the main themes of bullying that emerge on social media platforms. The combination of these two methods allows for the automatic recognition of bullying behavior patterns from big data, thus enabling the mapping of cyberbullying types into four categories: bodyshaming, racism, gender, and neutral. The neutral category in this dataset is populated by tweets or comments that contain the targeted keywords but are used in non-aggressive contexts, such as general discussions, news reporting, educational purposes, or anti-bullying campaigns. The inclusion of this neutral class is crucial to ensure that the machine learning system learns to distinguish between the malicious intent of these terms and their normal, benign usage.

Although the sequential integration of NER and LDA has been reported in previous studies, this research introduces several domain-specific adaptations for Indonesian cyberbullying detection on Platform X. These adaptations include slang normalization during preprocessing, cyberbullying-oriented keyword filtering, entity extraction focused on social media discussions, and manual semantic mapping of latent topics into four predefined cyberbullying categories (Bodyshaming, Racism, Gender, and Neutral). These adaptations enable the framework to process informal social media language more effectively than a direct implementation of the existing workflow. Therefore, this research can serve as a valuable reference for the development of more effective cyberbullying prevention policies and intervention strategies by providing comprehensive insights into the patterns, characteristics, and distribution of cyberbullying content on social media. Through the integration of Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA), the proposed approach enables the identification of both relevant entities and underlying discussion topics, allowing stakeholders to better understand the context and dynamics of cyberbullying. Furthermore, the findings of this study can contribute to enhancing public awareness, supporting future academic research, and assisting policymakers, educators, and social media platforms in designing evidence-based initiatives to reduce the prevalence and impact of cyberbullying in digital environments

LITERATURE REVIEW

Bullying

Bullying is a form of violence intentionally committed by an individual or a group that can have serious consequences for the well-being of others [11]. Bullying refers to deliberate and repeated acts of aggression, whether physical, verbal, or psychological, intended to harm, intimidate, or demean another individual [12]. Physical bullying is the most visible type of bullying, involving direct physical aggression such as hitting, kicking, or assaulting another person [2]. In contrast, verbal or psychological bullying is carried out through intimidation using offensive language, including insults, mockery, ridicule, or derogatory expressions intended to humiliate or belittle others [13].

Cyberbullying

Cyberbullying is a form of bullying that occurs through digital media, such as comments, messages, or posts on social media platforms, with the intention of humiliating, intimidating, or threatening other individuals [14]. In this study, cyberbullying is limited to three main categories: bodyshaming, racism, and gender-based cyberbullying.

Bodyshaming refers to insulting or mocking an individual's physical appearance or body condition, which may negatively affect mental health and self-confidence [15]. Racism refers to discriminatory behavior based on race or ethnicity that arises from stereotypes and beliefs regarding the superiority of a particular group [16]. Meanwhile, gender-based cyberbullying refers to acts of intimidation or harassment directed at individuals based on their gender or gender identity, which are influenced by social and cultural factors within digital environments [17].

Platform X

One of the most widely used social media platforms today is Twitter, now known as X. It is a microblogging platform that allows users to publish short posts of up to 280 characters. Twitter users come from diverse backgrounds, including government officials, celebrities, public figures, and the general public [18].

Web Scraping

Web scraping is a technique used to automatically extract information from websites. This process involves the use of software or algorithms capable of collecting required data from web pages, including text, images, or other types of information, which are generally presented in HTML or XML format. As the amount of online information continues to grow, web scraping has become an essential tool for research and data analysis across various disciplines [19]. In this study, web scraping was used to collect tweets and comments from the X platform.

Data Mining

Data mining is a process that integrates statistical methods, mathematics, artificial intelligence, and machine learning (a field that enables computers to operate and learn autonomously without explicit user guidance) to discover meaningful and valuable patterns from large datasets [20].

Python

Python is one of the most popular programming languages and is widely used in various applications today. Known for its simple and readable syntax, Python is an ideal choice for beginners learning programming. With the increasing demand for information technology solutions, Python offers great flexibility and is widely applied in web development, data analysis, artificial intelligence, and task automation [21].

Machine Learning

Machine Learning is a branch of Artificial Intelligence (AI) that enables computers to learn independently from available data and improve their performance over time without being explicitly programmed. Machine learning algorithms identify patterns within data and use these patterns to make predictions autonomously [22]. In this study, the machine learning process occurs during data analysis using the Latent Dirichlet Allocation (LDA) method.

Latent Dirichlet Allocation (LDA) is a prominent topic modeling method in Natural Language Processing (NLP) and text mining. First introduced by Blei, Ng, and Jordan in 2003, LDA is a probabilistic generative model designed to analyze large collections of textual data by identifying latent topics within document collections [23].

Named Entity Recognition

Named Entity Recognition (NER) is used to identify and classify named entities within text, such as person names, organizations, locations, and other entity types. Integrating NER into the keyword extraction process provides a deeper understanding of textual content while improving the relevance and accuracy of the extracted information [9].

Natural Language Processing

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that focuses on enabling computers to understand and interact with humans through natural language. In recent years, NLP has advanced rapidly due to the increasing availability of textual data and the growing demand for

more sophisticated human-computer communication that reflects the way humans naturally communicate [24].

Term Frequency-Inverse Document Frequency

Term Frequency–Inverse Document Frequency (TF-IDF) is a feature extraction technique that transforms text into high-dimensional numerical vectors based on the importance of each term across the entire corpus. This method is highly efficient for large-scale datasets because it emphasizes document-specific keywords while reducing the influence of commonly occurring words [25].

METHOD

Place and Time of Research Implementation

This study employed a quantitative descriptive research design. The research was conducted in Lhokseumawe, Aceh, Indonesia, using textual data collected from Platform X. The study was carried out from October 2025 until the completion of the research.

