Sentiment Analysis of Oppenheimer Movie Reviews: Naïve Bayes Algorithm for Public Opinion

Berliana Noviansyah1)*, Muhammad Makmun Effendi2), Yudianto Achmad3)
1)2)3) Informatics Engineering, Universitas Pelita Bangsa, Bekasi, Indonesia
1) noviansyahberliana@email.com, 2) effendiyan@pelitabangsa.ac.id, 3) yudianto.achmad@pelitabangsa.ac.id

ABSTRACT
The development of information and communication technology has revolutionized the way people consume and engage with media, particularly in the realm of film. Online platforms such as Netflix, Amazon Prime Video, and YouTube have transformed movie consumption habits, providing a vast array of options for viewers to explore and enjoy. A crucial aspect of this digital landscape is the proliferation of movie reviews, which serve as valuable guides for users seeking to discover films aligned with their preferences. However, the abundance of reviews, often varying in quality and objectivity, necessitates tools capable of effectively processing and understanding these textual data. This research delves into sentiment classification of Oppenheimer movie reviews, utilizing the Naïve Bayes algorithm to categorize reviews into positive, negative, and neutral sentiments. The dataset comprising audience reviews and numerical ratings undergoes preprocessing using the TF-IDF method to facilitate numerical representation. Subsequently, the Naïve Bayes algorithm is trained on this processed data to accurately classify sentiments. The model demonstrates exceptional performance, achieving an accuracy rate of 97.45% in distinguishing between positive, negative, and neutral sentiments within Oppenheimer movie reviews. This study underscores the efficacy of the Naïve Bayes algorithm in sentiment classification and emphasizes the significance of employing techniques like TF-IDF for enhancing sentiment analysis in the domain of movie reviews.

Keywords: Sentiment Classification; Oppenheimer; Naïve Bayes; TF-IDF; