Analysis of Public Sentiment Towards LGBT on Twitter Social Media Using Naïve Bayes Method

Yudhi Franata[1]*, Rizal[2], Rizki Suwanda[3]
University of Malikussaleh[1], [2], [3]
Lhokseumawe, Aceh
yudhi.180170124@mhs.unimal.ac.id[1], rizal@unimal.ac.id[2], rizkisuwanda@unimal.ac.id[3]

***Abstract*— The advancement of information technology and the widespread use of social media have provided a platform for individuals to express their views on various social issues, including those related to Lesbian, Gay, Bisexual, and Transgender (LGBT) topics. This study aims to assess public sentiment towards LGBT issues on Twitter by employing the Naïve Bayes classification algorithm. Relevant tweets were collected through web scraping based on specific LGBT-related keywords within a defined time frame. The collected data underwent several preprocessing stages, including data cleaning, tokenization, stopword removal, and stemming. The processed data were then categorized into three sentiment classes: positive, negative, and neutral. Naïve Bayes was chosen for its effectiveness and efficiency in handling large-scale textual data. The analysis revealed that negative sentiment toward LGBT issues was predominant, although a considerable portion of tweets expressed neutral and positive sentiments. These findings offer valuable insights for policymakers, social activists, and academics in understanding public perception and formulating more effective communication strategies related to LGBT discourse in Indonesia. The classification model achieved an accuracy of 57%, precision of 52%, recall of 100%, and an F1-score of 68%. While the Naïve Bayes approach proved capable in sentiment classification, the model's accuracy could be further enhanced through improved data preparation or the application of more advanced algorithms.**

Keywords*— *Sentiment Analysis, LGBT, Twitter, Naïve Bayes.

I. INTRODUCTION

Social media has become an integral part of modern society, serving as a platform for expressing opinions, sharing entertainment, and disseminating information. Among the many social issues discussed online, the topic of LGBT (Lesbian, Gay, Bisexual, and Transgender) has gained significant attention. Since the 1990s, the term LGBT has replaced the word "homosexual" to better represent the diversity of gender identities and sexual orientations. In Indonesia, the LGBT issue remains highly controversial, especially as members of the LGBT community have become increasingly visible through various campaigns and promotional activities on digital platforms [1]. This cultural shift has drawn considerable attention due to its perceived misalignment with the values held by the majority of Indonesians[2].

In particular, Twitter, a microblogging platform, has played a significant role in the conversation surrounding LGBT issues. With its fast dissemination and wide reach, Twitter allows users to share their thoughts in short but impactful messages known as tweets. The platform offers a unique window into the dynamics of public opinion. Events such as the proposed LGBT-themed "Paris Parade" in Indonesia [3] and the police intervention in an alleged LGBT community party in Bogor [4] demonstrate that social media functions not only as a discussion space but also as a stage for societal conflicts.

Sentiment analysis, a computational technique used to identify public attitudes towards specific topics, has become a valuable tool for analyzing these dynamics. By utilizing natural language processing techniques, sentiment analysis categorizes textual data into positive, negative, or neutral sentiments, while also detecting emotional states such as anger, sadness, or happiness. In this study, the Naïve Bayes classification algorithm was selected for its efficiency and accuracy in processing large-scale and diverse textual datasets.

Previous research has demonstrated the effectiveness of Naïve Bayes in sentiment analysis tasks. For instance, a study by [5] on public opinions regarding COVID-19 vaccination in Indonesia using Naïve Bayes achieved an accuracy of 93%. Another study by Ridwan (2019), which employed the Support Vector Machine (SVM) algorithm to analyze public sentiment on LGBT issues, found that public opinion was predominantly neutral and negative, with the highest accuracy of 74% achieved using a linear kernel.

The Naïve Bayes algorithm remains a preferred method due to its strong classification performance with relatively low computational complexity. According to [6], Naïve Bayes consistently delivers reliable results in real-world applications. In the context of public opinion toward LGBT issues, this algorithm can provide valuable insights and support the development of more effective communication strategies by policymakers.