Research Scheme

The research framework provides a comprehensive overview of the processes and steps implemented in this study. The following is the research framework for this study:

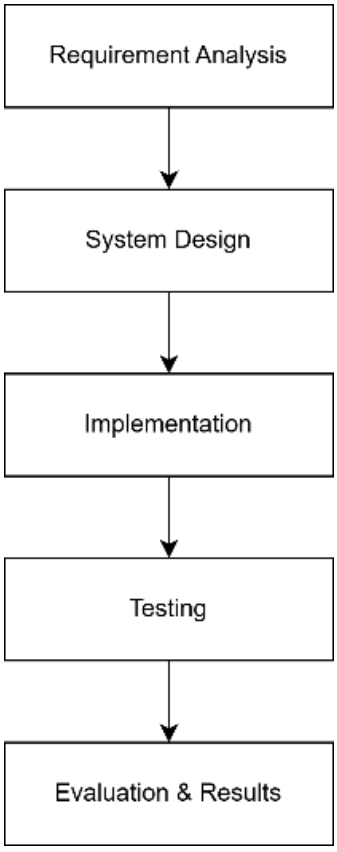

Figure 1. Research Scheme

a. Requirement Analysis: This stage is the initial phase of system development, where the necessary data for the research are collected. In this study, the process includes collecting tweets or comments from the X platform as the primary dataset.

b. System Design: At this stage, the system architecture is designed to identify and categorize cyberbullying conversations on the X platform by integrating the Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) topic modeling methods.

c. Implementation: The designed system is then implemented into a functional application. During this stage, data are collected from the X platform using an API, followed by a preprocessing process that includes removing punctuation, URLs, emojis, and other irrelevant words to improve data quality before analysis.

d. Testing: The testing phase is conducted to ensure that the integration of the Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) methods functions correctly and produces outputs that meet the objectives of the research.

e. Evaluation and Results: The final stage aims to evaluate the overall performance of the system in terms of both the accuracy of the analysis results and the clarity of data visualization. The outcome of this research is the classification of cyberbullying into four categories: Bodyshaming, Racism, Gender-based Cyberbullying, and Neutral.

System Requirements Analysis

A system requirements analysis was conducted to identify the tools used in the development of the integrated Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) system for extracting cyberbullying social issues on the X platform. The hardware used includes a laptop with an AMD Ryzen 5 processor, 8 GB of RAM, and a 512 GB SSD. Meanwhile, the software used consists of Windows 11 as the operating system, Visual Studio Code as the development environment, Google Colab for modeling and model training, XAMPP as a local server, and Microsoft Office 2021 for research documentation.

Training Data System Scheme

The Training Data System Schema is a workflow illustration that depicts how raw data is collected, processed, and utilized to train a machine learning model. The following are the steps of the training data system schema:

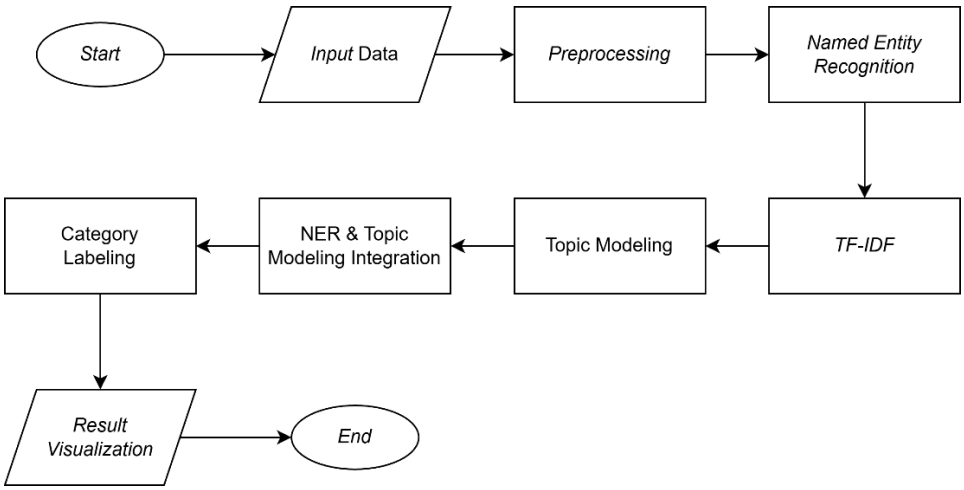

Figure 2. Training Data System Scheme

a. Start

The initial step indicating that the system will be executed.

b. Input Data

At this stage, the collected dataset of 2,000 tweets is divided into 80% development data (1,600 tweets) and 20% evaluation data (400 tweets). Although LDA is an unsupervised probabilistic topic modeling method rather than a supervised classifier, the 80% portion is used to build the corpus dictionary and fit the LDA topic model by estimating the latent topic distributions. The remaining

20% is reserved for topic inference on previously unseen documents. The inferred dominant topics are then manually mapped to the predefined cyberbullying categories and compared with the manually annotated ground-truth labels for evaluation.

c. Data Preprocessing

This stage consists of several steps, namely:

a. Case Folding: This step is useful for converting text into lowercase to facilitate the analysis.
b. Tokenizing: At this step, the text will be separated into individual tokens.
c. Stopword Removal: This step filters out words that have no significant meaning, such as conjunctions or prepositions (e.g., "dan", "di") and others.
d. Stemming: This step converts words into their base forms without affixes.
e. Filtering Data: This step selects important information and removes useless data so that the analysis results become more accurate.

d. Named Entity Recognition (NER)

This stage is useful for assigning labels to important entities in the text. Examples of entities include people (victims or perpetrators), mentioned organizations, and others.

e. TF-IDF

This stage converts text into a numerical representation (weight vectors). This process is useful for assessing the level of importance of a word, so that keywords characteristic of cyberbullying can be identified mathematically by the system.

