Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

Naive Bayes Algorithm for Sentiment Analysis on Spider-Man Movie: No Way Home

Ziddan Makarim[1)*], Ismasari Nawangsih[2)], Sanudin[3)]
[1*,2,3)]Informatic Engineering, Universitas Pelita Bangsa, Bekasi, Indonesia
[1)]zddnmkrm05@mhs.pelitabangsa.ac.id, [2)]ismasari.n@pelitabangsa.ac.id, [3)]sanudin@pelitabangsa.ac.id

ABSTRACT

The rapid development of streaming platforms has significantly changed the landscape of movie consumption. The ease of access and social interaction in online communities has led to the creation of a new pop culture around movies. One interesting phenomenon is the movie Spider-Man: No Way Home, which sparked heated and viral conversations on various social media platforms. This research aims to analyze audience sentiment towards the movie Spider-Man: No Way Home using Naïve Bayes algorithm. Review data collected from online platforms was processed to identify positive and negative sentiments. The choice of Naïve Bayes algorithm is based on its efficiency and ability to classify text. The results showed that the model built was able to classify sentiment with an accuracy of 72.34%. The model is more effective in identifying positive reviews than negative, indicating a positive response from the majority of viewers. However, the model still needs to improve its performance in classifying negative sentiments. This research makes an important contribution in understanding audience preferences and evaluating the success of a movie, especially in the context of the digital era. The results can be utilized by the film industry to improve production quality, marketing strategies, and content development that is more relevant to audience preferences. In addition, this research also opens up opportunities for further development, such as the use of more complex algorithms or combining with other sentiment analysis techniques, as well as application to various types of social media content.

Keywords: Movie Industry; Naive Bayes Algorithm; Sentiment Analysis; Spider-Man: No Way Home; Streaming Platform;

INTRODUCTION

The rapid development of information and communication technology has significantly changed the landscape of the entertainment industry, especially in terms of movie consumption. The emergence of streaming platforms and easy access to the internet has encouraged the formation of online communities that actively discuss and share opinions about movies. This phenomenon is further reinforced by the huge success of movies like Spider-Man: No Way Home that triggered heated and viral conversations on various social media platforms. This study aims to analyze audience sentiment towards the movie Spider-Man: No Way Home in more depth using the Naïve Bayes algorithm. By analyzing a large number of reviews available on online platforms, this study aims to fill the existing research gap related to sentiment analysis on popular movies, especially in the context of the pop culture phenomenon triggered by superhero movies. The main difference of this research is the focus on the movie Spider-Man: No Way Home as a case study, as well as the use of Naïve Bayes algorithm to automatically classify sentiment. The results are expected to make a significant contribution to the movie industry in understanding audience preferences and evaluating the success of a movie in the digital era. In addition, this research can also serve as a foundation for further research on sentiment analysis on various types of social media content.

In recent years, the movie industry has undergone a significant transformation due to technological disruption. Streaming platforms such as Netflix, Disney+ Hotstar, and Amazon Prime Video have changed the way audiences access and consume movies. This ease of access has led to the formation of highly active online communities, where viewers can share their opinions, leave reviews, and participate in discussions about their favorite films. The intense social interactions within these online communities have created a new pop culture around movies, where audiences are not only passive consumers, but also content producers who actively shape the narrative and meaning behind a movie.

One interesting phenomenon in the world of cinema is the huge success of the movie Spider-Man: No Way Home. The movie not only broke box office revenue records, but also sparked widespread conversations on various social media platforms. The popularity of this movie shows the power of nostalgia and fan service in attracting audiences.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

However, behind its commercial success, there is an interesting research potential to dig deeper into the audience's perception of the movie. How do audiences actually respond to the plot, characters, and visual elements in the movie? Questions like these are the main motivation for this research.

Sentiment analysis is one method that can be used to measure public opinion on a topic. By applying sentiment analysis techniques to audience reviews of the movie Spider-Man: No Way Home, we can gain a deeper understanding of the audience's responses to various aspects of the movie. The Naïve Bayes algorithm was chosen as the analysis method due to its ease of implementation and ability to classify text quickly and efficiently. Although it has some limitations, it has proven effective in various natural language processing applications.

Research on the movie "Ngeri-Ngeri Sedap" shows that the majority of viewers have a neutral view. The Naïve Bayes algorithm successfully classified the sentiment with good accuracy. However, why many viewers have a neutral view still needs further research(Nurtikasari et al., 2022). Sentiment analysis of the movie "Oppenheimer" showed high accuracy in classifying positive and negative sentiments. Problems arose with reviews with positive sentiments but low ratings, as well as the use of non-standard language. This indicates the need to develop a more robust model(Anuar et al., 2023). The movie "Agak Laen" was successfully analyzed for sentiment with high accuracy using Naïve Bayes. Future research can focus on applying this model to other movie genres, different social media platforms, or comparisons with other algorithms(Cahya Kamilla et al., 2024).

