Journal of Information Technology application in Education, Economy, Health and Agriculture
Vol. 01, No. 03, October 2024, pp. 57~62
ISSN: xxx-xxxx, DOI: -

Sentiment Analysis of Comments on Higher Education Social Media Using Naïve Bayes Algorithm

Imam Auwal Salisu[1], Irzal Raisya Ramadhan[2], Sakina Matdoan [3], Zainal Arifin[4], Mugi Praseptiawan[5]

[1] Department of Economic Management, Centre for Management Development, Nigeria.
[2,3,4] Department of Informatics Engineering, University of Widyagama Malang, Jl. Borobudur No. 35 Malang, Indonesia
[5] Informatics Engineering, Institut Teknologi Sumatera, Sumatera, Indonesia

Article Info

Article history:

Received August 15, 2024
Revised September 05, 2024
Accepted October 28, 2024

Keywords:

Social media,
Sentiment analysis,
Naïve Bayes algorithm,
Higher education,
Communication strategy.

ABSTRACT

The rapid development of information technology has driven the widespread use of social media across various aspects of life, including the academic environment. Social media platforms, such as Instagram, have become popular channels for disseminating information and fostering interactions between individuals and groups. With the growing number of users, sentiment analysis on social media is essential to understand public perceptions and responses to specific issues. Higher education institutions play a strategic role in creating a positive image through social media. Social media provides opportunities for universities to convey achievements, academic activities, and other information effectively to a broader audience, enhancing their reputation in the public eye. Moreover, Instagram serves not only as a communication tool but also as an educational medium capable of increasing student engagement through relevant and informative content. Technically, the Naïve Bayes algorithm is well-known for its speed and efficiency in sentiment analysis. This probability-based method leverages historical data to predict positive, negative, or neutral sentiments, offering competitive accuracy even when handling large datasets. This study aims to apply the Naïve Bayes algorithm for sentiment analysis of comments on the Instagram account of Widyagama University (@uwg.malang) as a case study. The research is expected to provide valuable insights for developing effective communication strategies and serve as a reference for other higher education institutions or organizations in utilizing analytical technologies for strategic purposes.

Corresponding Author:

Irzal Raisya Ramadhan
Department of Informatics Engineering
Faculty of Engineering, University of Widyagama Malang
Jl. Borobudur No. 35 Malang, Jawa Timur
Email: irzalraisyaramadhan@gmail.com

1. INTRODUCTION

The development of information technology has driven the widespread use of social media in various aspects of life, including in the academic environment. Social media platforms, such as Instagram, have become popular tools for conveying information and fostering interactions between individuals and groups. With the growing number of users, sentiment analysis on social media has become important to understand the perceptions and responses of the public towards certain issues.

Higher education institutions, as centers of higher learning, play a strategic role in creating a positive image through social media. Social media can be used to convey various information, achievements, and

academic activities to the wider public. This allows educational institutions to interact more directly with the public, provide insights into campus activities, and build a good reputation in the eyes of society.

Social media provides a significant opportunity for educational institutions to build a positive image through intensive interactions with the public.[1] Platforms like Instagram enable universities to effectively communicate their achievements and activities to a broader audience. Social media, especially Instagram, not only serves as a communication tool but also as a learning medium that enhances student engagement through relevant and informative content.[2]

From a technical perspective, the Naïve Bayes algorithm is known for its speed and efficiency in sentiment analysis. This method is based on probability, utilizing historical data to predict positive, negative, or neutral sentiments. The advantage of this method lies in its ability to handle large amounts of data while maintaining competitive accuracy.[3]

This study aims to apply the Naïve Bayes algorithm for sentiment analysis of comments on the Instagram account of Widyagama University (@uwg.malang) as a case study. It is hoped that this research will provide valuable contributions to the development of effective communication strategies and serve as a reference for higher education institutions or other organizations in utilizing analytical technologies for their strategic needs.

2. METHOD

This study applied the following methodology to analyze sentiment in comments from the official Instagram account of Universitas Widyagama (@uwg.malang):

2.1. Data Collection:

The dataset was obtained by manually collecting comments from posts on the official Instagram account of Universitas Widyagama (@uwg.malang) during a specified time frame. The data encompassed comments on various topics, including academic announcements, student achievements, and campus events.