f. Topic Modeling

After labels are assigned to important entities in the text, this section applies Latent Dirichlet Allocation (LDA) Topic Modeling to discover the topics and themes discussed in the cyberbullying acts.

g. Integration

This stage combines Named Entity Recognition and Latent Dirichlet Allocation Topic Modeling, which is useful for better understanding the social issue of bullying, as well as identifying the perpetrators, victims, and frequently emerging topics.

h. Category Labeling

It is important to distinguish between the dataset preparation phase and the final live system implementation. While the final web application automatically categorizes cyberbullying issues without human intervention, the initial preparation of the entire dataset involved a rigorous manual annotation process. All 2,000 text data points were manually evaluated and categorized by the researcher into one of the four predefined classes: bodyshaming, racism, gender, or neutral, to establish the ground truth (actual) labels. This comprehensive manual labeling was essential to provide a definitive "answer key" for evaluating the model's accuracy on the testing set (as presented in the confusion matrix) and to reliably map the unsupervised topic clusters generated by the LDA model to the actual cyberbullying categories.

To ensure annotation consistency, the manual labeling process followed predefined annotation guidelines established before the annotation stage. Each tweet was assigned to only one category based on its dominant cyberbullying context. Bodyshaming labels were assigned to texts targeting physical appearance, Racism labels to texts containing racial, ethnic, or religion-based attacks,

Gender labels to texts containing insults related to gender or gender identity, and Neutral labels to texts discussing the monitored keywords without cyberbullying intent. The same annotation criteria were consistently applied throughout the entire dataset.

Table 1. Category Labeling

Category	Annotation Criteria
Bodyshaming	Tweets containing insults or ridicule directed at physical appearance or body characteristics.
Racism	Tweets containing insults or discriminatory expressions targeting race, ethnicity, nationality, or religion.
Gender	Tweets containing insults directed at gender, gender identity, or sexual orientation.
Neutral	Tweets containing general words that do not contain elements of bullying

i. Result Visualization

This stage will display the analysis results in a visual form to make them easier to understand. In this system, the output displays the entities discussed in the tweets or comments, shows the topics present in those tweets or comments, and displays the discussed topics categorized into bodyshaming, gender, or neutral.

j. End

Indicates the end of the system process.

User System Scheme

The user system schema is a system workflow design that demonstrates how users interact with the system. The following is the user system schema:

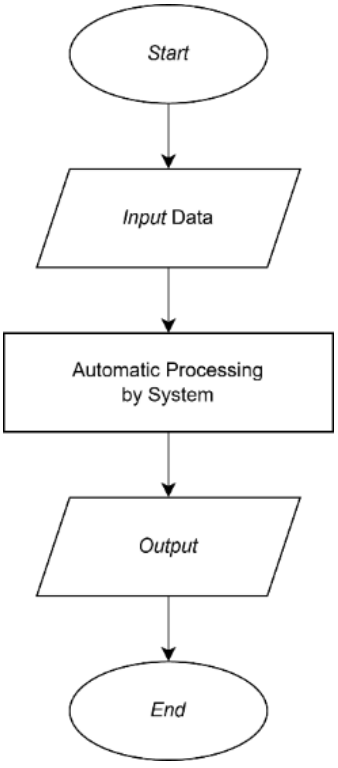

Figure 3. User System Scheme

a. Start

The initial step indicating that the system testing process will begin.

b. User

This stage is the first point of interaction for the user with the system.

c. Input Data

At this stage, the user must input the sentence data to be processed.

d. Automatic Processing by System

This process is the stage where the system independently processes information from tweets or comments without human involvement. This includes reading the content, cleaning the data, identifying main entities, discovering the discussed themes, and generating an output in the form of bullying discussion patterns. All steps, such as preprocessing, entity extraction (NER), topic modeling (LDA), and the integration of both processes, are performed automatically by the algorithm.

e. Output

This is the final stage in the system that provides the results of the previously analyzed text. In this research, the output displayed will be the topics discussed in the tweet or comment, categorized into cyberbullying types such as bodyshaming, racism, gender, or neutral, and will also display the entities discussed within the tweet or comment.

f. End

Indicates the end of the system process.

RESULTS AND DISCUSSION

Dataset Collection

The dataset collection phase in this study was conducted using text crawling techniques directly from the social media Platform X. The data extraction focused on tweets or replies from users containing keywords relevant to indications of cyberbullying issues, such as racist sentiments, body shaming, and gender.

Table 2. Raw Data

No.	Raw Data
1.	@geloraco SURYO BANCI
2.	@wayangndeso77 @yappingfess Pasti lu orang Jawa tolol percaya cerita hoax bgni. Kresten emang goblok
3.	@tanyakanrl Bete jink ada cewe matre
4.	tina kebanyakan makan bakso...,,,,,,---

Data Cleaning

Table 3. Data Cleaning

No.	Data Cleaning
1.	suryo banci
2.	kamu orang jawa bodoh percaya cerita hoax kristen emang bodoh
3.	kesal anjing cewe matre
4.	tina banyak makan bakso

Table 3 presents the final results of the data cleaning or text preprocessing stage on the raw text data. At this stage, the text has been transformed through a case folding process, converting all characters to lowercase. Furthermore, elements that act as noise in the text data, such as punctuation, symbols, excessive spacing, and other social media attributes, have been completely removed. Slang word handling and stemming processes have also been applied to transform non-standard words or abbreviations into standard Indonesian forms. The resulting clean text from this stage minimizes word ambiguity and serves as an optimal data representation to be executed in the subsequent stages, namely Named Entity Recognition (NER) extraction and word feature weighting.