This research is expected to make a significant contribution to several parties. First, for the movie industry, the results of this research can be used to better understand audience preferences, which can help in product development and more effective marketing strategies. Second, for researchers, this research can serve as a foundation for further research on sentiment analysis on various types of social media content. Third, for the general public, this research can provide a deeper insight into the pop culture phenomenon surrounding movies and how technology can be used to analyze public opinion.

Overall, this research aims to fill the existing research gap related to sentiment analysis on popular movies, particularly in the context of the pop culture phenomenon triggered by superhero movies. By using the Naïve Bayes algorithm, this research is expected to make a valuable contribution to the development of science and the creative industry.

LITERATURE REVIEW

Definition of Data Mining

Data mining is a process that involves finding hidden patterns, relationships, and valuable information from large and complex data sets. This process allows us to dig deeper into the data and discover insights that cannot be obtained by just looking at the data in its raw form (Komariyah et al., 2020). The main goal of data mining is to transform data into useful information for decision-making.

The techniques and methods used in data mining are very diverse, ranging from statistical methods, artificial intelligence, to machine learning. The selection of the right method depends largely on the type of data, the purpose of the analysis, and the resources available.

Definition of Text Mining

Text mining is a process to extract patterns from an unstructured document, and then get an interesting pattern that can be used (Mulyani & Novita, 2022). In other words, text mining is a data analysis technique that aims to find hidden information, patterns and relationships in large amounts of text. This unstructured text can be documents, articles, reviews, social media posts, or other text data sources. Several techniques in text mining are used to handle problems such as clustering, classification, information extraction, and also information retrieval (Darwis et al., 2020). These techniques allow researchers to group documents that have similar topics (clustering), classify documents into certain categories (classification). Extracting specific information from text (information extraction), and finding documents relevant to a query (information retrieval).

Definition of Sentiment Analysis

Sentiment analysis is a computational approach that automatically identifies emotional polarity contained in text, such as tweets on Twitter (Safira et al., 2023). In another journal, it is explained that sentiment analysis is one of the branches of text mining research, where sentiment analysis is related to a broader field such as data processing in certain activities (Darwis et al., 2021). In other words, sentiment analysis seeks to understand whether a text contains positive, negative emotions. In the context of this research, sentiment analysis is applied to Twitter tweets about the movie Spider-Man: No Way Home to measure public response to the movie.

The main purpose of sentiment analysis is to gain a deeper understanding of public perception on a topic. Through

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

sentiment analysis, researchers can identify opinion trends, measure the level of public satisfaction or dissatisfaction. As well as identify factors that influence the sentiment. The results of sentiment analysis can be utilized by various parties, such as companies, governments, and researchers to make better decisions based on objective data.

Definition of Classification

Classification is the process or stage of grouping new data or objects into labels or classes based on certain attributes (Noviansyah et al., 2024). In other words, classification is an attempt to predict the class or category of new data based on patterns that have been learned from existing data. In classification, the technique used is to look at variables from previously existing data groups, then build a model that can map new data into the appropriate class.

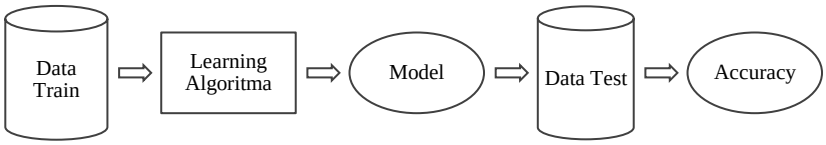

Fig. 1 Machine Learning Model Classification Process

The general process in classification to build and evaluate a machine learning model is as follows:

a. Data Train, the process begins by collecting and preparing data that will be used to train the model. This data is referred to as training data. The training data contains examples that have known labels or classes. For example, in the case of email classification, the training data will contain emails that have been labeled "spam" or not "spam".
b. Learning Algorithm, once the training data is ready the learning algorithm will be used to build the model. This algorithm will learn the patterns in the training data so that it can make predictions for new data. In the case of Naive Bayes, this algorithm will calculate the probability of occurrence of a class based on the features in the data.
c. Model. The result of the learning process is a model. This model represents the knowledge that the algorithm has learned from the training data. This model can then be used to make predictions on data that has never been seen before.
d. Test data. Once built, the model will be tested on different data, namely test data. The test data contains examples that have never been used in the training process. The goal is to measure how well the model can predict the class of the new data.
e. Accuracy, the result of evaluation on test data is accuracy. Accuracy is a metric used to measure model performance. The higher the accuracy value, the better the model can make predictions.

Definition of Naive Bayes Algorithm

Naïve Bayes is one of the most popular algorithms used for data mining purposes. This algorithm is widely used because of its ease of use and fast processing, easy to implement with a structure that is considered quite simple, also with a high level of effectiveness (Khotimah & Utami, 2022). These advantages make Naïve Bayes one of the main choices in data mining applications, especially for text classification problems, spam filters, and recommendation systems. The Naïve Bayes algorithm can also be interpreted as a method that has no rules. This algorithm uses a branch of mathematics called probability to find the greatest chance of all possible classifications by looking at the frequency of each classification in data mining (Retnosari et al., 2021).