Based on this background, the present study is entitled "Public Sentiment Analysis on LGBT Issues on Twitter Using the Naïve Bayes Method." This research aims to explore Indonesian public opinion towards LGBT-related discussions on Twitter and contribute to a better understanding of societal perspectives, which can inform academic research, social

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2400, Copyright ©2025
Submitted : Mey 26, 2025, Revised : June 4, 2025, Accepted : June 13, 2025, Published : July 28, 2025

advocacy, and policy development.

II. LITERATURE REVIEW

A. Previous Research

Many studies have shown that sentiment analysis using machine learning—especially the Naïve Bayes algorithm—can be highly effective. For example, Nurdin et al. (2021) used Naïve Bayes to classify student academic papers and achieved an average accuracy of 86.68% [7]. Similarly, Adek et al. (2020) applied the same method to analyze perfume product reviews on Bukalapak.com, reaching an impressive accuracy of 96.44% [8]. In the realm of social media, Adek et al. (2018) explored sentiment classification on Twitter using unigram, bigram, and trigram models, finding that accuracy improved as the size of the training data increased [9]. Kosasih and Alberto (2021) also applied Naïve Bayes in their analysis of game product reviews on Shopee, combining it with TF-IDF and achieving 80.22% accuracy [10]. Meanwhile, Karami et al. (2021) developed an automatic classification system for LGBT Twitter profiles, including those featuring explicit content, and reported around 88% accuracy [11]. These findings collectively underscore the adaptability and dependability of the Naïve Bayes classifier in a wide range of applications. Its strength lies in its solid probabilistic foundation, straightforward implementation, and reliable performance in predictive tasks.

B. LGBT

LGBT stands for Lesbian, Gay, Bisexual, and Transgender, a term that has been used since the 1990s to replace the more limited term "homosexual." Each identity within LGBT represents different sexual orientations or gender expressions, and in Indonesia, the existence of LGBT individuals remains controversial. Although human rights advocates support equal treatment, resistance still prevails due to cultural and religious factors. Campaigns promoting LGBT awareness face significant challenges in a Muslim-majority country like Indonesia[12].

C. Sentiment Analysis

Sentiment analysis is a method for evaluating public opinions on specific topics by categorizing text as positive or negative. Typically applied to social media data, this technique uses machine learning algorithms to classify large volumes of textual content. Naïve Bayes is suitable for this task as it allows for manual evaluation of training data and delivers clear sentiment labels such as P for positive and N for negative sentiments[13].

D. Text Mining

Text mining refers to the process of extracting valuable insights and patterns from unstructured textual data. Unlike structured data mining, text mining must first undergo preprocessing to clean and standardize the text. Hermanto & Noviriandini A (2021) explained that this process includes information extraction, clustering, classification, and visualization, enabling researchers to analyze opinions, emotions, and behaviors related to certain topics or entities[14].

E. Text Preprocessing

Text preprocessing is an essential step in text mining aimed at eliminating noise and improving classification accuracy[15]. This process involves several steps, including case folding, tokenization, stopword removal, and stemming, to prepare the text for further analysis. It helps standardize text input, making it more suitable for machine learning models. The effectiveness of sentiment analysis often hinges on the quality of preprocessing applied to the raw text data.

F. Naïve Bayes Classification

The Naïve Bayes algorithm is a popular choice for text classification because it's both simple and surprisingly effective. By applying Bayes' theorem, it estimates how likely a piece of text belongs to a certain category based on how often specific words appear. Simorangkir and Lhaksmana highlighted how well this method handles large volumes of tweets, offering fast processing and dependable predictions. What makes Naïve Bayes especially appealing is its solid foundation in probability and how easy it is to implement in real-world applications.[16].

$$P(H|X) = \frac{P(H|X)P(H)}{P(X)} \quad (1)$$

Unknown:

X	: Unknown class data
H	: Special class data (Hypothesized data X)
$P(H\|X)$	: Probability of hypothesis H on X
$P(H)$	: Probability of hypothesis H
$P(X\|H)$	: Probability of hypothesis X on H
$P(X)$	: Probability of hypothesis X

