Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Web-Based Chatbot Development and User Satisfaction Analysis using the Naive Bayes Method Through Online Questionnaires

Nurholis[1]*, Willy Prihartono[2], Fathurrohman[3]

[1,2,3] *STMIK IKMI Cirebon*
olisdragneel@gmail.com [1*]

Abstract

This study aims to develop a web-based chatbot using Natural Language Processing (NLP) technology and the Naive Bayes algorithm to enhance digital interaction quality. User satisfaction was evaluated through an online survey involving 202 university students, focusing on ease of use, response speed, and relevance. The research followed the CRISP-DM framework, including data preprocessing (case folding, tokenization, stopword removal, and stemming), text transformation using the TF-IDF method, and implementation of a Naive Bayes classification model. an F1-score of 84%. Sentiment analysis revealed predominantly positive feedback, reflecting user satisfaction with the chatbot's ease of use and response accuracy. However, some limitations, such as insufficient contextual understanding, were identified. These findings provide valuable insights into NLP-based chatbot development to support effective digital interactions. The proposed chatbot demonstrates potential applications in customer service, education, and e-commerce, with future improvements suggested to enhance contextual comprehension and scalability.

Keywords: *Web-Based Chatbot, Natural Language Processing, Naive Bayes, User Satisfaction, Sentiment Analysis, TF-IDF*

1. Introduction

The rapid advancement of digital technology has driven innovation in user-service interactions, with chatbots emerging as one of the leading tools. A web-based chatbot powered by Natural Language Processing (NLP) can automatically understand and respond to human language, improving service efficiency[1]. However, the success of a chatbot is not solely determined by technological sophistication but also by user satisfaction during interactions [2]. Poor user experiences may reduce engagement, highlighting the need for thorough evaluation and improvement of chatbot performance [3].

Web-based chatbots often face challenges in providing relevant responses due to limited contextual understanding and variations in user input [4]. Despite advancements in NLP technology, many chatbots fail to deliver natural and responsive interactions, which can decrease user satisfaction[5]. Users generally expect fast and easy-to-use services, but poorly designed chatbots may leave them feeling confused or dissatisfied[6]. This makes it critical to evaluate user satisfaction systematically and incorporate user feedback into the development process[7].

This study focuses on developing a web-based chatbot using the Naive Bayes algorithm for NLP and evaluating user satisfaction through an online survey. The research aims to provide insights into chatbot improvement, ensuring effective, responsive, and user-friendly digital interactions. By addressing these challenges, this study contributes to enhancing the adoption of chatbots across various sectors[8], [9].

2. Research Methods

Study employs a quantitative method to develop a web-based chatbot using the Naive Bayes algorithm and to evaluate user satisfaction through an online questionnaire. The quantitative approach was chosen to ensure that the results could be analyzed objectively and measurably.Naive Bayes is an efficient classification algorithm for processing text data to determine relevant responses based on the probability of word occurrences. This algorithm was implemented to enable the chatbot to process text quickly and provide appropriate answers. [10], [11], [12].

2.1. Research Design

This study employs a descriptive research design to evaluate user satisfaction and analyze sentiment regarding the chatbot. The research focuses on text data collected through open-ended questionnaire responses. The data is processed using Natural Language Processing (NLP) techniques, with the Naive Bayes algorithm implemented to classify user feedback into sentiment categories (positive and negative).

2.2. Place and Time of Research

The study was conducted online as the chatbot operates as a web-based system. Data collection took place during October to November 2024, involving participants who interacted with the chatbot and provided their feedback

2.3. Population and Sample

The population in this study consists of chatbot users, particularly those who tested the chatbot for casual conversation. The sample was selected using a purposive sampling technique, involving 202 participants who voluntarily provided feedback through an online questionnaire. Participants were chosen based on their engagement with the chatbot and their willingness to share their impressions.

2.4. Research Instruments

The main research instrument used in this study was an **online questionnaire**, which consisted of a single open-ended question:

- "Bagaimana kesan anda secara keseluruhan tentang penggunaan chatbot untuk kegiatan sehari-hari?"

This question was designed to gather detailed and qualitative user opinions about the chatbot's performance, relevance, and usability.