The working principle of Naïve Bayes is based on Bayes' theorem, which allows researchers to calculate the probability of a class (for example, spam or non-spam) based on the features in the data. The main assumption in Naïve Bayes is that the features are mutually independent, although this assumption is often not fully met in real data. Nonetheless. Naïve Bayes still gives pretty good results in most cases.

Naive Bayes classification algortima has an equation that can be used as a reference that can be used to calculate the value of probability in decision making, as for the equation is (Anwar, 2022):

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

Description:

X : Data with unknown class
H : Hypothesized data X is a specific class

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

P(H|X) : Probability of hypothesis H based on condition X
P(H) : Probability of hypothesis H
P(X|H) : Probability of X based on condition in hypothesis H
P(X) : Probability of X

The stages in the Naive Bayes algorithm are in the following figure:

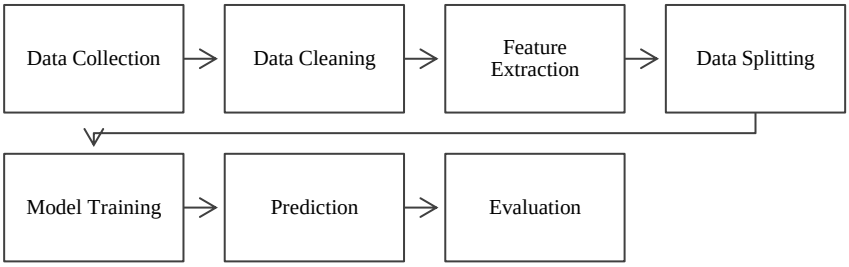

Fig. 2 Steps in the Naïve Bayes Algorithm

a. Data collection. the process begins with collecting data that is relevant to the problem to be solved. This data can be text, numbers, or a combination of both. Good data quality and quantity greatly affect the performance of the model.
b. Data cleaning, once collected, the data needs to be cleaned to remove noise, inconsistent data, or missing data. This process is important to ensure that the model is trained with accurate and quality data.
c. Feature extraction. Not all features in the data have the same contribution to the prediction results. Therefore, feature selection is performed to select the features that are most relevant to the class to be predicted. This process helps to reduce the dimensionality of the data and improve the efficiency of the model.
d. Data splitting The data that has been cleaned and feature selected is then divided into two parts, namely training data and test data. The training data is used to train the model, while the test data is used to evaluate the performance of the model.
e. Model training The Naïve Bayes model is trained using the training data. The model will learn the relationship between the features and the class to be predicted.
f. Prediction, once the model is trained, it can be used to predict the class of new data that has never been seen before.
g. Evaluation, the performance of the model is evaluated by comparing the predicted results with the actual values on the test data. Some commonly used metrics for evaluation are accuracy, precision, recall, and f1-score.

Spider-Man: No Way Home

Fig. 3 Spider-Man: No Way Home

Spider-Man: No Way Home, is one of the most anticipated and commercially successful superhero films in recent years. The film brought together three generations of Spider-Man played by Tobey Maguire, Andrew Garfield, and Tom Holland, and presented multiverses that drew fans' attention to the interactions between iconic characters from various Marvel universes. The popularity of Spider-Man: No Way Home is not only evident from the number of viewers in theaters, but also from the very active conversations on social media. Fans actively shared their opinions and feelings about the movie through various platforms such as Twitter, Instagram, and discussion forums. This results

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

in a huge volume of data that can be used for sentiment analysis.

Sentiment analysis of the movie Spider-Man: No Way Home has several potential benefits. First, it can provide a clearer picture of the public's perception of the movie. Second, it can identify the aspects of the movie that are most liked or disliked by the audience. Third, it can provide input for the film industry in developing sequels or other superhero films.

In addition, the movie Spider-Man: No Way Home also has several characteristics that make it an interesting object of research:

1. Fan-focused, this movie is highly anticipated by loyal Spider-Man fans, resulting in very diverse and intense reactions.
2. Nostalgic elements, the presence of three generations of Spider-Man evokes nostalgia in fans who grew up with this character.
3. The concept of multiverses, the complex concept of multiverses provides a lot of room for interpretation and discussion.

METHOD

This study delves into audience sentiment towards "Spider-Man: No Way Home" using sentiment analysis. Leveraging a rich dataset of 1,551 reviews from Kaggle.com, the research employs the Naïve Bayes algorithm to classify sentiment within the reviews. The pre-processing stage cleans and extracts features like keywords for effective model training. Evaluation metrics like accuracy and F1-score assess the model's performance, while further analysis explores the words and phrases most common in positive and negative reviews. This research not only contributes to sentiment analysis but also offers insights into movie audience behavior, potentially benefiting the film industry and future research. The following are the steps carried out in this research.