The flowchart of the classification process is shown in the following figure:

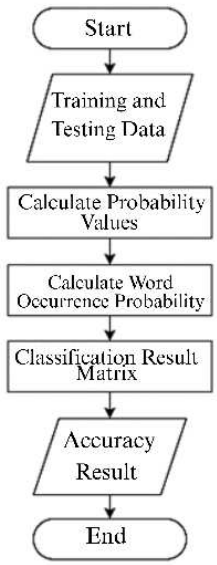

Fig. 1. flowchart of the classification process Naïve Bayes

The steps in the Naïve Bayes classification method are as follows:

- Calculate the probability of each document category

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2400, Copyright ©2025
Submitted : Mey 26, 2025, Revised : June 4, 2025, Accepted : June 13, 2025, Published : July 28, 2025

using:

$$P(cj) = \frac{Nc}{N} \quad (2)$$

- Calculate the probability of each word appearing in a category using:

$$P(wi|cj) = \frac{f(wi|cj)+1}{f(cj)+|V|} \quad (3)$$

- Determine the class with the highest probability using:

$$VMAP = \text{argmax}_{Cj \pounds v} P(cj)\ \Pi i \text{P(wi|cj)} \quad (4)$$

- Evaluate model performance using a confusion matrix consisting of accuracy, precision, and recall.

G. Confusion Matrix

To assess the performance of classification models, a confusion matrix is commonly used. It measures accuracy, precision, and recall by comparing the predicted outcomes with the actual values. Muktafin et al. (2020) emphasized that high precision and recall indicate a strong model, while additional metrics like AUC (Area Under Curve) and ROC (Receiver Operating Characteristic) curves help visualize classification accuracy and performance across different thresholds[17].

TABLE I. CONFUSION MATRIX MODEL

Correct Classification	Classified as	
	+	-
+	True positives (A)	False negatives (B)
-	False positives (C)	True negatives (D)

The classification results are evaluated using a confusion matrix, with the following metrics:

- **Accuracy** = (A + B) / (A + B + C + D)
- **Precision** = A / (C + A)
- **Recall** = A / (A + D)

III. RESEARCH METHODS

A. Time of Research Implementation

This research was conducted during the period June 2023 to September 2023. In this time span, all stages ranging from data collection, system design, algorithm implementation, to model evaluation are carried out gradually and systematically.

B. Problem Formulation

The LGBT phenomenon in Indonesia is a topic that generates both support and opposition in society. Therefore, this research aims to develop a sentiment analysis system for public opinion on Twitter using the Naïve Bayes algorithm, in order to identify trends in public attitudes towards the issue.

C. Research Steps

This research uses a waterfall approach that includes the stages of data collection, data preprocessing, classification, and system testing. The process flow is explained visually through a flow chart to facilitate understanding of the series of research activities.

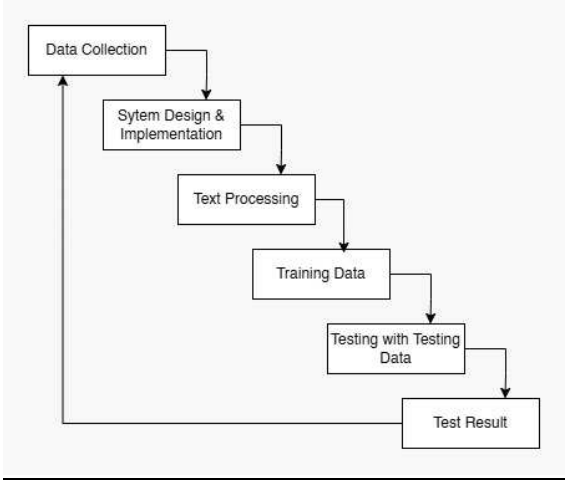

Fig. 2. Research Steps

D. System Requirements Analysis

System requirements analysis includes identifying the hardware and software specifications used during the system development and testing process. The hardware used is a laptop with AMD A10 processor, 8 GB RAM, and 1 TB HDD, which is sufficient to handle data crawling, text preprocessing, and classification model training. Meanwhile, the software used includes Windows 10 Home 64-bit operating system, Visual Studio Code as a text editor, and Python as the main programming language supported by various additional libraries for web scraping, machine learning, and text preprocessing needs. All of these requirements become a reference in ensuring the system can run efficiently and optimally.

