Journal of Applied Linguistics (ALTICS)
https://www.e-journal.unper.ac.id/index.php/ALTICS
ISSN 2721-995X (Printed)
ISSN 2721-0316 (Online)
Vol 6, No 1 (2024), pp.41-51

# Sentiment Analysis of Characters Romeo and Juliet by Shakespeare Using Naïve Bayes Algorithm

Ni Made Dewi Cahyadi[1], Dede Brahma Arianto[2]
[1]*English Literature, Faculty of Humanity, Universitas Jenderal Soedirman*
[2]*Master of Informatics, Faculty of Industrial Technology, Universitas Islam Indonesia*
*dewicahyadii@gmail.com*

## Abstract

Sentiment analysis has evolved into an intriguing field of computational literary studies, merging literature and technology to gain fresh insights into classic w\orks. This paper explores the sentiment analysis of characters in Romeo and Juliet by Shakespeare using the Naïve Bayes algorithm, aiming to understand how character sentiments shape the story's development. In this tragic play, emotions among characters play a pivotal role, influencing the narrative's course. However, no prior research has delved into the sentiment analysis of characters in this classic work. This research addresses this gap by providing a profound understanding of character emotions within the broader context of literary works. Moreover, it contributes to the continually evolving field of computational literary studies, emphasizing the importance and urgency of employing contemporary computational techniques to examine emotional nuances in classic literary works. Utilizing Naïve Bayes based on Bayes' theorem and a probabilistic approach, this study achieves an accuracy rate of 81% with precision values of 82% for neutral and 79% for positive, recall scores of 96% for neutral and 73% for positive, and F1-scores 88% for neutral and 76% for positive, demonstrating the model's effectiveness in classifying sentiments with the dataset.

**Keywords:** Applied Linguistic, Sentiment Analysis, Naïve Bayes, Romeo and Juliet, hakespeare.

## INTRODUCTION

The convergence of literature and technology, historically separate domains, has opened new avenues for generating fresh insights. In the field of computational literary studies, sentiment analysis has transcended years of disinterest and neglect to become a dynamic and compelling subject of discourse. Liu (in Pavan, 2022) stated sentiment analysis a study that examines people's opinions, sentiments, emotions, attitudes, and feelings concerning specific subjects. Its primary aim to discern sentiment tendencies, whether they are positive, negative, or neutral. The remarkable ability of sentiment analysis to investigate sentiment trends within various forms of

written text constitutes the link that connects it to literature. Therefore, this paper seeks to explore sentiment analysis concerning characters in literary works.

Sentiment analysis, as defined by Feldman (in Mata et al, 2021), is the task of uncovering the author's opinions on specific entities. Liu (in Pavan, 2022), further elaborates that sentiment analysis is the field of study that examines people's opinions, sentiments, appraisals, attitudes, and emotion towards entities and their attributes. In essence, sentiment analysis is focused on exploring the subjective content in text to gain insights into how individuals perceive and evaluate various entities. Moreover, sentiment analysis is an application of Natural Language Processing (NLP) technology, which trains computer software to comprehend text in a manner like humans.

According to Devi et al. (2020), There are different ways to perform sentiment analysis, from lexicon-based analysis to artificial intelligence and machine learning. Hota et al. (2021) stated the distinction lies in the methodology employed: the lexicon-based approach calculates polarity based on words from a document, while the artificial intelligence and machine learning approach involve training a model using machine learning algorithms based on training data stated by Therdoost et al. (2023). Although these two techniques can be used in conjunction and offer promising results, their combination is rarely utilized in research due to its complexity. Nonetheless, their shared objective is to provide insights into the classification of sentiment tendencies (positive, negative, or neutral.

In the process of sentiment classification, a variety of tools and techniques are utilized to process text and determine sentiment. One aspect of NLP is semantic interpretation, which aims to represent the meaning of a sentence in a context-independent manner for further purposes, including the assignment of sentiment scores to these words. These sentiment scores are subsequently used to classify sentiment tendencies. For instance, confidences scores range from 1 to 0, with scores closer to 1 indicating higher confidence in label classification, while lower score signify lower confidence, and the total predicted score related to labels (positive, negative, and neutral) always sums up to 1 for each document or sentence (Aahil, 2023).