2.5. Rating Sheet

The research followed these steps to collect data:

1. **Chatbot Testing**: Users interacted with the web-based chatbot developed for casual conversation.
2. **Feedback Submission**: Participants submitted their feedback through an online questionnaire.
3. **Data Compilation**: The responses from 202 participants were gathered and prepared for analysis.

2.6. Analysis Results

The data obtained from the questionnaires were analyzed quantitatively. A descriptive analysis was conducted to summarize user satisfaction levels based on their responses

2.7. Data Analysis Techniques

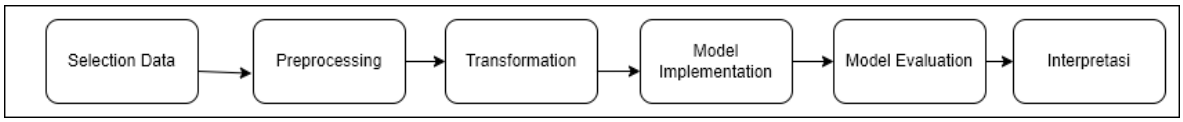

Fig 1: KDD diagram

Analysis process in this study followed these steps:

1. **Selection Data**: Responses from **202 users** were gathered using an online questionnaire. Participants provided feedback on their overall impression of the chatbot.
2. **Preprocessing:** The responses were cleaned and organized by **case folding**, **tokenization**, **stopword removal**, and **stemming** to prepare them for further analysis.
3. **Transformation**: The text data were transformed into numerical values using the **TF-IDF** method, which converts text into a structured format suitable for analysis by machine learning algorithms.
4. **Model Implementation**: The **Naive Bayes classifier** was applied to the transformed data to classify user feedback into categories such as **positive** or **negative**.
5. **Model evaluation**: The model was evaluated based on **accuracy**, **precision**, **recall**, and **F1 score** to measure its performance in classifying user feedback.
6. **Interpretation**: The evaluation results were interpreted to provide insights into user satisfaction and identify areas for improvement in the chatbot's design and functionality.

4. Result and Discussion

3.1. Result

This study aims to develop a web-based chatbot by implementing Natural Language Processing (NLP) technology. The analysis was conducted through a series of stages, including data collection from questionnaires, data preprocessing, Naive Bayes model implementation, model evaluation, and result interpretation. The study focuses on measuring the chatbot's accuracy and response relevance, as well as evaluating user satisfaction when interacting with the developed system. The steps undertaken and the findings of this research are explained in detail in this report.

3.1.1. Selection Data

At this stage, data selection is carried out to ensure that the data used in the research meets the relevance and quality criteria. Research data was obtained from an online questionnaire that was distributed to active semester 8 students at STMIK IKMI Cirebon.

Table 1: Selection data

Cap waktu	Nama	Ulasan	Rating
2024/11/16 7:01:20 PM GMT+7	Tia Kawaii	Lumayan	4
2024/11/16 7:02:59 PM GMT+7	Lixie	Bagus banget,mayan lahh buat temen chat kalau lagi gabut	5
2024/11/16 7:03:19 PM GMT+7	Malik Ibrahim	Desain tampilan udah bagus tapi masih ada beberapa respon yang tidak sesuai	3
2024/11/16 7:10:05 PM GMT+7	Rahmat Abdillah	Keren	5

3.1.2. Preprocessing data

At this stage, data preprocessing is carried out to prepare the data obtained from the questionnaire so that it can be used in further analysis. The preprocessing process is carried out to ensure the data is clean, relevant, and can be processed with the Naive Bayes algorithm. The preprocessing steps carried out include:

3.1.1.1. Case folding

The technique of changing each letter in text to lowercase is called case folding (lowercase).

Table 2: Case folding

No	Ulasan	Case folding
1.	Bagus banget,mayan lahh buat temen chat kalau lagi gabut:)	bagus banget,mayan lahh buat temen chat kalau lagi gabut
2.	Desain tampilan udah bagus tapi masih ada beberapa respon yang tidak sesuai	desain tampilan udah bagus tapi masih ada beberapa respon yang tidak sesuai
3.	Sangat membantu dalam mencari tugas atau keperluan lainnya	sangat membantu dalam mencari tugas atau keperluan lainnya

3.1.1.2. Tokenize

The next step involves breaking the text into individual words (tokens) using **tokenization techniques**. Tokenization helps in understanding the structure of sentences and prepares the data for feature analysis.