E. System Scheme

The general sentiment analysis system scheme can be seen in the figure below:

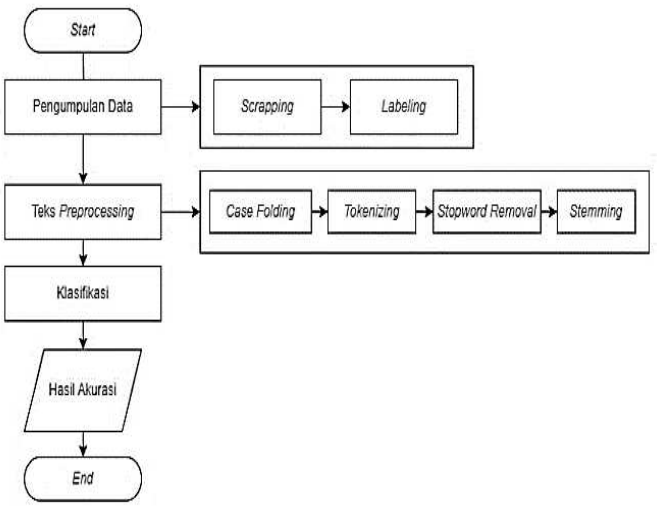

Fig. 3. System Scheme

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2400, Copyright ©2025
Submitted : Mey 26, 2025, Revised : June 4, 2025, Accepted : June 13, 2025, Published : July 28, 2025

F. System Schematic of Naïve Bayes Method

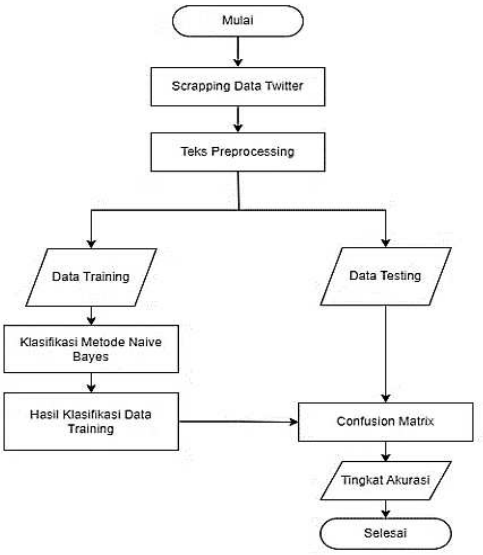

Fig. 4. System Schematic of Naïve Bayes Method

The diagram above shows the overall workflow of the process. It starts with collecting data through scraping using the Twitter API. Once the data is gathered, it goes through preprocessing and is labeled based on sentiment. After that, the dataset is split—80% is used to train the model, while the remaining 20% is reserved for testing. A Naïve Bayes classifier is then built using the training data and evaluated on the test set. To assess how well the model performs, a confusion matrix is used to calculate its accuracy (Afifah & Voutama, 2023).

IV. RESULTS AND DISCUSSION

A. Research Results

This study conducted sentiment analysis of Indonesian public opinion related to LGBT using data from the X.com platform. Data was collected through crawling techniques using Python, with authentication in the form of auth_token from X.com. The collected sentiments cover the period 2023 to 2024 and are processed using the Naïve Bayes algorithm. The implementation results produced three sentiment categories, namely positive, negative, and neutral. Model evaluation is done by measuring the level of accuracy as an indicator of the feasibility of the Naïve Bayes method in classifying public opinion on LGBT issues.

B. System Analysis

System analysis is carried out to decompose the problem into smaller components so that it is easy to understand and evaluate. The main problem in this research is how to identify public opinion on LGBT automatically. Therefore, a sentiment analysis system based on Naïve Bayes algorithm was built by utilizing crawling data from X.com. The system receives input in the form of tweets, then performs preprocessing to clean the text, before finally being classified into sentiment categories. The system is developed using PHP and Python programming languages and MySQL database, and the final result is sentiment prediction and its accuracy level.