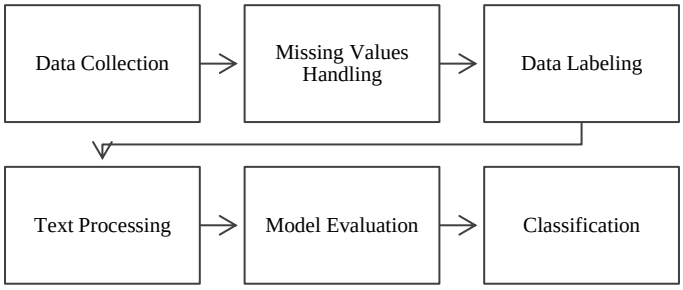

Fig. 4 Research Steps

1. The initial data used in this study amounted to 1,800 reviews on the movie Spider-Man: No Way Home obtained from the dataset on the Kaggle.com website.
2. Next, researchers handle missing values in the initial data used by deleting data that does not have values in it. By removing data with missing values, researchers reduced the sample size. However, this step is considered necessary to maintain data quality and avoid bias in further analysis.
3. Next, we labeled the data. To convert the rating scale into sentiment (positive, negative), the researcher set a threshold value of 5. The selection of this threshold is based on the assumption that ratings above 5 reflect high satisfaction with the film, so it is categorized as a positive sentiment. Conversely, a rating below 6 indicates dissatisfaction or neutralism, thus categorized as negative sentiment. Thus, the 'rating' column, which was originally numerical, is converted into a 'sentiment' column which is categorical.
4. Preprocessing is the first step in text processing that aims to clean and simplify the text so that it is better prepared for further analysis (Najjichah et al., 2019). This process is crucial, especially in sentiment analysis, where researchers want to identify positive or negative sentiments in a text. Raw text often contains noise such as emoticons, abbreviations, or spelling mistakes that can confuse algorithms. Through preprocessing, researchers can normalize the text, remove stopwords, and perform stemming or lemmitization to reduce words to their base form. For example, the word “interested” can be converted to its base form “interest”. Thus, sentiment analysis models can more easily identify key words that indicate positive or negative sentiment. In addition, preprocessing also helps in overcoming the class imbalance problem, where the number of positive and negative data is not balanced, so that the model can be trained more accurately.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

a. Data cleaning, is the process of removing inaccurate, inconsistent, or irrelevant data from the dataset. The goal is to produce high-quality, clean data that is reliable for further analysis(Watratan et al., 2020). This process is crucial because dirty data can produce biased and misleading analysis results. Some types of data that need to be cleaned include missing values, outliers, and data that is inconsistent with the predetermined format. In addition, data cleaning also involves handling irrelevant data, such as variables that do not contribute significantly to the analysis.
b. Tokenizing, is the process of decomposing sentences into individual words as the smallest unit in text analysis, allowing each word to be processed separately (Darmawan et al., 2023). This process is a fundamental process in text preprocessing because it allows computers to understand and process text more deeply. By dividing the text into individual tokens, researchers can perform various advanced analyses such as calculating word frequency, performing stemming or lemmitization, and building language models.
c. Filtering. After word parsing, the next stage is to filter out words that are considered important for further analysis, so that only relevant words are retained (Yuyun et al., 2021). Words that are often filtered out by stopwords are common words such as "or", "at", "and" that often appear in everyday language but do not carry deep meaning in the context of analysis. In addition to stopwords, special characters, numbers, and words that are too common or appear too infrequently are also often filtered out. By filtering, researchers can reduce the dimensionality of the data and improve computational efficiency in the subsequent analysis process.
d. Stemming, which is the process of reducing words into their basic form by removing unnecessary affixes (Nurtikasari et al., 2022; Sundara & Ekaputri Arnas, 2020). It is an important step in preprocessing. The main purpose of stemming is to combine words that have the same root but have different morphological forms. For example, the words "walking" and "walked" will be reduced to their base form "walk". Thus, stemming can increase the effectiveness in information retrieval and text analysis.
e. Transform cases, is the process of adjusting the word form by changing all characters to lowercase. This process aims to standardize the word form so that it facilitates the analysis process (Esa Tiffani, 2020). By equalizing all words into lowercase letters, we can reduce the variation of the same word form so that text processing algorithms can work more efficiently. For example, the words "Tom" and "tom" will be considered as the same word after transform cases.
f. Stopword removal, is a step in text processing that aims to filter out connecting words such as "a", "the". These words, although frequently used, are considered to carry no meaningful information in text analysis (Prasetyo et al., 2023). These words, often called stopwords, are very commonly used in everyday language but do not contribute significantly to identifying patterns or themes in a text. By removing stopwords, researchers can reduce the dimension of the vector space and improve the efficiency of machine learning algorithms. Thus, stopword removal helps researchers to focus on words that actually carry meaning in the text.

5. After the review data has been cleaned and the features extracted, the next step is to evaluate the performance of the model that has been built. This is where cross validation plays an important role. Cross validation is a technique in machine learning that is used to evaluate the performance of a model more objectively and avoid overfitting. Overfitting occurs when the model is too complex and too adapted to the training data. Therefore, it performs poorly when applied to new data.
6. Once the model has been trained and tested using cross validation techniques, researchers need to have a way to measure how well the model performs. This is where the confusion matrix comes in. Confusion matrix is a table that presents the predicted results of the built model compared to the actual class labels. In the context of sentiment analysis, the confusion matrix can be used to measure the model's performance in classifying a thread as positive or negative. For example, if a researcher wants to know how often the model misclassifies positive reviews as negative, the researcher can look at the false negative value.