Table 3: Tokenize

No	Ulasan	Token
1.	Bagus banget mayan lahh buat temen chat kalau lagi gabut	Bagus, banget, mayan, lahh, buat, temen, chat, kalua, lagi, gabut
2.	Lumayan seru untuk temen gabut	Lumayan, seru, untuk, temen, gabut
3.	Desain tampilan udah bagus tapi masih ada beberapa respon yang tidak sesuai	Desain, tampilan, udah, bagus, tapi, masih, ada, beberapa, respon, yang, tidak, sesuai

3.1.1.3. Stopword

Common words such as *"and," "or,"* and *"in,"* which do not carry significant meaning for analysis, are removed using a **stopword list**. This removal helps highlight more meaningful words for further analysis.

Table 4: Stopword

No	Ulasan	Stopword
1.	desain tampilan udah bagus tapi masih ada beberapa respon yang tidak sesuai	desain tampilan udah bagus respon tidak sesuai
2.	lumayan seru untuk temen gabut	lumayan seru temen gabut
3.	bagus banget mayan lahh buat temen chat kalau lagi gabut	bagus banget mayan lahh temen chat gabut

3.1.1.4. Stemming

After the processes of case folding, tokenization, and stopword removal, the next step is **stemming**, which involves removing affixes from words to determine their root forms.

Table 5: Stemming

No	ulasan	stemming
1.	desain tampilan udah bagus tapi masih ada beberapa respon yang tidak sesuai	desain tampil udah bagus masih beberapa respon tidak sesuai
2.	lumayan seru untuk temen gabut	lumayan seru temen gabut
3.	lumayan membantu	Lumayan bantu

3.1.3. Transformation

The data transformation stage is a crucial step in data processing, aimed at converting raw data into a format suitable for analysis. In this study, the Term Frequency-Inverse Document Frequency (TF-IDF) method was applied, which assigns weights to each word in the document based on its frequency of occurrence and its uniqueness across the entire document. As a result, the text derived from the questionnaire responses can be represented as numerical vectors, which are then utilized by the Naive Bayes algorithm for classification. The transformation process involves the following steps:

1. **Term Frequency (TF):**
   Term Frequency calculates the frequency of a word (term) in a document.
   Formula:
   $$TF(t,d) = \frac{Number\ of\ occurrences\ of\ term\ t\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d}$$
   Example: If the word "good" appears 5 times in a document containing 50 words:
   $$TF = \frac{5}{50} = 0.1$$
2. **Inverse Document Frequency (IDF):**
   IDF reduces the influence of common words in the analysis by assigning higher weights to less frequently used words across all documents.
   Formula:
   $$IDF(t) = log\ log\ \frac{N}{n_t}$$
   Where:
   N = total number of documents
   n_t = number of documents containing term ttt
   Example: If the word "chatbot" appears in 10 out of 100 documents:
   $$IDF(t) = log\ log\ \frac{100}{10} = log\ log\ 10$$
3. **TF-IDF**
   The final weight of each word, which reflects its importance in a specific document, is the product of TF and IDF.
   Formula:
   $$TF - IDF(t,d) = TF(t,d) \times IDF(t)$$

3.1.4. Model Implementasion

Before training the model, the data needs to be split into two parts: training data and testing data. This split is performed using the train_test_split method from the scikit-learn library. The following are the steps and results of the split:

1. **Train data:** The training data, which accounts for 70% of the total dataset, is used to train the Naive Bayes model. In the example above, the size of the training data is 141 rows.
2. **Test data:** The testing data, which accounts for 30% of the total dataset, is used to evaluate the model's performance. In the example above, the size of the testing data is 61 rows.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 0)
print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(141, 3)
(61, 3)
(141,)
(61,)
```

Fig. 2: Train data and test data

- **X_train.shape:** Displays the dimensions of the training data features, which is (141, 3).
- **X_test.shape:** Displays the dimensions of the testing data features, which is (61, 3).
- **y_train.shape:** Displays the dimensions of the training data labels, which is (141,).
- **y_test.shape:** Displays the dimensions of the testing data labels, which is (61,).

3.1.5. Model Evaluation

The model is evaluated using standard metrics such as **accuracy**, **precision**, **recall**, and **F1-score**. **Accuracy** indicates how often the model correctly classifies the responses based on the questionnaire data. **Precision** and **recall** are used to assess how well the model can identify specific categories within the questionnaire data. The **F1-score**, which combines precision and recall, provides a comprehensive overview of the model's performance.