C. Basic Model Design

1) Context Diagram

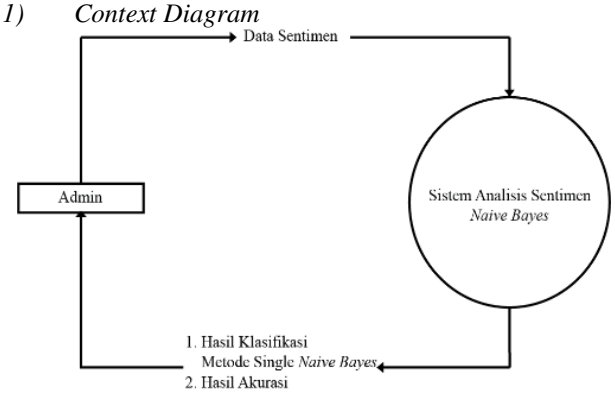

Fig. 5. Context Diagram

2) Data Flow Diagram (DFD)

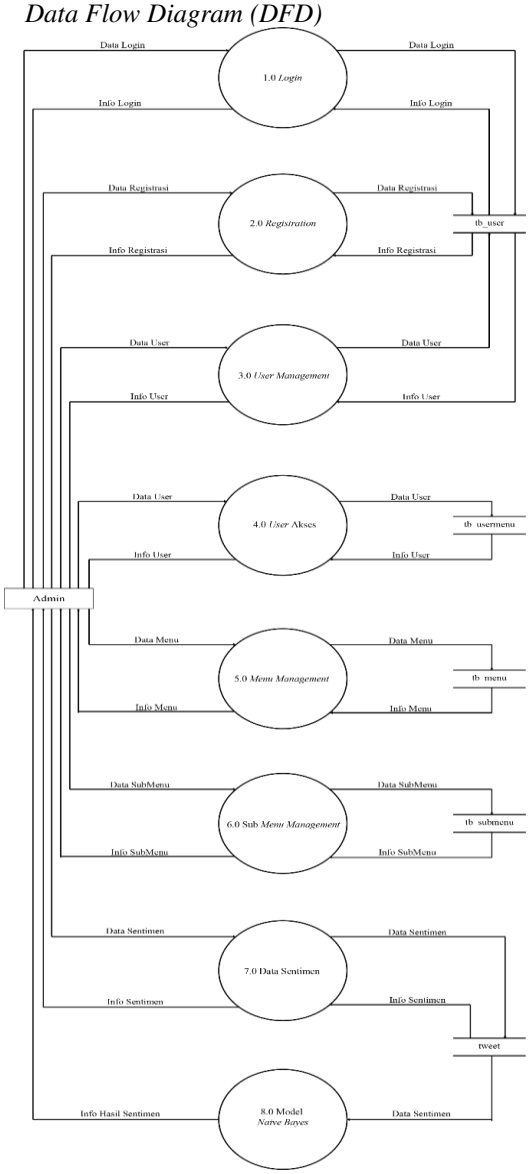

Fig. 6. Data Flow Diagram (DFD)

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2400, Copyright ©2025
Submitted : Mey 26, 2025, Revised : June 4, 2025, Accepted : June 13, 2025, Published : July 28, 2025

3) Entity Relationship Diagram (ERD)

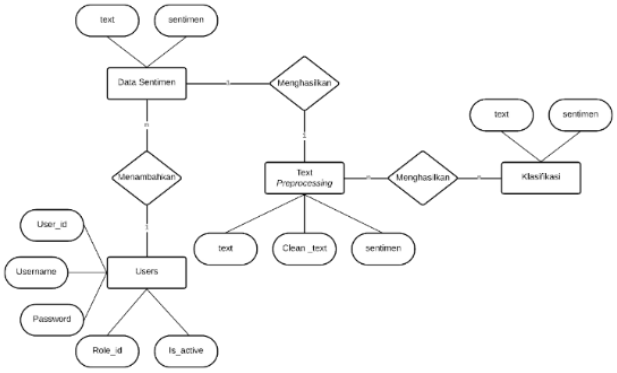

Fig. 7. Entity Relationship Diagram (ERD)

4) Performance Evaluation of the Naïve Bayes Model Using a Confusion Matrix

To evaluate the performance of the Naïve Bayes model, a confusion matrix is used that measures accuracy, precision, recall, and F1-score, and provides an overview of the model's ability to classify positive, negative, and neutral sentiments.