In the realm of sentiment analysis, the Naïve Bayes algorithm emerges as a prominent and accurate classification method rooted in Bayes' theorem. Falling within the realm of machine learning, this algorithm utilizes probability calculations, treating a document as a collection of words without considering their order. This makes it particularly suitable for solving multi-class

problems involving input categories rather than numerical data. The application of the Naïve Bayes algorithm in sentiment analysis involves classifying text based on the probability of words appearing in each sentiment label (Tineges et al, 2023). For instance, if a text contains words like 'love', 'happy', and 'joyful', the Naïve Bayes algorithm tends to classify the text as having a positive sentiment due to the high probability of these words in the positive label. While the Naïve Bayes algorithm is praised for its fast computation, simplicity, and generally high accuracy, it comes with limitations. Siregar et al. (2020) pointed out a weakness in the method, specifically the assumption of class independence, leading to potential inaccuracies. Therefore, additional evidence is often required to substantiate the analysis result when using the Naïve Bayes algorithm.

The play of sentiments among characters is a common element of drama scripts. This occurrence is frequently caused by the necessity for dramatized characters to express a range of emotions, which defines their roles as either antagonist or protagonists. One notable example of a drama script that employs sentiment play is Romeo and Juliet by Shakespeare. This drama, set in the city of Veronica, was first performed in 1957. Romeo and Juliet have garnered significant attention from various audiences due to its tragic love story. The conflict begins with the feud between the Montague and Capulet families, leading to battles and heart-wrenching deaths. Despite the ongoing conflict between these two families, Romeo and Juliet, one from each family, fall in love. They manage to maintain their relationship by secretly marrying, but a series of misunderstandings and fatal events bring their love story to a tragic end. Due to its tragic ending, Romeo and Juliet continues to be one of the most memorable works in the world of literature.

Applying sentiment analysis to literary works provides a unique perspective on understanding the emotional dynamics of characters, as exemplified in the iconic drama, Romeo and Juliet. Romeo is passionate young member of the Montague family who has a complicated conflict with the Capulets. Romeo is known for being an intensely romantic and emotional person. As the story continues, he is portrayed as a passionate and fierce person. Romeo's love for Juliet grows quickly and deeply, demonstrating the height of sincere and intense love. Beyond his passionate desire for marrying Juliet, he also intends to make peace between his family and the rival Capulets, signifying his transformation from idealistic idealist to committed and self-sacrificing partner. Juliet, a member of Capulet family, is shown as a perceptive, devout, and deeply spiritual young lady who grows to love Romeo. After meeting Romeo, she gradually gains courage and the confidence to

follow her heart even though at first, she was initially obedient to her parents. Her persona is one of bravery, unwavering love, and a strong desire to be with her true live—even if it means going against social norms and taking big chances. Juliet struggles to maintain a balance between her love for Romeo and her family's expectations, which leads to the tragedy that accompanies their love story. Together, in Shakespeare's play, Romeo and Juliet symbolize intense, real, and pure love.

## METHOD

The procedures used in research related to a dataset to solve research questions or accomplish research goals are known as research methods. Sugiyono (2013) defines a research method as a scientific approach to gathering data for a particular purpose and use. On the other hand, Priyono (2016) defines a research method as a means of accomplishing a task by carefully applying one's intellect to reach a goal. Thus, research methods can be defined as a methodical approach to gathering data from scientific sources or as a means of conducting a deliberate action to accomplish research goals.

This research employs the qualitative descriptive method because the researcher aims to gain a more detailed and comprehensive understanding of the sentiment outcomes of the characters Romeo and Juliet. By utilizing the qualitative descriptive method, the researcher able to provide a detailed explanation of the context, background, and other supporting factors that contribute to the sentiment outcomes of both Romeo and Juliet.

1. **Data Collecting**

   The initial step in gathering relevant data or facts for the study or analysis that will be conducted.

2. **Teks Processing**

   There are multiple steps in this section, such as:

   1) Tokenization

      The tokenization technique is a technique of breaking down a sentence into individual word units or phrases that utilizes NLP, a machine learning technology that allows computers to understand natural language like humans do. This method is used to accurately represent text in sentiment analysis, which improves algorithm analysis.

2) Removal of special characters

Eliminating punctuation, periods, and commas makes the text to be cleaner and easier to process.

3) Normalization of text

This procedure changed the entire text to lowercase. This is done to standardize the data and prevent mistakes in meaning interpretation. For example, “Good” and “good” will have various interpretations if the language is not normalized.

**3. Sentiment Analysis**

Sentiment analysis, also known as opinion mining, is a process used to determine the emotional tone or attitude expressed within a piece of text. It helps to identify whether the sentiment conveyed is positive, negative, or neutral. In this journal, sentiment analysis is performed using the NLTK (Natural Language Toolkit) library.