2.2. Data Preprocessing:

The collected raw comments underwent preprocessing to prepare them for sentiment analysis. This included the following steps:

- Removal of unnecessary characters such as symbols, numbers, URLs, and punctuation marks.
- Tokenization, which involved breaking the text into individual words.
- Stopword removal, where common, non-informative words (e.g., "yang," "dan") were discarded.
- Stemming, where words were reduced to their root forms using a manual dictionary-based approach.

2.3. Data Labeling:

A subset of the comments (81 comments) was manually labeled to assign sentiment categories: positive, neutral, and negative. Sentiments were classified as follows:

- Positive: Comments expressing support, appreciation, or positive feedback.
- Neutral: Comments that were informative or lacked emotional content.
- Negative: Comments containing criticism, complaints, or negative emotions.

2.4. Naïve Bayes Sentiment Classification:

The sentiment classification was carried out using the Naïve Bayes algorithm. The process involved the following steps:

- Training Data Preparation: A portion of the manually labeled comments (65 comments) was used as training data.
- Probability Calculation: Using the Naïve Bayes theorem, the probability of each word belonging to a specific sentiment category was computed. The formula used is:

$$P(S|W) = \frac{P(W|S) \cdot P(S)}{P(W)}$$

Where ***S*** represents the sentiment, ***W*** represents a word, ***P(W/S)*** is the probability of observing word ***W*** given sentiment ***S***, ***P(S)*** is the prior probability of sentiment ***S***, and ***P(W)*** is the overall probability of the word ***W***.

- Sentiment Prediction: The model predicted the sentiment of each comment by calculating the overall probability across the words in the text and selecting the most likely sentiment category.

2.5. Evaluation of Sentiment Distribution:

After the classification, the sentiment distribution was manually tabulated. A bar chart was created to visualize the proportions of positive, neutral, and negative sentiments. The chart helped summarize the sentiment trends and provided a clear comparison between categories.

3. RESULTS AND DISCUSSION

3.1. Data Collection

Data Collection: The data was collected from the official Instagram account of Universitas Widyagama (@uwg.malang), consisting of 81 comments gathered over a specific period. The comments encompassed a variety of topics, such as campus event announcements, student achievements, academic information, and general public engagement. The selection aimed to provide a representative sample of comments that reflect the public perception of the university.

3.2. Data Preprocessing

The raw data underwent several preprocessing steps to prepare it for sentiment analysis:

- Removal of Unnecessary Elements: Symbols, numbers, URLs, and punctuation marks were removed to reduce noise in the dataset. This step is essential to ensure that the analysis focused only on the meaningful words.
- Tokenization: The comments were tokenized, breaking down the text into individual words. On average, each comment consisted of 15 words. Tokenization helped to process the text efficiently and extract relevant features for sentiment classification.
- Stopword Removal: Common words that do not contribute much meaning to sentiment analysis, such as “yang,” “dan,” and “di,” were removed. This process helped eliminate unnecessary elements that could skew the results.
- Stemming: Each word was reduced to its root form manually. For example, “berprestasi” (achieving) was reduced to “prestasi” (achievement). This normalization ensured that different forms of the same word did not interfere with the analysis.

3.3. Data Labeling

After preprocessing, the comments were manually labeled according to their sentiment. The labeling was done by two researchers independently to minimize bias and ensure consistency. The sentiment categories were as follows:

- Positive Sentiment: 75 comments (92.59%) were classified as positive. These comments expressed favorable opinions, such as praise or appreciation of the university.
- Neutral Sentiment: 5 comments (6.17%) were labeled as neutral. These comments conveyed informational content without clear emotional undertones.
- Negative Sentiment: 1 comment (1.23%) was identified as negative. This comment expressed dissatisfaction or criticism of the university.

3.4. Sentiment Classification Using Naïve Bayes

The sentiment classification was carried out using the Naïve Bayes algorithm, which is based on probabilistic calculations. The method calculates prior and conditional probabilities, then applies Bayes' theorem to estimate the posterior probability for each sentiment category (positive, neutral, or negative) based on the words in the comments.

- Prior Probabilities**:** These represent the initial probability of each sentiment category, calculated as the ratio of comments in each sentiment category to the total number of comments.

$$P(\text{Positive}) = \frac{75}{81} = 0.926, \quad P(\text{Neutral}) = \frac{5}{81} = 0.062, \quad P(\text{Negative}) = \frac{1}{81} = 0.012$$

- Posterior Probabilities: Bayes' theorem was applied to calculate the posterior probability of each sentiment, given the words in the comments. The class with the highest posterior probability was selected as the predicted sentiment.