TABLE II. USERS TABLE

Actual \ Predicted	Positive	Negative	Neutral	Total
Positive	100	0	0	100
Negative	50	30	20	100
Neutral	7	5	40	100
Total	192	47	61	300

a) Accuracy

$$Accuracy = \frac{100 + 40}{100 + 40 + 57 + 50} = 0.5676 \approx 57\%$$

b) Precision

$$Precision = \frac{100}{100 + 92} = 0.5208 \approx 52\%$$

c) Recall

$$Recall = \frac{100}{100 + 0} = 100\%$$

d) F1-Score

$$F1 - Score = 2 \times \frac{0.52 \times 1}{0.52 + 1} = \times 0.3421 \approx 68\%$$

D. Database Management

This research uses MySQL as a database management system to facilitate structured data processing and storage. The use of the right DBMS supports the relationship between data and facilitates the process of managing the system.

1) Users Table

TABLE III. USERS TABLE

No	Name	Type	Width	Notes
1	*id_ulselr*	*int*	11	*Primary Key*
2	*ulselrnamel*	*varchar*	50	
3	*password*	*varchar*	256	
4	*id_role*	*int*	11	
5	*is_active*	*int*	11	

2) MCenu Table

TABLE IV. MCENU TABLE

No	Name	Type	Width	Notes
1	id_menu	*int*	11	*Primary Key*
2	menu	*varchar*	50	

3) Sub Menu Table

TABLE V. SUB MENU TABLE

No	Name	Type	Width	Notes
1	id_sub_menu	*int*	11	*Primary Key*
2	submenu	*varchar*	50	
3	url	*varchar*	100	
4	id_menu	*int*	11	*Foreign Key*

4) User Menu Table

TABLE VI. USER MENU TABLE

No	Name	Type	Width	Notes
1	id_user_menu	*int*	11	*Primary Key*
2	id_user	*int*	11	*Foreign Key*
3	id_menu	*int*	11	*Foreign Key*

5) Role Table

TABLE VII. ROLE TABLE

No	Name	Type	Width	Notes
1	id_role	*int*	11	*Primary Key*
2	role	*varchar*	50	

E. Discussion

1) Testing the Naïve Bayes Method

The testing carried out in this research is to use the Naïve Bayes model. In the testing process, the system requires research data, namely sentiment data obtained by crawling the X.com application.

a) Research Data

Research data is data taken by crawling the X.com application. The data in the research is in the form of X user sentiment data on LGBT from 2023 to 2024. In this process the system also performs automatic data labeling using AutoTokenizer.

b) Preprocessing Data

Data preprocessing is the initial stage in data processing which aims to clean, change, and prepare raw data so that it is ready to be used in the Naïve Bayes modeling process.

c) Naïve Bayes Implementation

The implementation of the Naïve Bayes algorithm in this study produces a classification model which is then evaluated using the following performance metrics:

Accuracy = 57%
Precision = 52%
Recall = 100%
F1-score = 68%

F. System Implementation

The implementation of the sentiment analysis system follows a structured process as outlined in the flowchart. Each stage of the system plays a crucial role in ensuring the model functions as expected.

1) Data Collection

The data collection process involves scraping 300 tweets related to LGBT from Twitter over a three-month period (January to March 2023). This data was collected using Python's web scraping libraries with specific LGBT-related keywords.

2) System Design & Implementation

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2400, Copyright ©2025
Submitted : Mey 26, 2025, Revised : June 4, 2025, Accepted : June 13, 2025, Published : July 28, 2025

The system was developed using Python for machine learning tasks and PHP for the web interface. The design focuses on creating a user-friendly system for data collection, processing, and sentiment analysis with Naïve Bayes.