NLTK is a widely used library designed to assist in various text processing tasks. One of its notable features is the SentimentIntensityAnalyzer, a pre-trained model that facilitates sentiment analysis on text data. This tool employs a lexicon-based approach, where words within sentence are cross-referenced with a predefined sentiment lexicon to assign sentiment scores.

**4. Machine Learning Development**

Machine learning comprises a range of techniques that prove beneficial for handling and predicting extensive datasets by processing them through learning algorithms (Danukusumo, 2017). One widely used machine learning algorithm is Naïve Bayes Multinomial, which is a probabilistic classification method based on Bayes’ theorem. In this journal, this algorithm is applied in sentiment analysis.

**5. Data Visualization**

The step at which the outcomes of the sentiment analysis of the main characters can be graphically displayed using elements such as charts, diagrams, or other forms of visual representation.

## FINDINGS AND DISCUSSION

During the initial stage of this study, the dataset for the "Romeo and Juliet" script was obtained from Kaggle, consisting of 3313 raw data entries. Kaggle is a website containing a collection of

raw datasets that can be processed. The collected data was then meticulously processed to eliminate NaN (Not a Number), also known as missing values. This resulted in a clean dataset with 3096 entries suitable for further analysis. It is essential to remove columns that contain NaN values, as ignoring it may lead to insufficient data analysis. Furthermore, Warnes (2021) emphasizes that neglecting missing values can cause the conclusions drawn from the analysis results to differ from those obtained with cleaned data.

Afterwards, text processing occurs. Firdaus et al. (2021) emphasizes that text processing is crucial as it helps eliminate unnecessary parts of the data. Therefore, text processing must be conducted to generate more accurate data analysis. The column focused on for text processing is 'Playerline.' This column contains the dialogue between the characters in the play, making it an essential section that will provide insights into the sentiment each character possesses. This column underwent a conversion to the string data type as it originally possessed an object data type. This conversion was necessary for subsequent tokenization, which involves breaking down the text into words or elements based on specific criteria. It also means special characters, punctuation, symbols, and non-alphanumeric characters were removed to ensure the relevance of the text for analysis. Moreover, normalization to lowercase was applied to maintain consistency in text processing.

The next step involved text analysis; to prepare for sentiment analysis, a 'stopwords.words' code was executed. Common words, sometimes referred to as useless words, were removed from the data using the 'stopwords.words' code. Examples of these words include "I," "the," and "be." Ganesan (2023) asserts that these terms are so frequently used that they convey very little meaningful information. Consequently, the data eventually includes only the words required for analysis, providing analysis results that are more precise. After the 'stopwords' code was executed, it was applied to the 'Playerline' column, resulting in the column being cleaned of common words and prepared for further analysis. Following that, sentiment analysis was conducted for every row in the ‘Playerline’ column using the TextBlob library, a Python library typically used for natural language processing (NLP) tasks, in this case, applied for sentiment analysis. The words in the 'Playerline' column were measured with a polarity score that describes the extent to which the text is positive, negative, or neutral. Finally, a new column was added to the data as the result of sentiment analysis to assign sentiment labels.

Continuing with the research, a machine learning model was constructed using Naïve Bayes Multinomial. This is the pivotal aspect of the entire analysis, where constructing this model will offer more precise insights into the sentiments of the characters Romeo and Juliet. Initially, the 'Playerline' column and 'sentiment labels' were extracted as variables x and y; subsequently, they were divided into training and test data. Text from both datasets was transformed into vectors using CountVectorizer, a method employed to calculate word frequencies in the data and represent them in vector form. This process also serves to make the data readable for machine learning.

In the evaluation stage, various metrics, including precision, recall, f1-score, and accuracy, were employed to measure the performance of the sentiment analysis machine learning model. This evaluation stage is crucial to assess the extent to which the sentiment analysis model can predict accurately and consistently, allowing us to gauge the accuracy of our analysis results. The model classified texts into Negative, Neutral, and Positive classes, achieving good precision for Neutral (0.82) and Positive (0.79) classes, while indicating room for improvement in the Negative class (0.78). Recall was very good for Neutral (0.96) and quite good for Positive (0.73) classes but low for Negative (0.22). The F1-score indicated difficulty in achieving a balance between precision and recall for the Negative class (0.34). Overall accuracy stood at 81%, demonstrating the model's capability to correctly classify most texts in the dataset.

Finally, data visualization was employed through a table to present the sentiment analysis of the characters in "Romeo and Juliet." A table is a highly effective visualization method for depicting data distribution. In this context, the table aids in understanding how character sentiment evolves throughout the script and provides a clear picture of sentiment distribution. Additionally, the table assists in comprehending the development of sentiment changes throughout the story. The table of the analysis result can be seen below, negative in nature. Finally, sentiment labels are assigned to the analysis result by adding new column.