RESULT

Dataset Used

After a data cleaning process by eliminating empty or irrelevant data (missing values), the amount of data that can be used in further analysis becomes 1,551 reviews. This reduction in the amount of data aims to ensure the quality and reliability of the analysis to be carried out, using only relevant complete data.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

Table 1Dataset Used

No	text	rating
1	Awful movie. The plot sucked and had awful pacing humour, and dialogue were subpar(like the octopus…	5
2	It's hard to discuss this movie and not get into the spoilers because the biggest moments of this movie…	10
3	Review By Kamal KI don't think this film was written well. It had already established Spider-man charac…	8
…	…	…
…	…	…
1550	The amount of feelings and love for this movie and the performance was great from the crew of this movie…	?
1551	This movie feels suprisingly somber for a marvel movie. Not necessarily in the things that happen but…	9

Remove Missing Values

The review data about the movie Spider-Man: No Way Home has gone through text cleaning and processing using RapidMiner. After handling missing values, then handling the characters and symbols contained in the data, after that all the steps contained in the preprocessing activities are also carried out, so that the sentiment analysis model can be more accurate in identifying the sentiment of the movie audience. The process is depicted in the figure below.

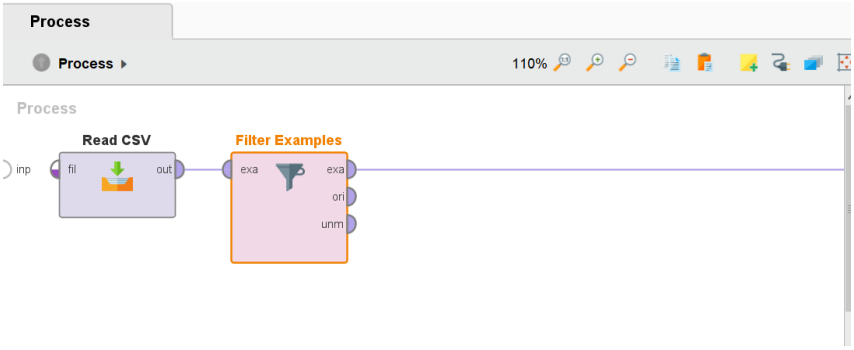

Fig. 5 Missing Values Removing

Table 2
No Missing Values

No	text	rating
1	Awful movie. The plot sucked and had awful pacing humour, and dialogue were subpar(like the octopus…	5
2	It's hard to discuss this movie and not get into the spoilers because the biggest moments of this movie…	10
3	Review By Kamal KI don't think this film was written well. It had already established Spider-man charac…	4
…	…	…
…	…	…
1550	The originals have always been my fav Spiderman movies, then I saw this movie! All around a great m…	9
1551	This movie feels suprisingly somber for a marvel movie. Not necessarily in the things that happen but…	9

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

Labeling

As described in the method, after the activity of removing missing values, in this study, the data labeling process is carried out. Process can be seen in the figure below.

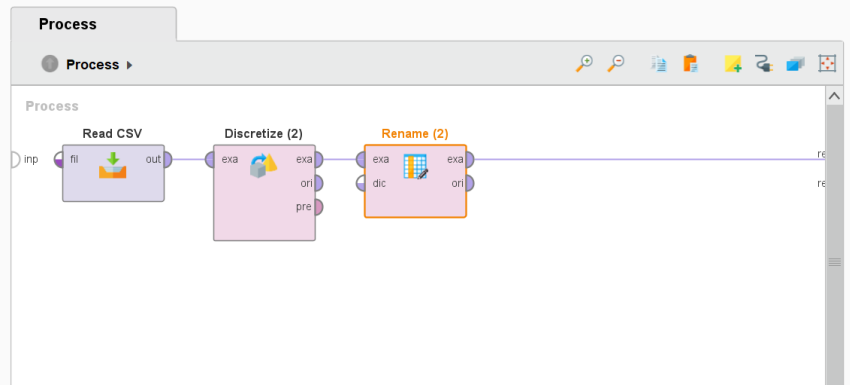

Fig. 6 Labeling Process

Table 3
Data Labeling Result

No	text	rating
1	Awful movie. The plot sucked and had awful pacing humour, and dialogue were subpar(like the octopus…	negative
2	It's hard to discuss this movie and not get into the spoilers because the biggest moments of this movie…	positive
3	Review By Kamal KI don't think this film was written well. It had already established Spider-man charac…	negative
…	…	…
…	…	…
5326	The originals have always been my fav Spiderman movies, then I saw this movie! All around a great m…	positive
5327	This movie feels suprisingly somber for a marvel movie. Not necessarily in the things that happen but…	positive

Preprocessing and Feature Extraction

After the sentiment labeling process, the review data goes through another round of preprocessing to ensure the quality of the data to be used in the sentiment analysis model. This process includes the removal of irrelevant characters and symbols, such as emoticons, hashtags, and other special characters. In addition, trimming is also performed to remove excess spaces at the beginning or end of words. These steps aim to produce cleaner and more consistent data, so that the model can more accurately identify the sentiment of a review.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