```
Evaluasi Multinomial Naive Bayes
Accuracy: 0.8524590163934426
F1 Score: 0.8517777304662552
Precision: 0.8515801138751959
Recall: 0.8524590163934426

Classification Report:
              precision    recall  f1-score   support

     negatif       0.82      0.78      0.80        23
     positif       0.87      0.89      0.88        38

    accuracy                           0.85        61
   macro avg       0.84      0.84      0.84        61
weighted avg       0.85      0.85      0.85        61
```

Fig. 3: Classification report

The classification report presents the **precision**, **recall**, and **F1-score** for each class (positive and negative). With an average **F1-score** of **85%** and an overall accuracy of **85%**, the model has performed well in classifying sentiment.

Evaluation using a matrix

- **Accuracy**
  The model's accuracy is **85.25%**, indicating that it performs well in classifying both positive and negative sentiments in the given data.
- **Precision**

$$Precision = \frac{TP}{TP + FP} = \frac{34}{34 + 5} = 0.8516$$

  The model's **precision** is **85.16%**, which indicates the proportion of positive predictions that are truly positive.
- **Recall**

$$Recall = \frac{TP}{TP + FN} = \frac{34}{34 + 5} = 0.8525$$

  Recall model adalah 85.25%, Ini menunjukkan jumlah data positif yang berhasil dikenali dengan benar oleh model.

- **F1-score**

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = 2 \cdot \frac{0.8516 \cdot 0.8525}{0.8516 + 0.8525} = 0.8518$$

  The model's **F1-Score** is **85.18%**, which provides a balanced measure between precision and recall.

In this study, to evaluate the performance of the Naive Bayes model, a Confusion Matrix is used. This matrix helps to visualize the comparison between correct (true) and incorrect (false) predictions, both for positive and negative sentiment categories. The Confusion

Matrix shows how many data points were misclassified (false positive and false negative) and how many were correctly classified (true positive and true negative). With this information, we can assess the quality of the model built and ensure whether it works optimally on the provided data.

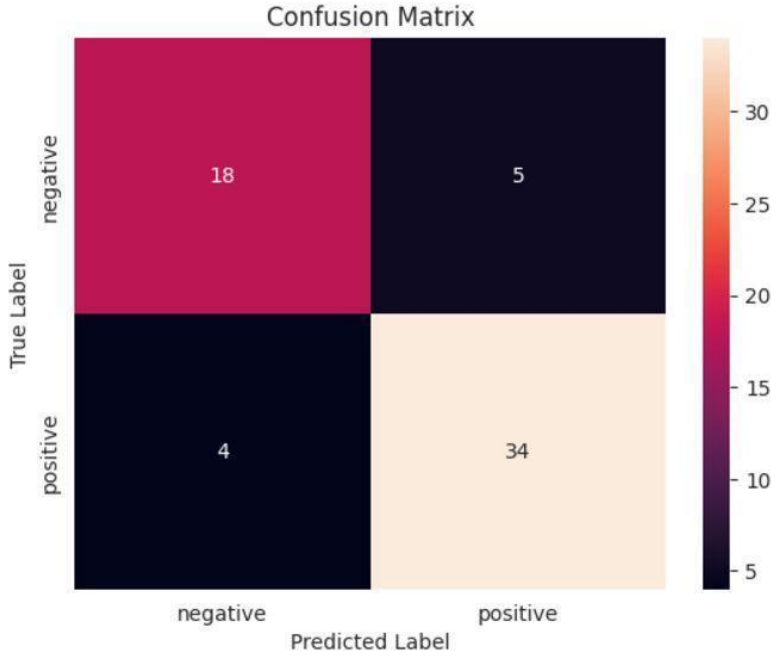

Fig. 4: Confusion matrix

From the **Confusion Matrix** above, we can identify the following values:

- True Positive (TP) = 34
- True Negative (TN) = 18
- False Positive (FP) = 5
- False Negative (FN) = 4

Based on the evaluation results, it can be concluded that the **Naive Bayes algorithm** performs well in classifying sentiment from the provided questionnaire data. Although there are some minor errors in the predictions, this model can be relied upon for sentiment analysis on similar data in the future.

3.1.6. Interpretation

After the evaluation, the sentiment analysis results are interpreted to provide an understanding of user opinions regarding the chatbot program. The majority of the feedback indicates positive sentiment, reflecting satisfaction with using the chatbot as a conversation partner. However, some negative feedback identifies issues such as responses that the bot couldn't understand and the limited number of keywords. These findings offer a clear picture of the aspects that users appreciate and areas that require improvement in the Olis AI chatbot. This can be seen in the table below.

Table 6: Interpretation

	Precision	**Recall**	**F1-score**
Negatif	82%	78%	80%
Positif	87%	89%	88%
Accuracy			85%
Marco Avg	84%%	84%	84%