Entity Extraction with Named Entity Recognition

Table 4. Entity

Text	Entity Detected	Entity Label
suryo banci	Suryo	Orang
kamu orang jawa bodoh percaya cerita hoax kristen emang bodoh	Jawa	Lokasi
kesal anjing cewe matre	-	-
tina banyak makan bakso	Tina	Orang

Table 3 presents the entity extraction results using the Named Entity Recognition (NER) method on the cleaned text data. At this stage, the model spaCy automatically scans the text to identify and classify specific words into designated entity labels, such as names of individuals (Person) and places (Location). For instance, the system successfully extracted the word "Suryo" and labeled it as a 'Person' entity in the first row, and detected "Jawa" as a 'Location' entity in the second row. This entity extraction process is essential for identifying the subjects or targets being discussed within the cyberbullying context, and this information will subsequently be integrated with topic modeling in the next stage to produce a more comprehensive analysis.

Unlike previous studies that applied NER in formal text domains, this research applies standard entity extraction to informal social media text on Platform X. Despite the presence of non-standard writing styles and slang expressions, the extracted Person, Organization, and Location entities provide contextual information that complements the topic modeling results in identifying cyberbullying-related discussions.

Word Weighting with Term Frequency-Inverse Document Frequency

Table 5. TF-IDF

Term	TF	IDF	TF-IDF
anjing	0.250	0.6021	0.1505
bakso	0.250	0.6021	0.1505
banci	0.500	0.6021	0.3010
banyak	0.250	0.6021	0.1505
bodoh	0.250	0.6021	0.1505
cewe	0.250	0.6021	0.1505
cerita	0.125	0.6021	0.0753
hoax	0.125	0.6021	0.0753
jawa	0.125	0.6021	0.0753
kesal	0.250	0.6021	0.1505
kresten	0.125	0.6021	0.0753
kulup	0.125	0.6021	0.0753
makan	0.250	0.6021	0.1505
matre	0.250	0.6021	0.1505
percaya	0.125	0.6021	0.0753

Term	TF	IDF	TF-IDF
suryo	0.500	0.6021	0.3010
tina	0.250	0.6021	0.1505

Table 4 displays the final calculation results of the word feature matrix weighting using the Term Frequency-Inverse Document Frequency (TF-IDF) method. The final TF-IDF value in the table is obtained by multiplying the term's occurrence frequency within a single document (TF) by its uniqueness value across the entire document corpus (IDF). Based on the calculations, terms such as "banci" and "suryo" have the highest weight scores of 0.3010. This indicates that these words are highly dominant in specific documents but relatively rare in others, leading the algorithm to evaluate them as highly strong and representative keywords or features. Conversely, words with lower TF-IDF weights, such as "cerita", "hoax", and "jawa" (0.0753), hold less significance in distinguishing context between documents. This numerical weighting result subsequently serves as the input data for the Machine Learning model in the classification stage.

Latent Dirichlet Allocation Modeling

Table 6. Words

Document	Text
D1	suryo banci
D2	orang jawa bodoh percaya cerita hoax kresten kulup bodoh
D3	kesal anjing cewe matre
D4	tina banyak makan bakso

Next, frequency calculations are carried out for the 4 previously determined datasets.

Table 7. Frequency

Term	Frequency
bodoh	2
suryo	1
banci	1
jawa	1
percaya	1
cerita	1
hoax	1
kresten	1
kulup	1
kesal	1
anjing	1
cewe	1
matre	1
tina	1
banyak	1
makan	1
bakso	1

Total words = 18

a. Calculation of Dominant Topics

1. Topic 1 (cyberbullying)

Table 8. Frequency topic 1

Words	Total
bodoh	2
banci	1
anjing	1
matre	1
kesal	1
jawa	1

Total topic frequency = 7

Probability:

$$P(\text{bodoh} \mid T_1) = \frac{2}{7} = 0.286$$
$$P(\text{banci} \mid T_1) = \frac{1}{7} = 0.143,$$

For the word "banci" the probability obtained is 14.3%, the higher the probability of the word, the more the word represents the topic sentence.

2. Topic 2 (General/Neutral Words)

Table 9. Common word frequency

Text	Frequency
tina	1
banyak	1
makan	1
bakso	1

Total topic frequency = 4

Probability:

$$P(\text{bakso} \mid T_2) = \frac{1}{4} = 0.25$$

To convert the LDA's continuous probabilistic output into a single, definitive discrete label (predicted label) for the confusion matrix, the system applies an argmax function. The system assigns a single predicted label to each document based on its dominant topic—the topic that yields the highest probability distribution score in that specific text. Furthermore, the mapping of the unsupervised hidden topics generated by LDA to the four predefined researcher categories (Bodyshaming, Racism, Gender, Neutral) was conducted manually. The researchers inspected the top 15 most frequent and dominant keywords (terms) within each generated LDA topic cluster and qualitatively aligned them with the semantic context of the targeted cyberbullying categories. For instance, a cluster dominated by terms like "cina", "jawa", was manually mapped to the "Racism" class.

Table 10. Manual Topic-to-Category Mapping Based on Dominant LDA Keywords

LDA Topic	Top Keyword	Category	Mapping Rationale
Topic 1	jelek, gendut, muka hitam	Bodyshaming	Dominated by words that insult physical appearance.
Topic 2	cina, jawa, papua, batak, kafir	Racism	Most keywords refer to race, ethnicity, or religion.

LDA Topic	Top Keyword	Category	Mapping Rationale
Topic 3	banci, bencong, cewe, cowo, gay	Gender	Dominant words are related to gender identity and gender-based insults.
Topic 4	makan, bakso, rumah, sekolah, pagi	Neutral	common words that do not lead to cyberbullying

Model Evaluation

The evaluation of the system's performance was conducted by comparing the category labels obtained from the dominant inferred topic after manual topic-to-label mapping with the manually annotated ground-truth labels. Although the underlying model is based on unsupervised topic modeling, the dominant inferred topic of each document was manually mapped to one of the predefined cyberbullying categories, enabling the calculation of standard evaluation metrics such as accuracy, precision, recall, and F1-score. The results of this comparison are presented in the confusion matrix below, which subsequently serves as the basis for calculating precision, recall, F1-score, and overall accuracy across the four predefined categories.