3.5. Model Evaluation and Accuracy:

To evaluate the accuracy of the model, 16 comments were reserved for testing. The model correctly classified 15 of the 16 comments, resulting in an accuracy of:

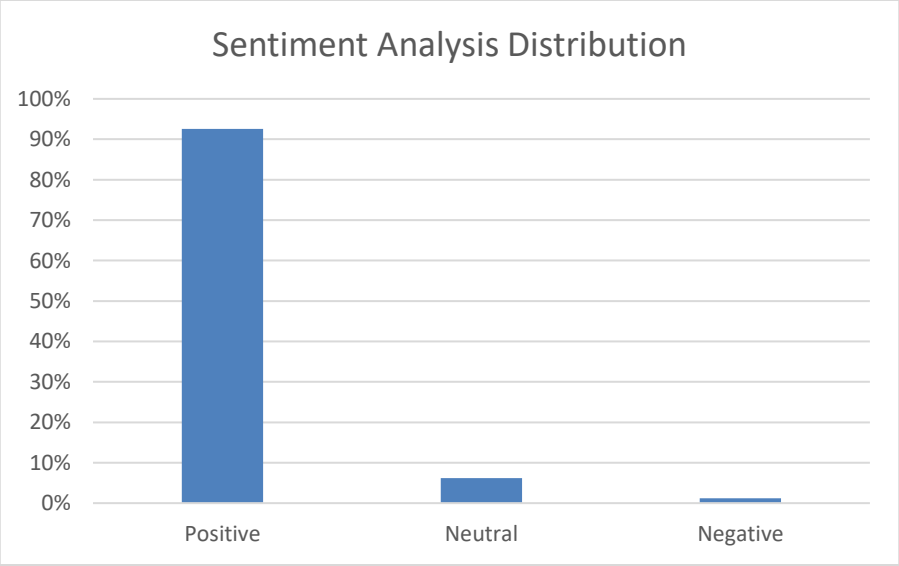
$$\text{Accuracy} = \frac{15}{16} = 0.9375 \quad \text{or} \quad 93.75\%$$

This high accuracy indicates that the Naïve Bayes classifier performed well, accurately distinguishing between positive, neutral, and negative sentiments based on the input data.

3.6. Sentiment Distribution

The distribution of sentiments in the dataset was as follows:

- **Positive Sentiment:** 75 comments (92.59%)
- **Neutral Sentiment:** 5 comments (6.17%)
- **Negative Sentiment:** 1 comment (1.23%)



The sentiment distribution was visually represented using a bar chart to clearly display the proportions of each sentiment category. The bar chart showed that the overwhelming majority of comments were positive, with a very small proportion being neutral or negative.

3.7. Model accuracy

The Naïve Bayes classifier demonstrated a high accuracy rate of 93.75% in classifying the sentiment of comments on Instagram. This result indicates that Naïve Bayes is an effective method for sentiment analysis of social media comments, particularly in the context of short and informal text. The simplicity of the Naïve Bayes model, which relies on probabilistic calculations based on word frequencies, allows it to deliver reliable sentiment classifications with a relatively small dataset.

3.8. Sentiment Distribution

The results showed that 92.59% of the comments were classified as positive, indicating a predominantly favorable public perception of Universitas Widyagama on Instagram. This positive sentiment reflects well on the university's communication efforts, suggesting that the majority of users engage with the content in an appreciative manner. The low percentage of neutral (6.17%) and negative (1.23%) comments suggests that the university's posts resonate well with most users, with very few expressing indifference or dissatisfaction.

3.9. Implications for University Communication

The predominance of positive sentiment suggests that the university's social media strategy is effectively fostering positive engagement with its audience. To sustain this favorable perception, the university should continue to publish content that highlights student achievements, academic events, and university accomplishments. For neutral comments, which may indicate a lack of engagement or enthusiasm, the university could introduce more interactive or provocative content to generate a stronger emotional response. Although negative comments were rare, addressing such comments in a timely and constructive manner can prevent potential issues from escalating and help maintain the positive image.