Weighted Avg	85%	85%	85%

The table provides insights into the performance of the Naive Bayes model, highlighting its strength in classifying positive sentiment while identifying areas for improvement in handling negative sentiment. Overall, the model demonstrates good performance for sentiment analysis. However, improvements in the keyword database and contextual understanding are necessary to enhance the chatbot's accuracy in generating appropriate responses.

3.2. Disussion

The findings of this study highlight the performance of the Naive Bayes model in classifying user sentiment regarding the chatbot. With an accuracy of 85.25%, the model demonstrated strong performance in differentiating between positive and negative sentiments. The

precision, recall, and F1-score values for each sentiment class further reinforce the model's reliability. Specifically, the model achieved higher precision and recall for positive sentiment compared to negative sentiment, suggesting that the chatbot effectively meets user expectations for most interactions.

The Confusion Matrix analysis reveals that the model performs better in predicting positive sentiment, as indicated by a higher F1-score of 88% for the positive class compared to 80% for the negative class. This discrepancy highlights areas where the chatbot struggles, such as handling more complex or ambiguous user inputs.

Qualitative feedback from users provides additional insights into the chatbot's strengths and limitations. Positive responses emphasize the chatbot's role as an engaging and helpful conversational partner, particularly in its ease of use and speed of response. However, negative feedback identifies specific challenges, including:

1. The chatbot's limited ability to understand complex questions or contextually ambiguous inputs.
2. The minimal keyword database, which restricts its capacity to provide relevant and accurate responses.

These findings suggest that while the chatbot delivers satisfactory performance overall, there are areas requiring improvement. Enhancing the chatbot's keyword database and refining its contextual understanding capabilities will be critical in addressing the identified shortcomings. Furthermore, incorporating more advanced NLP techniques, such as contextual embeddings or deep learning models, could improve the chatbot's ability to process and respond to diverse inputs more effectively.

Overall, the results demonstrate that the Naive Bayes algorithm provides a solid foundation for sentiment analysis in chatbot development. However, continuous improvements based on user feedback are necessary to optimize the chatbot's performance and ensure it meets the evolving expectations of its users.

4. Conclusion

This study successfully developed a web-based chatbot using Natural Language Processing (NLP) and evaluated its performance through user sentiment analysis. The Naive Bayes algorithm demonstrated its effectiveness in classifying user feedback into positive and negative sentiments, achieving an accuracy of 85.25%. The evaluation metrics, including precision, recall, and F1-score, further confirmed the reliability of the model, with the chatbot performing particularly well in identifying positive sentiment.

The majority of user feedback indicated positive experiences, highlighting the chatbot's strengths in providing relevant responses and serving as a helpful conversational companion. However, the study also identified areas for improvement, such as the chatbot's limited understanding of complex queries and a restricted keyword database, which led to some inaccurate responses.