Table 11. Model Evaluation

Actual	Prediction			
	Bodyshaming	Gender	Neutral	Racist
Bodyshaming	80	2	25	0
Gender	7	99	2	1
Neutral	31	1	57	5
Racist	2	1	2	85

$$Accuracy = \frac{Total\ Correct\ Predictions}{Total\ of\ All\ Data} (1)$$

$$TP = 321$$

Total Data = 400

$$= \frac{321}{400} = 0{,}8025\ (80{,}25\%)$$

a. Bodyshaming Class Calculation

Precision measures the machine's accuracy when predicting the Body shaming class, while Recall measures the machine's sensitivity in identifying all existing Body shaming data.

TP=80

FP=7+31+2 = FP=40

$$FN = 27$$

$$\text{Precision} = \frac{80}{80 + 40}$$

$$= \frac{80}{120}$$

$$= 0{,}6667$$

$$= 66{,}67\%$$

$$Recall = TP + FNTP \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$= \frac{80}{80 + 27}$$

$$= \frac{80}{107}$$
$$= 0{,}7477$$
$$= 74{,}77\%$$
$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \text{.....(3)}$$
$$= 2 \times \frac{0{,}6667 \times 0{,}7477}{0{,}6667 + 0{,}7477}$$
$$= 0{,}7048$$
$$= 70{,}48\%$$

b. Gender Class Calculation

$$\text{TP}=99$$
$$FP = 2 + 1 + 1 = 4$$
$$\text{FN}=7+2+1=10$$
$$Precision = \frac{99}{99 + 4}$$
$$= \frac{99}{103}$$
$$= 0{,}9612$$
$$= 96{,}12\%$$
$$Recall = \frac{99}{99 + 10}$$
$$\frac{99}{109} = 0{,}9083$$
$$= 90{,}83\%$$
$$= 2 \times \frac{0{,}9612 \times 0{,}9083}{0{,}9612 + 0{,}9083}$$
$$=0.9340 = 93{,}40\%$$

c. Neutral Class Calculation

$$\text{TP}=57$$
$$\text{FP}=25+2+2=29$$
$$\text{FN}=31+1+5=37$$
$$Precision = \frac{57}{57 + 29} = 0{,}6628 = 66{,}28\%$$
$$Recall = \frac{57}{57 + 37} = 0{,}6064\ (60{,}64\%)$$
$$F1\text{-}Score = 2 \times \frac{66{,}28 \times 60{,}64}{66{,}28 + 60{,}64} = 0{,}6334 = 63{,}34\%$$

d. Racism Class Calculation

$$\text{TP}=85$$
$$\text{FP}=0+1+5=6$$
$$\text{FN}=2+1+2=5$$
$$Precision = \frac{85}{85 + 6} = 0{,}9341\ (93{,}41\%)$$
$$Recall = \frac{85}{85 + 5} = 0{,}9444\ (94{,}44\%)$$

$$F1\text{-}Score = 2 \times \frac{0{,}9341 \times 0{,}9444}{0{,}9341 + 0{,}9444} = 0.9392\ (93{,}92\%)$$

System Implementation

The following is the result of the system implementation in the form of a web-based application developed using the Python programming language and the Flask framework. This application is designed to detect and extract cyberbullying social issues from text data on Platform X by integrating Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) methods. Through this application, users can comprehensively observe the stages of the analytical process, ranging from text cleaning (text preprocessing) and target entity extraction, to the final results of interactive cyberbullying categorization.

a. Dashboard Page
The dashboard page is the initial page for users who have successfully logged into the system.

Figure 4. Dashboard

b. Detection Page
This is the results page, where beforehand the user must input the dataset they wish to detect into the system. After the user uploads the dataset in Excel format, the system will process it and display the results as shown in the following image.