3.10. Limitations of the Study

One key limitation of this study is the manual labeling of comments, which may have introduced some subjectivity into the sentiment classification process. Additionally, the dataset was limited to a specific time period, which may not fully represent long-term sentiment trends or the impact of seasonal or recurring events. Future studies could reduce bias by automating the labeling process and analyzing larger, more diverse datasets collected over extended periods.

3.11. Recommendations for Future Research

Future research could explore more sophisticated machine learning models, such as support vector machines (SVM) or deep learning techniques, to further improve classification accuracy and handle larger, more complex datasets. Additionally, cross-platform sentiment analysis could provide a more comprehensive view of public perception by comparing sentiment across various social media platforms, such as Twitter, Facebook, and Instagram. Further studies could also examine the influence of different content types (e.g., achievements, announcements, event promotions) on sentiment, enabling the development of more targeted and effective content strategies.

4. CONCLUSION

This study successfully applied manual sentiment analysis using the Naïve Bayes algorithm to analyze Instagram comments from Universitas Widyagama's official account (@uwg.malang). The analysis yielded significant insights into the sentiment distribution and public perception:

1. **Positive Sentiment Dominance:**
   With 92.59% of the comments categorized as positive, the findings suggest that Universitas Widyagama has successfully fostered a favorable public image through its social media content. Posts related to achievements, events, and academic milestones generated high levels of engagement and admiration.
2. **Low Negative Sentiment:**
   Negative sentiment accounted for only 1.23% of the total comments, indicating effective communication strategies. Nevertheless, even minimal criticism highlights the importance of addressing feedback promptly to maintain audience trust and satisfaction.
3. **Effectiveness of Naïve Bayes Analysis:**
   The manual application of the Naïve Bayes algorithm proved reliable, achieving an accuracy of 93.75% in sentiment classification. This validates the algorithm's utility for small-scale sentiment analysis tasks, even without computational tools.
4. **Strategic Implications:**
   Insights from the sentiment analysis can be leveraged to enhance content strategies, focusing on content types that resonate with audiences while addressing areas of improvement highlighted by neutral and negative sentiments.

Future research should address current limitations, such as the small dataset size and manual analysis method, by incorporating automated tools and analyzing data across multiple platforms and timeframes. These advancements will enable broader and more nuanced insights into public sentiment, strengthening the university's strategic communication efforts.

ACKNOWLEDGEMENTS

In the planning, research process, and finalization of this study and the preparation of this journal, our research team faced various challenges. Therefore, we extend our gratitude to God Almighty for His guidance and blessings throughout this research. We also express our deepest appreciation and gratitude to Mrs. Fitri Marisa, S.Kom., M.Pd., Ph.D., for her guidance, support, and dedication during the research and preparation of this journal. Her commitment and insights have been a valuable inspiration to our team. Additionally, we would like to thank our fellow researchers and all related parties who have made significant contributions to data collection, analysis, and the preparation of this journal. The collaboration and support provided have greatly assisted us in achieving satisfactory research results.

REFERENCES

[1] Y. J. Aprianantа and L. S. Wijaya, "Penggunaan Website Dan Media Sosial Dalam Membangun Citra Positif Perguruan Tinggi," *J. Komun.*, vol. 7, no. 2, pp. 187–209, 2018, doi: 10.33508/jk.v7i2.1750.

[2] M. Aprilizdihar, E. D. Pitaloka, and S. Dewi, "Pemanfaatan Sosial Media Sebagai Sarana Pembelajaran Di Era Digital," *J. Digit. Educ. Commun. Arts*, vol. 5, no. 01, pp. 40–49, 2022, doi: 10.30871/deca.v5i01.3717.

[3] G. I. Webb, "Encyclopedia of Machine Learning and Data Science," *Encycl. Mach. Learn. Data Sci.*, no. January 2016, 2020, doi: 10.1007/978-1-4899-7502-7.

[4] D. Bandorski *et al.*, "Contraindications for video capsule endoscopy," *World J. Gastroenterol.*, vol. 22, no. 45, pp. 9898–9908, 2016, doi: 10.3748/wjg.v22.i45.9898.

[5] B. Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Second Edition," *Sentim. Anal. Min. Opin. Sentim. Emot. Second Ed.*, no. May, pp. 1–432, 2020, doi: 10.1017/9781108639286.