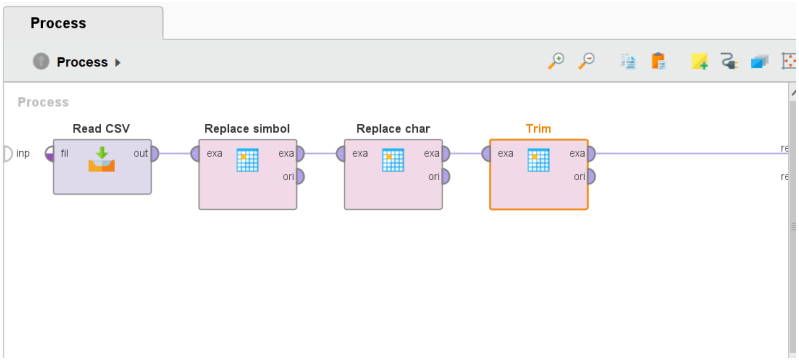

Fig. 7 Preprocessing Data

Once the review data has been cleaned of characters and symbols, the next stage is feature extraction. This process starts with tokenizing to break the text into individual words. Then, the words are converted into their basic form through the stemming process. Next, stopwords that are common and do not carry much semantic information, such as “and”, “or”, “with”, are removed. To simplify the analysis, all words are converted into lowercase letters through the transform cases process. Finally, tokens that are too short or too long are removed through token by length filtering to obtain more relevant features.

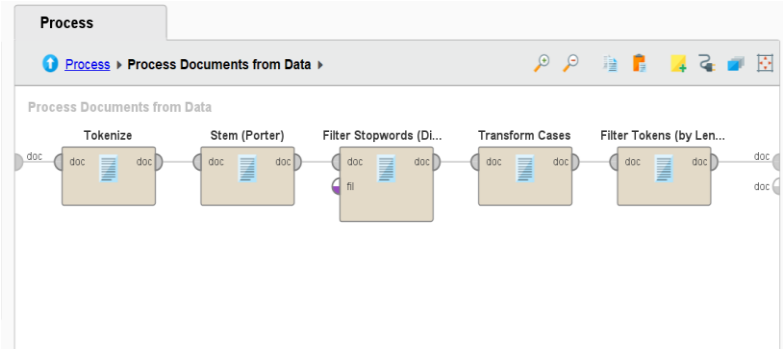

Fig. 8 Feature Extraction

Table 4Dataset After Preprocessing and Feature Extraction

No	text	rating
1	movi plot suck pace humour…	negative
2	hard discuss thi movi spoiler becaus biggest moment thi movi necessit ruin sever reveal love everyth trail…	positive
3	review kamal thi film written alreadi establish spiderman charact involv stark…	negative
…	…	…
…	…	…
1550	origin alwai fav spiderman movi…	positive
1551	thi movi feel suprisingli somber marvel movi necessarili happen wai thei minut mcu movi leav feel…	positive

Cross Validation

After removing missing values, labeling, feature extraction and preprocessing, then cross validation is performed, here are the steps.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

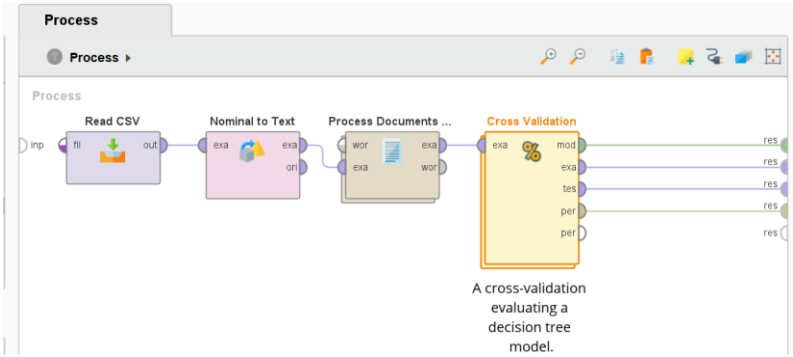

Fig. 9 Cross Validation

Classification with Confussion Matrix

After cross validation, classification with confusion matrix is performed to obtain model performance values in the form of accuracy, precision, and recall. Here is the application.

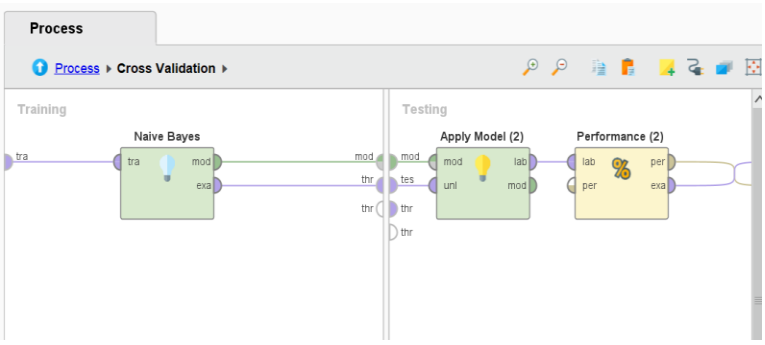

Fig. 10 Classification with Confussion Matrix

Based on the confusion matrix presented, it can be seen that the model built has achieved an accuracy of 72.34%. This means that the model managed to correctly predict the sentiment of 72.43% of all test data. However, if you look in more detail at the precision and recall values, there are some interesting findings.