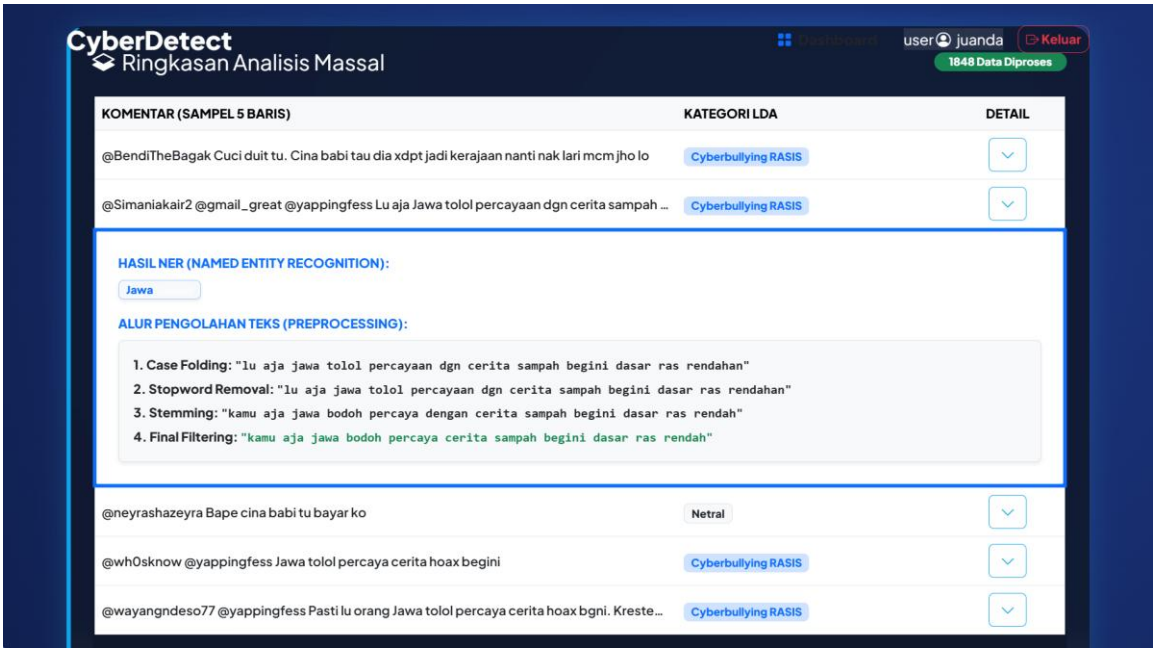

Figure 5. Detection

Figure 6. Word Cloud

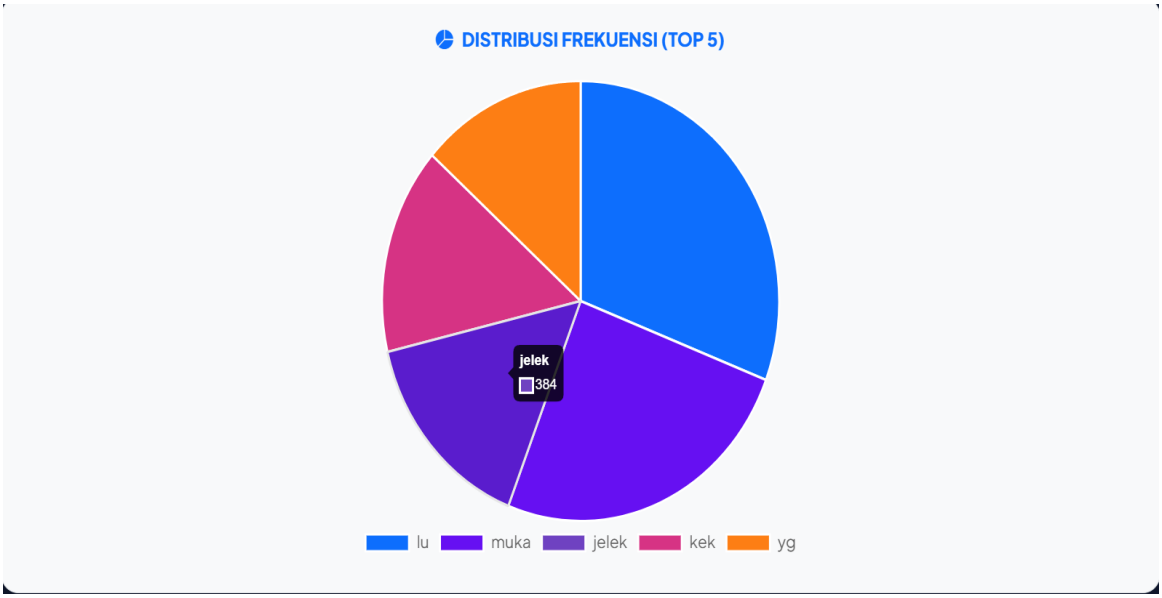

Figure 7. Frequency Distribution

The integration of Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) provides a robust framework for identifying cyberbullying social issues within unstructured text data from the X platform, yielding an overall classification accuracy of 80%. This finding confirms that the synergy between entity extraction and thematic modeling offers superior predictive capabilities compared to utilizing individual methods in isolation. The performance of the proposed model aligns with and extends the findings of Taher et al., who observed that the integration of NER and LDA significantly enhances accuracy—specifically reaching 71.97% in their research on address extraction. The current study achieves a higher accuracy of 80% within the context of cyberbullying detection, suggesting that this integration remains highly effective even when applied to the complex, informal, and slang-heavy language characteristic of the X platform. This demonstrates the scalability and adaptability of the integrated NER-LDA pipeline across different domains, from data extraction to social issue classification.

Furthermore, this study addresses the methodological limitations observed in prior research, such as the work by [7] While that study successfully employed NER and topic modeling to analyze telemedicine news, it treated these methods as separate analytical components rather than as a fully integrated, sequential pipeline. In contrast, the current research leverages NER as an essential preprocessing filter to identify key actors, organizations, and entities, which then informs the LDA process. This structural integration allows for a more granular mapping of cyberbullying types, enabling the system to categorize content into distinct classes like racism, gender, and bodyshaming with greater precision.

The high F1-Scores achieved in the Racism (94%) and Gender (93%) categories indicate that the model is particularly adept at detecting specific, keyword-driven social issues. However, the relatively lower performance in the Bodyshaming (70%) and Neutral (63%) categories highlights the inherent challenges of linguistic ambiguity in social media interactions. Furthermore, this performance gap may not be exclusively due to linguistic ambiguity; it is highly likely that the keyword-based data collection and annotation strategies inherently favor categories characterized by strong, explicit lexical markers such as racial or gender slurs. Consequently, these categories naturally achieve higher classification scores regardless of the deeper semantic pattern recognition capabilities of the NER-LDA pipeline. Bodyshaming often relies on figurative or context-dependent language that may not always contain explicit entities or distinct keyword markers, leading to lower sensitivity. These results suggest that while the current model is highly effective for detecting overt forms of aggression, future work could incorporate more context-aware word embeddings to better handle the subtle or metaphorical expressions often found in bodyshaming or neutral-toned conversations.