3) Text Processing

Text preprocessing steps included tokenization, stopword removal, and stemming. After cleaning the raw data, around 250,000 words remained, ready for use in training the model.

4) Training Data

80% of the cleaned data, about 240,000 words, was used for training. The data was categorized into three sentiment classes: positive, negative, and neutral.

5) Testing with Testing Data

The remaining 20% of the data (60,000 words) was used for testing. The model's predictions were compared with actual results to assess its accuracy and effectiveness.

6) Test Result

The model's performance was evaluated, resulting in an accuracy of 57%, precision of 52%, recall of 100%, and an F1-score of 68%.

The system developed in this research is a web-based application that serves to perform sentiment analysis of public opinion on LGBT using Naïve Bayes algorithm. The system is built using PHP and Python programming languages, and supported by MySQL database. Users can login and access various features, such as managing sentiment data, preprocessing text, and viewing sentiment classification results. One of the important parts of this system is the dashboard page that displays a summary of the amount of data, classification results, and model performance in the form of an informative and easy-to-use interface, as shown in Figure below:

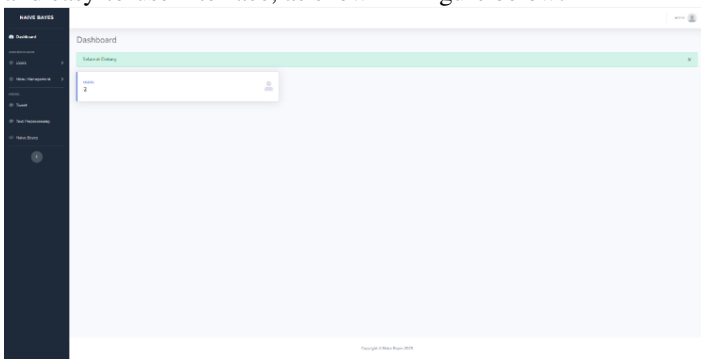

Fig. 8. Dashboard

The following is also the appearance of the Access User Page:

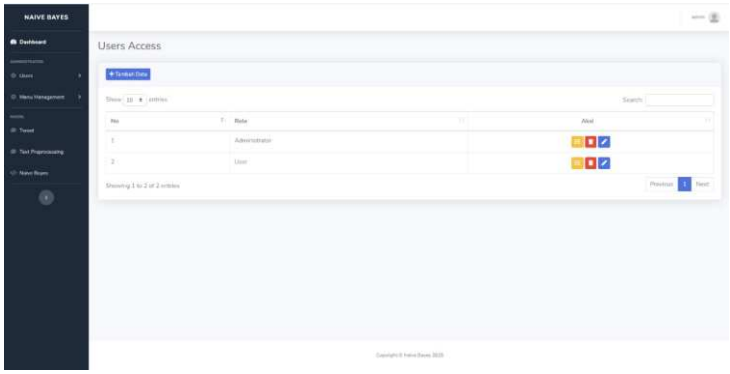

Fig. 9. Access User Pages

V. CONCLUSION

This study successfully developed a sentiment analysis system to capture Indonesian public opinion on LGBT issues through Twitter. The system was built using the Naïve Bayes algorithm and involved several key stages: data collection, text preprocessing, sentiment labeling, and classification. The evaluation results indicate that the Naïve Bayes algorithm can classify tweets into positive, negative, and neutral categories, achieving an accuracy of 57%, a precision of 52%, a perfect recall of 100%, and an F1-score of 68%. These findings suggest that while the Naïve Bayes method shows promise in handling sentiment classification tasks, particularly in identifying relevant data, there is still significant room for improving overall accuracy—potentially through model refinement or by exploring more advanced algorithms.

REFERENCES

[1] R. Palupi, M. H. Rahmansyah, G. M. Arasta, and G. Irhamdhika, "Isu LGBT Dalam Bingkai Media Online (Analisis Framing Robert Entman Pada Pemberitaan RKUHP LGBT Pada Tempo. co Dan BBBCIndonesia. com)," *Jurnal Media Penyiaran*, vol. 2, no. 2, pp. 148–156, 2022.