The visualization used to present the sentiment analysis of the Romeo and Juliet character is a table. A table is a highly effective visualization method for depicting data distribution. In this context, the table aids in understanding how character sentiment evolves throughout the script and provides a clear picture of sentiment distribution. Additionally, the table assists in comprehending the development of sentiment changes throughout the story. The table of the analysis result can be seen below,

*Table 1. Romeo and Juliet' Sentiment Total*

| No. | Character | Positive | Negative | Netral |
|---|---|---|---|---|
| **1** | Romeo | 145 | 126 | 316 |
| **2** | Juliet | 106 | 106 | 318 |

Based on the table above, both Romeo and Juliet tend to exhibit neutral sentiment. Which means, Romeo and Juliet generally express emotions or statements that are neither overwhelmingly positive nor negative. However, Romeo tends to have a more positive sentiment rather than negative sentiment, which indicates a generally optimistic or favorable tone in his dialogues. On the other hand, in the above table, it is evident that for most of the dialogues, Juliet remains in the neutral sentiment category, even more so than Romeo, which indicates a tendency towards balanced and emotionally restrained expressions. Then, the score of positive and negative sentiment for Juliet is equal, suggesting a harmonious blend of positive and negative emotions in her dialogues, creating a balanced emotional portrayal.

In Shakespeare's play, Juliet is portrayed as an intelligent and clever character. Furthermore, Juliet's sentiment throughout the play tends to be neutral. One of the data supporting Juliet's neutral sentiment is found at data line 777 with the player line, 'A rhyme I learn'd even now,' labeled as neutral sentiment. It can be concluded that Juliet's intelligence enables her to perceive situations from various perspectives, making her a character who remains impartial in her dialogues due to her knowledge. However, in addition to the reasons, Juliet's tendency towards neutral sentiment is also influenced by her confusion in facing the conflict between the Capulet and Montague families.

Romeo, on the other hand, commits several misdeeds, such as sneaking into Juliet's house and even killing Juliet's cousin, Tybalt. However, among Romeo's positive and negative sentiments, the positive ones tends to dominate. One of the reasons positive sentiments dominates Romeo, as expressed by Benvolio in data line 1683, is a scene in the play that mentions, 'Romeo, that spoke him fair … All this uttered with gentle breath, calm look, knees humbly bowed …' this indicates a character development in Romeo, where he initially attempts to achieve peace with Tybalt and resolve the long-standing family feud. Unfortunately, Romeo's peaceful efforts are rebuffed by Tybalt, as he is, still in the perspective of Benvolio, '… Tybalt, deaf to peace…'. One sentence

spoken by Romeo when he wants to reconcile with Tybalt, which has positive sentiment label, is data line 1574, stating 'Tybalt, the reason that I have to love thee.' Furthermore, it can be concluded that Romeo's shift from positive to neutral sentiment is not highly significant, indicating that Romeo's character remains stable throughout the story of Romeo and Juliet.

An intriguing aspect of this data analysis and visualization is the presence of low-level negative sentiment in both Romeo and Juliet's dialogues. This suggest that both characters, intentionally and unintentionally, occasionally express negative sentiments. This observation underscores that there are no characters who are entirely good or wholly evil; instead, it emphasizes the human tendency to express negative emotions, even in moments of vulnerability. Shakespeare's skillful character development demonstrates that negative sentiment doesn't equate to inherent evil; it rather reflects the emotional turmoil experienced by these characters during their challenging circumstances. His portrayal adeptly captures the complexity of human emotions, revealing the depth and intricacy of their personalities.

The result and the corresponding discussion demonstrate that there are numerous techniques for performing analysis in the field of literature. Analysis of character dynamics in drama can still be done with programming languages. This procedure confirms that a computational approach offers fresh and insightful perspectives on the dynamics, speech patterns, and characters in literary works. We can explore and understand the emotional nuances of characters by utilizing machine learning algorithms and sentiment analysis techniques, providing a more thorough understanding of the emotions aspects depicted in literary texts. This demonstrates how technological developments create new avenues for literary investigation and in-depth comprehension.

## CONCLUSION

This study embarked on exploring sentiment analysis within the realm of literature, particularly focusing on characters in "Romeo and Juliet" by Shakespeare. Sentiment analysis, a field that investigates emotions, opinion, and attitudes within textual data, provided a compelling lens through which to understand the emotional dynamics of literary characters. The sentiment analysis process involved several key steps, including data collection, text processing, sentiment analysis, machine learning using the Naïve Bayes Multinomial algorithm, and data visualization. The result

and discussions revealed interesting insight into the sentiments of the main characters, Romeo and Juliet.