In conclusion, this study validates the use of an integrated NER-LDA approach as a practical, automated tool for monitoring cyberbullying. It provides a structured foundation that outperforms disjointed method applications, offering a pathway for more effective policy formulation and intervention strategies on social media platforms. While the findings demonstrate the efficacy of the proposed model, certain limitations must be recognized. Primarily, the dataset was exclusively sourced from Platform X and is confined to Indonesian text. This single-platform and single-language focus may constrain the applicability of the model to other social networks or languages. Additionally, The ground-truth annotation for the entire dataset was performed by the researcher following predefined annotation guidelines to ensure consistent labeling across all samples. Nevertheless, because this study did not involve multiple independent annotators or calculate inter-annotator agreement metrics such as Cohen's Kappa, some degree of subjective bias may remain, particularly in context-dependent cases such as sarcasm, figurative language, or implicit

cyberbullying. Future studies should involve multiple annotators and report formal inter-annotator agreement statistics to further strengthen annotation reliability.

CONCLUSION

The integration of Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) methods has proven effective in producing a comprehensive analysis of the cyberbullying phenomenon on the X platform. This approach combines NER's capability to extract key entities—such as the names of individuals, organizations, and locations with LDA's reliability in modeling the main topic patterns of conversations. The evaluation results show that this model achieves an overall accuracy of 80%, with excellent classification performance in the Racism (F1-Score 94%) and Gender (93%) categories. Although the Body shaming (70%) and Neutral (63%) categories were slightly hindered by the contextual similarities of social media language, the system as a whole successfully clustered topics and entities in a structured manner, thus holding great potential to be utilized as a foundation for analysis and decision-making. For future research, it is recommended to expand the dataset across multiple platforms and languages while establishing formal inter-annotator reliability metrics during the labeling process. Additionally, comparing this integrated NER-LDA pipeline against supervised deep learning baselines could provide deeper insights into its comparative effectiveness.

ACKNOWLEDGMENT

The author expresses their deepest gratitude to Universitas Malikussaleh, particularly the Informatics Engineering Study Program, Faculty of Engineering, for the academic support and facilities provided throughout the course of this research. Sincere appreciation is also extended to Dr. Eng., Defri Hamdhana, S.T., M.Kom., as the Main Supervisor, and Zara Yunizar, S.Kom., M.Kom., as the Co-Supervisor, for their time, direction, and invaluable guidance that led to the successful completion of this study, as well as to all parties who have assisted in the writing of this article.

AUTHOR CONTRIBUTION STATEMENT

JP conceived the study, collected the dataset from Platform X, performed data preprocessing, implemented the Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) methods, developed the web-based application, conducted the experiments, analyzed the results, and prepared the original manuscript; DH supervised the research, contributed to the study design and methodology, reviewed the experimental procedures, validated the results, and critically revised the manuscript for important intellectual content; and ZY contributed to the research methodology, assisted in data analysis and interpretation, reviewed the manuscript, provided technical feedback, and approved the final version of the manuscript.

AI DISCLOSURE STATEMENT

The authors used AI-based language assistance tools (Gemini/ChatGPT) solely to improve English grammar, language clarity, and sentence structure during manuscript preparation. The AI tools were not used for the study design, data collection, data analysis, interpretation of results, or scientific decision-making. All research methods, experimental procedures, analyses, interpretations, and conclusions were developed and verified by the authors, who take full responsibility for the accuracy and integrity of the manuscript.

REFERENCES

[1] F. Erkurnia, T. N. Putri, Y. N. Dianningsih, and I. Rachmawati, “Analisis Profil Perilaku Bullying Pada Siswa Tingkat Sekolah Dasar Negeri Di Kabupaten Bantul,” *G-Couns: Jurnal Bimbingan dan Konseling*, vol. 8, no. 3, pp. 1254–1259, 2024, https://doi.org/ 10.31316/gcouns.v8i3.5059.

[2] M. H. T. Pelangi Dea Sri Damayanti, Fitri Handayani, Yuli Ramahwati, Suhernah, Anisa Dian Cahyani, “Peranan Psikologi Pendidikan untuk Pencegahan Perundungan Siswa Sekolah Dasar The Role of Educational Psychology in Preventing Bullying of Elementary School Students,” *Jurnal Bimbingan Konseling Pendidikan Islam*, vol. 4, no. 1, pp. 1–10, 2023, https://doi.org/10.31943/counselia.v4i1.60.

[3] J. M. Tamaela, N. Amir, M. J. Sanggel, and R. S. Nompo, "The Impact of Non-Physical Bullying on Adolescents' Self-Concept: An Observational Study in the Adolescent Community in Sausapor, Tambrauw Regency, West Papua, Indonesia," *Scientia Psychiatrica*, vol. 5, no. 3, pp. 509–519, 2024. https://doi.org/10.37275/scipsy.v5i3.171.

[4] R. Pribadi, "Pengaruh Cyberbullying Terhadap Harga Diri Pada Dewasa Awal Korban Cyberbullying Di Twitter," *Jurnal Riset Psikologi*, vol. 6, no. 3, p. 113, 2023, https://doi.org/10.24036/jrp.v6i3.15036.

[5] I. S. Arfan, S. Fauziah, and I. Nawangsih, "Analisis Sentimen Terhadap Cyber Bullying di X Menggunakan Algoritma Naïve Bayes," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1411–1419, 2024, https://doi.org/10.57152/malcom.v4i4.1550.

[6] Puja Setia Kirana, Marsha Jihan Alisah, Marta Erika Putri, Risma Anita Puriani, and Rizki Novirson, "Cyberbullying : Pola Perilaku, Faktor Pemicu, dan Dampak Psikologis," *Jurnal Nakula : Pusat Ilmu Pendidikan, Bahasa dan Ilmu Sosial*, vol. 3, no. 3, pp. 192–199, 2025, https://doi.org/10.61132/nakula.v3i3.1813.