[2] E. Andina, "Faktor psikososial dalam interaksi masyarakat dengan gerakan lgbt di indonesia," *Aspirasi: Jurnal Masalah-masalah Sosial*, vol. 7, no. 2, pp. 173–185, 2019.

[3] S. Amalia, "Kelompok LGBT Indonesia Ikut Serta dalam Paris Pride," 2019.

[4] M. Sholihin, "Polisi Tak Beri Izin Kegiatan LGBT Berkedok Gathering di Puncak Bogor.," vol. 1, 2022.

[5] W. Yulita, E. D. Nugroho, and M. H. Algifari, "Analisis sentimen terhadap opini masyarakat tentang vaksin covid-19 menggunakan algoritma naïve bayes classifier," *Jurnal Data Mining dan Sistem Informasi*, vol. 2, no. 2, pp. 1–9, 2021.

[6] W. Zhang and F. Gao, "An improvement to naive bayes for text classification," *Procedia Eng*, vol. 15, pp. 2160–2164, 2011.

[7] N. Nurdin, M. Suhendri, Y. Afrilia, and R. Rizal, "Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (NBC)," *SISTEMASI: Jurnal Sistem Informasi*, vol. 10, no. 2, pp. 268–279, 2021.

[8] R. Rizal, M. Fikry, and A. Helmina, "Opinion Mining About Parfum on E-Commerce Bukalapak. Com Using the Naïve Bayes Algorithm," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 6, no. 1, pp. 107–114, 2020.

[9] R. Tjut Adek and S. Nasution, "Tweet Clustering in Indonesian Language Twitter Social Media using Naive Bayes Classifier Method," *Eurasian Journal of Analytical Chemistry (Abbrev. Eurasian J Anal Chem. or EJAC)*, vol. 13, no. 6, pp. 277–284, 2018.

[10] R. Kosasih and A. Alberto, "Sentiment analysis of game product on shopee using the TF-IDF method and naive bayes classifier," *ILKOM Jurnal Ilmiah*, vol. 13, no. 2, pp. 101–109, 2021.

[11] A. Karami, M. Lundy, F. Webb, H. R. Boyajieff, M. Zhu, and D. Lee,

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2400, Copyright ©2025
Submitted : Mey 26, 2025, Revised : June 4, 2025, Accepted : June 13, 2025, Published : July 28, 2025

"Automatic Categorization of LGBT User Profiles on Twitter with Machine Learning," *Electronics (Basel)*, vol. 10, no. 15, p. 1822, 2021.

[12] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," *Procedia Comput Sci*, vol. 161, pp. 765–772, 2019.

[13] L. A. Fudholi, N. Rahaningsih, and R. D. Dana, "Sentimen Analisis Perilaku Penggemar Coldplay Di Media Sosial Twitter Menggunakan Metode Naive Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 3, pp. 4150–4159, 2024.

[14] H. Hermanto and A. Noviriandini, "Analisa Sentimen Terhadap Belajar Online Pada Masa Covid-19 Menggunakan Algoritma Support Vector Machine Berbasis Particle Sarm Optimization," *Jurnal Informatika Kaputama (JIK)*, vol. 5, no. 1, pp. 129–136, 2021.

[15] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text preprocessing for text mining in organizational research: Review and recommendations," *Organ Res Methods*, vol. 25, no. 1, pp. 114–146, 2022.

[16] H. Simorangkir and K. M. Lhaksmana, "Analisis Sentimen pada Twitter untuk Games Online Mobile Legends dan Arena of Valor dengan Metode Naïve Bayes Classifier," *eProceedings of Engineering*, vol. 5, no. 3, 2018.

[17] E. H. Muktafin, K. Kusrini, and E. T. Luthfi, "Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing," *Jurnal Eksplora Informatika*, vol. 10, no. 1, pp. 32–42, 2020.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2400, Copyright ©2025
Submitted : Mey 26, 2025, Revised : June 4, 2025, Accepted : June 13, 2025, Published : July 28, 2025