The precision value for the positive class (92.18%) is much higher than the negative class (10.85%). This indicates that when the model predicts a review as positive, the prediction tends to be correct. However, when the model predicts a review as negative, the accuracy is lower, indicating that the model may be better at identifying positive reviews compared to negative reviews.

The recall value for the positive class (76.23%) was also quite good, meaning that the model was able to capture most of the actual positive reviews. However, the recall for the negative class is very low (30.83%). This means that the model often fails to identify actual negative reviews. The results can be seen in the figure below.

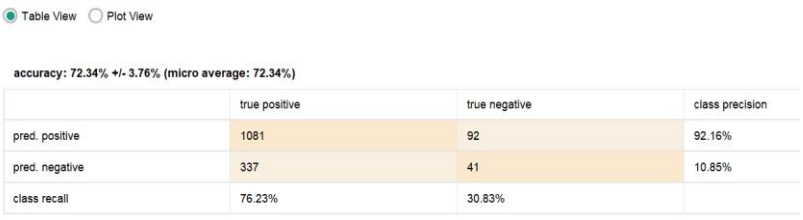

Table View Plot View

accuracy: 72.34% +/- 3.76% (micro average: 72.34%)

	true positive	true negative	class precision
pred. positive	1081	92	92.16%
pred. negative	337	41	10.85%
class recall	76.23%	30.83%	

Fig. 11 Classification Accuracy Result

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

DISCUSSIONS

Model Performance Evaluation

The sentiment classification model developed in this study shows a fairly good performance in predicting the sentiment of Spider-Man movie reviews: No Way Home. with an accuracy rate of 72.34%. This figure indicates that the model is able to predict review sentiment with a fairly good level of success. However, further analysis of the other evaluation metrics, namely precision and recall, provides a more detailed picture of the strengths and weaknesses of the model.

Precision and Recall Evaluation

The precision value for the positive class of 92.18% indicates that the model predicts a review as positive, a prediction that has a high level of confidence. This means that the model tends to be correct in identifying reviews that do contain positive sentiment. In contrast, the relatively low precision value for the negative class (10.85%) indicates that the model often misclassifies positive reviews as negative.

The recall value for the positive class of 76.23% indicates that the model is able to identify most of the positive reviews in the dataset. However, the low recall value for the negative class (30.83%) indicates that the model struggles to identify negative reviews. This could be due to several factors, such as class imbalance in the dataset or the complexity of the features contained in the negative reviews.

Implications and Suggestions for Improvement

The evaluation results show that the developed model has the potential to be used in classifying movie review sentiment. However, there are still some areas that need improvement. To solve the class imbalance problem, oversampling or undersampling techniques can be applied to the training data. In addition, the use of more complex algorithms such as Support Vector Machine (SVM) or deep learning can be considered to improve the accuracy of the model in classifying minority classes.

CONCLUSION

Based on the results of sentiment analysis of Spider-Man movie reviews: No Way Home using the Naïve Bayes algorithm, it can be concluded that the majority of viewers gave a positive response to the movie. The model built is able to classify reviews with fairly good accuracy, especially for the positive class. This indicates that elements such as the multiverse concept, nostalgia, and spectacular visual effects succeeded in attracting viewers and providing a satisfying viewing experience. However, the model still needs to improve its performance in identifying negative reviews, which suggests there is room for future development of more complex models.

This research contributes to the understanding of superhero movie audience preferences and the influence of pop culture phenomena on audience perceptions. The results of this study can be utilized by the movie industry to develop more effective marketing strategies, as well as for researchers to conduct further studies on sentiment analysis on various types of social media content. In addition, this research also shows the potential of using the Naïve Bayes algorithm in analyzing sentiment on natural language text.

As a limitation of this research, it can be mentioned that the model built still has some shortcomings, such as difficulty in identifying more complex sentiments or sarcasm. In addition, the dataset used in this research comes from one source, namely Kaggle.com, so further research needs to be done using more diverse datasets to generalize the research results.

The research question regarding audience sentiment towards the movie Spider-Man: No Way Home has been partially answered. This study identified that the majority of viewers responded positively to the movie. However, to gain a deeper understanding of the factors that influence audience sentiment, further research is needed using more sophisticated analysis methods and larger datasets.

REFERENCES

Anuar, F., Putra, R., Firman Fadilah, F., Enri, U., & Karawang, U. S. (2023). *ANALISIS SENTIMEN ULASAN FILM OPPENHEIMER PADA SITUS IMDB MENGGUNAKAN METODE NAIVE BAYES*.