Overall, the findings demonstrate that the chatbot performs well for its intended purpose and can serve as a foundation for further development. Future improvements should focus on expanding the keyword database, enhancing contextual understanding, and incorporating more advanced NLP techniques to handle broader and more complex conversational contexts. These enhancements will improve the chatbot's accuracy, user satisfaction, and adaptability for real-world applications.

References

[1] R. Nur Aziza *et al.*, "Pembangunan Aplikasi dan Klasifikasi Pertanyaan Chatbot Informasi Akademik Menggunakan Metode Cosine Similarity dan Naïve Bayes," vol. 12, no. 2, pp. 169–179, 2024, [Online]. Available: https://doi.org/10.33322/kilat.v12i2.1921

[2] R. A. Sekarwati, A. Sururi, R. Rakhmat, M. Arifin, and A. Wibowo, "Survei Metode Pengujian Chatbot pada Media Sosial untuk Mengukur Tingkat Akurasi," *Sisfotenika*, vol. 11, no. 2, p. 172, 2021.

[3] M. Muliyono and S. Sumijan, "Identifikasi Chatbot dalam Meningkatkan Pelayanan Online Menggunakan Metode Natural Language Processing," *Jurnal Informatika Ekonomi Bisnis*, vol. 3, pp. 142–147, 2021, doi: 10.37034/infeb.v3i4.102.

[4] Y. Christian and M. Erline, "Web-Based Chatbot With Natural Language Processing and Knuth-Morris-Pratt (Case Study: Universitas Internasional Batam)," *JST (Jurnal Sains dan Teknologi)*, vol. 11, no. 1, pp. 132–141, 2022, doi: 10.23887/jstundiksha.v11i1.43258.

[5] A. L. Maitri and J. Sutopo, "Rancangan Bangun Chatbot Sebagai Pusat Informasi Lembaga Kursus Dan Pelatihan Menggunakan Pendekatan Natural Language Processing," *Eprints.Uty.Ac.Id*, pp. 1–9, 2019, [Online]. Available: http://eprints.uty.ac.id/

[6] D. Apriliani, S. F. Handayani, and I. T. Saputra, "Implementasi Natural Language Processing (NLP) Dalam Pengembangan Aplikasi Chatbot Pada SMK YPE Nusantara Slawi," *Techno.Com*, vol. 22, no. 4, pp. 1037–1047, 2023, doi: 10.33633/tc.v22i4.9155.

[7] Y. Yunefri, Y. E. Fadrial, and S. Sutejo, "Chatbot Pada Smart Cooperative Oriented Problem Menggunakan Natural Language Processing dan Naive Bayes Classifier," *INTECOMS: Journal of Information Technology and Computer Science*, vol. 4, no. 2, pp. 131–140, 2021, doi: 10.31539/intecoms.v4i2.2704.

[8] C. Diantoni, R. Mufidah, and H. Triana, "Membangun Chatbot Untuk Informasi Magang Dan Studi Independen Kampus Merdeka Dengan Algoritma Naive Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 2, pp. 1389–1397, 2024, doi: 10.36040/jati.v8i2.8962.

[9] Pardede, A. M. H. (2018). Perancangan Sistem Pakar Diagnosa Penyakit Tanaman Kelapa Sawit Dengan Metode Bayes Study Kasus PT. Ukindo Blankahan Estate.

[10] Abdullah, D., Zarlis, M., Pardede, A. M. H., Anum, A., Suryani, R., Parwito, ... & Setiyadi, D. (2019, November). Expert System Diagnosing Disease of Honey Guava Using Bayes Method. In *Journal of Physics: Conference Series* (Vol. 1361, No. 1, p. 012054). IOP Publishing.

[11] Sawitri, S., Simanjuntak, M., & Pardede, A. M. H. (2024). APPLICATION OF NAIVE BAYES METHOD TO DIAGNOSE FMD DISEASE IN GOATS. *Journal of Mathematics and Technology (MATECH)*, *3*(2), 103-111.

[12] Lestari, S. A. M., Pardede, A. M., & Simanjuntak, M. (2024). Prediksi Disleksia pada Anak menggunakan Metode Naive Bayes. *Jurnal Kajian dan Penelitian Umum*, *2*(5), 37-51.