Both characters exhibited a notable tendency toward neutral sentiments. Juliet's intelligence and her capacity to perceive situations from various perspectives contributed to her consistently neutral demeanor. However, her confusion in facing the Capulet and Montague feud also played a role in her neutral sentiment. Romeo's sentiment, on the other hand, showed a slight shift from positive to neutral during the play. His initial attempts to reconcile with Tybalt, even after Tybalt's provocations, demonstrated the character development. This shift was indicative of Romeo's commitment to peace and his refusal to perpetuate the family feud.

A fascinating revelation from the data analysis was the presence of low-level negative sentiment in both Romeo and Juliet's dialogues. This observation underlines the complexity of human emotions and the fact that characters in literature are not entirely good or evil. It emphasizes the natural human tendency to express negative emotions, even during vulnerable moments. Shakespeare's masterful character development captured these nuances, shedding light on the depth and intricacy of his characters' personalities.

This research contributes to the field of computational literary studies by using sentiment analysis to delve into the emotions of iconic literary characters. The study achieved an accuracy rate of 81% using the naïve bayes algorithm. These findings can provide valuable insights for future researchers interested in character sentiment analysis within novels. As a suggestion for future research, it is recommended to explore the use of the TF-IDF method in sentiment analysis, which may potentially enhance the model's accuracy. Additionally, further research is needed to identify the factors that can influence character sentiments in other classic literary works and to apply different sentiment analysis methods for a more in-depth comparison.

**REFERENCES**

Aahill, & eric-urban. (2023, July 20). Cara: Gunakan Analisis Sentimen dan Penggalian Opini. Retrieved from Microsoft Ignite: https://learn.microsoft.com/id-id/azure/ai-services/language-service/sentiment-opinion-mining/how-to/call-api

ADMINLP2M. (2022). Analisis Sentimen (Sentiment Analysis): Definisi, Tipe dan Cara Kerjanya.

Danukusumo, Kefin Pudi (2017) Implementasi Deep Learning Menggunakan Convolutional Neural Network Untuk Klasifikasi Citra Candi Berbasis Gpu. S1 thesis, UAJY.

Firdaus, A., & Firdaus, W. (2021). Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi: (Sebuah Ulasan). Jurnal JUPITER, 67.

Ganesan, K. (2023). What Are Stop Words? Retrieved from Opinosis Analytics: https://www.opinosis-analytics.com/knowledge-base/stop-words-explained/

Hota, H., Sharma, D., & Verma, N. (2021). Lexicon-based sentiment analysis using Twitter data. National Library of Medicine.

Mata, P., Mata, M., Martins, J., Rita, J., & Correia, A. (2021). Sentiment analysis - A literature review. Academy of Entrepreneurship Journal.

Maulud, D., Zeebaree, S., Jacksi, K., Sadeeq, M., & Sharif, K. (2021). State of Art for Semantic Analysis of Natural Language Processing. Qubahan Academic Journal, 21.

Pavan, L. (2022). A Survey of Some Italian Literature Works using Sentiment Analysis. International Journal of Linguistics, Literature, and Translnation, 117.

Priyono. (2016). Metode Penelitian Kuantitatif. Sidoarjo: Zifatama.

Rebora, S., & Gutenberg, J. (2023). Sentiment Analysis in Literary Studies. A Critical Survey. digital humanities quarterly.

Siregar, N., Siregar, R., & Sudirman, M. (2020). Implementasi Metode Naive Bayes Classifier (NBC) Pada Komentar Warga Sekolah Mengenai Pelaksanaan Pembelajaran Jarak Jauh (PBB). Jurnal Teknologia.

Sugiyono. (2013). Metodelogi Penelitian Kuantitatif, Kualitatif Dan R&D. Bandung: ALFABETA.

Therdoost, H., & Madanchian, M. (2023). Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research. MDPI.

Tineges, R., & Davita, A. (2022, May 23). Mengenal Naive Bayes Sebagai Salah Satu Algoritma Data Science. Retrieved from DQLab: https://dqlab.id/mengenal-naive-bayes-sebagai-salah-satu-algoritma-data-science

Warnes, Z. (2021, July 10). Missing Value Handling - Missing Data Types. Retrieved from medium: https://towardsdatascience.com/missing-value-handling-missing-data-types-a89c0d81a5bb.