[7] S. B. Panuntun, D. Krismawati, S. Pramana, and E. T. Astuti, "Analisis Teks Pemberitaan Telemedicine di Indonesia: Pendekatan Sentimen, NER, Topic Modeling, dan Social Network dalam Memahami Isu dan Persepsi," *Indonesian of Health Information Management Journal (INOHIM)*, vol. 11, no. 1, pp. 56–67, 2023, https://doi.org/10.47007/inohim.v11i1.500.

[8] H. A. Taher, N. N. Alabid, and B. M. Hasan, "Integration Named Entity Recognition and Latent Dirichlet Allocation to Enhance Topic Modeling," *Annals of Emerging Technologies in Computing*, vol. 9, no. 2, pp. 20–30, 2025, https://doi.org/10.33166/AETiC.2025.02.002.

[9] M. Theofany Aulia Anwar, S. Hadi Wijoyo, and W. Hayuhardhika Nugraha Putra, "Implementasi Metode TextRank dan Named Entity Recognition Untuk Ekstraksi Kata Kunci Pada Media Online Berita," *Jurnal Sistem Informasi, Teknologi Informasi, dan Edukasi Sistem Informasi*, vol. 5, no. 1, pp. 34–41, 2024, https://doi.org/10.25126/justsi.v5i1.401.

[10] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, S. M. Al-Ghuribi, and F. A. Ghanem, "Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling," *IEEE Access*, vol. 10, no. October, pp. 105328–105351, 2022, https://doi.org/10.1109/ACCESS.2022.3211396.

[11] A. J. Sugiarto, "Perlindungan Tindak Bullying Yang Terjadi Di Kalangan Pelajar," *Jurnal Inovasi Global*, vol. 1, no. 1, pp. 26–31, 2023, https://doi.org/10.58344/jig.v1i1.4.

[12] S. Aini and W. Rahardjo, "Perilaku Cyberbullying pada Remaja Ditinjau Dari Empati dan Regulasi Emosi," *Jurnal Ilmu Perilaku*, vol. 7, no. 2, pp. 121–139, 2024, https://doi.org/10.25077/jip.7.2.121-139.2023.

[13] W. Septina and S. Q. Ain, "Kecerdasan Interpersonal Siswa dengan Perilaku Verbal Bullying di Kelas V Sekolah Dasar," *Jurnal Imiah Pendidikan dan Pembelajaran*, vol. 6, no. 3, pp. 536–547, 2022, https://doi.org/10.23887/jipp.v6i3.54285.

[14] Alia Firda Rayani, Anastasya Dumyanti, Okta Febri Lestari, and Raja Oloan Tumanggor, "Strategi Edukasi Anti-Bullying Anak Usia Dini melalui Psikoedukasi di RPTRA Satria Jelambar," *Jurnal Pendidikan dan Sastra Inggris*, vol. 5, no. 2, pp. 216–228, 2025, https://doi.org/10.55606/jupensi.v5i2.5201.

[15] F. S. A. Ningsih, H. Hudaniah, and S. N. Rokhmah, "Pengaruh body shaming terhadap body image remaja perempuan," *Cognicia*, vol. 11, no. 1, pp. 79–85, 2023, https://doi.org/10.22219/cognicia.v11i1.24983.

[16] P. E. Oktavia, V. Yunitasari, and Y. Wulandari, "Kebijakan Hukum Terkait Tindakan Rasisme Yang Melumpuhkan Sistem Keadilan Di Indonesia," *Jurnal Rechten : Riset Hukum dan Hak Asasi Manusia.*, vol. 1, no. 2, pp. 29–35, 2021, https://doi.org/10.52005/rechten.v1i2.46.

[17] M. S. A. Park, K. J. Golden, S. Vizcaino-Vickers, D. Jidong, and S. Raj, "Sociocultural values, attitudes and risk factors associated with adolescent cyberbullying in east asia: A systematic review," *Cyberpsychology (Brno).*, vol. 15, no. 1, pp. 1–19, 2021, https://doi.org/10.5817/CP2021-1-5.

[18] U. Khaira, R. Johanda, P. E. P. Utomo, and T. Suratno, "Sentiment Analysis Of Cyberbullying On Twitter Using SentiStrength," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 3, no. 1, p. 21, 2020, https://doi.org/10.24014/ijaidm.v3i1.9145.

[19] G. Machado, "Web Scraping: Data search and retrieval methodology in information science," *Academic Education Navigating the Path of Knowledge*, 2023, https://doi.org/10.56238/sevened2023.008-002.

[20] A. Srirahayu and L. S. Pribadie, "Review Paper Data Mining Klasifikasi Data Mining," *Jurnal Ilmiah Informatika Global*, vol. 14, no. 1, 2023, doi: https://doi.org/10.36982/jiig.v14i1.2981.

[21] Budhi Gustiandi, *Langkah Awal Menguasai Bahasa Pemrograman Python*. 2023, https://doi.org/10.55981/brin.656.

[22] P. Chyan *et al.*, *Pengantar Machine Learning* T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997, ISBN: 0-07-042807-7.

[23] X. Pan and Y. Xue, "Advancements of Artificial Intelligence Techniques in the Realm About Library and Information Subject - A Case Survey of Latent Dirichlet Allocation Method," *IEEE Access*, vol. 11, no. October, pp. 132627–132640, 2023, https://doi.org/10.1109/ACCESS.2023.3334619.

[24] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the Real World: A Survey on NLP Applications," *Information (Switzerland)*, vol. 14, no. 4, 2023, https://doi.org/10.3390/info14040242.

[25] M. Adi Widianto, Eka Pebriyanto, Fitriyanti, "Kesamaan Dokumen menggunakan Frekuensi Istilah-Dokumen Terbalik Representasi Frekuensi dan Kesamaan Kosinus," *Jurnal Dinda*, vol. 4, no. 2, pp. 149–153, 2024, https://doi.org/10.20895/dinda.v4i2.1589.