Anwar, K. (2022). *KLIK: Kajian Ilmiah Informatika dan Komputer Analisa sentimen Pengguna Instagram Di Indonesia Pada Review Smartphone Menggunakan Naive Bayes*. *2*(4), 148–155. https://djournals.com/klik

Cahya Kamilla, A., Priyani, N., Priskila, R., Handrianus Pranatawijaya, V., Yos Sudarso, J., Jekan Raya, K., Palangka Raya, K., & Tengah, K. (2024). ANALISIS SENTIMEN FILM AGAK LAEN DENGAN KECERDASAN BUATAN: TEXT MINING METODE NAÏVE BAYES CLASSIFIER. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 8, Issue 3).

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 4, October 2024
https://doi.org/10.47709/cnahpc.v6i4.4845

Submitted : Oct 10, 2024
Accepted : Oct 20, 2024
Published : Oct 25, 2024

Darmawan, G., Alam, S., Imam Sulistyo, M., Studi Teknik Informatika, P., Tinggi Teknologi Wastukancana Purwakarta, S., & Artikel, R. (2023). *ANALISIS SENTIMEN BERDASARKAN ULASAN PENGGUNA APLIKASI MYPERTAMINA PADA GOOGLE PLAYSTORE MENGGUNAKAN METODE NAÏVE BAYES INFO ARTIKEL ABSTRAK*. *2*(3), 100–108. https://doi.org/10.55123

Darwis, D., Shintya Pratiwi, E., Ferico, A., & Pasaribu, O. (2020). PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA. In *Jurnal Ilmiah Edutic* (Vol. 7, Issue 1).

Darwis, D., Siskawati, N., & Abidin, Z. (2021). *Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional*. *15*(1).

Esa Tiffani, I. (2020). *Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review*.

Felicia Watratan, A., Puspita, A. B., Moeis, D., Informasi, S., & Profesional Makassar, S. (2020). Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. In *JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST)* (Vol. 1, Issue 1). http://journal.isas.or.id/index.php/JACOST

Khotimah, A. C., & Utami, E. (2022). COMPARISON NAÏVE BAYES CLASSIFIER, K-NEAREST NEIGHBOR AND SUPPORT VECTOR MACHINE IN THE CLASSIFICATION OF INDIVIDUAL ON TWITTER ACCOUNT. *Jurnal Teknik Informatika (JUTIF)*, *3*(3). https://doi.org/10.20884/1.jutif.2022.3.3.254

Kokom Komariyah, Rahaditya Dasuki, Dias Bayu Saputra, Saeful Anwar, & Gifthera Dwilestari. (2020). Klasifikasi Stok Barang Menggunakan Algoritma Naïve Bayes Pada Pt.Dharma Electrindo Manufacturing. *KOPERTIP : Jurnal Ilmiah Manajemen Informatika Dan Komputer*, *4*(2), 35–41. https://doi.org/10.32485/kopertip.v4i2.117

Mulyani, S., & Novita, R. (2022). IMPLEMENTATION OF THE NAIVE BAYES CLASSIFIER ALGORITHM FOR CLASSIFICATION OF COMMUNITY SENTIMENT ABOUT DEPRESSION ON YOUTUBE. *Jurnal Teknik Informatika (Jutif)*, *3*(5), 1355–1361. https://doi.org/10.20884/1.jutif.2022.3.5.374

Najjichah, H., Syukur, A., & Subagyo, H. (2019). PENGARUH TEXT PREPROCESSING DAN KOMBINASINYA PADA PERINGKAS DOKUMEN OTOMATIS TEKS BERBAHASA INDONESIA. In *Jurnal Teknologi Informasi* (Vol. 15, Issue 1). http://research.

Noviansyah, B., Makmun Effendi, M., & Achmad, Y. (2024). Sentiment Analysis of Oppenheimer Movie Reviews: Naïve Bayes Algorithm for Public Opinion. *Architecture and High Performance Computing*, *6*(3). https://doi.org/10.47709/cnapc.v6i3.4393

Nurtikasari, Y., Syariful Alam, & Teguh Iman Hermanto. (2022). Analisis Sentimen Opini Masyarakat Terhadap Film Pada Platform Twitter Menggunakan Algoritma Naive Bayes. *INSOLOGI: Jurnal Sains Dan Teknologi*, *1*(4), 411–423. https://doi.org/10.55123/insologi.v1i4.770

Retnosari, R., Studi, P., Informasi, S., Nusa, S., & Jakarta, M. (2021). *ANALISIS KELAYAKAN KREDIT USAHA MIKRO BERJALAN PADA PERBANKAN DENGAN METODE NAIVE BAYES*.

Safira, A., Masyarakat…ν, A. S., & Hasan, F. N. (2023). ANALISIS SENTIMEN MASYARAKAT TERHADAP PAYLATER MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER. *Jurnal Sistem Informasi*, *5*(1).

Sundara, T. A., & Ekaputri Arnas, S. (2020). *Naïve Bayes Classifier untuk Analisis Sentimen Isu Radikalisme*.

Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, & Fitri Nurapriani. (2023). Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN. *Jurnal KomtekInfo*, 1–7. https://doi.org/10.35134/komtekinfo.v10i1.330

Yuyun, Nurul Hidayah, & Supriadi Sahibu. (2021). Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(4), 820–826. https://doi.org/10.29207/resti.v5i4.3146

* Corresponding author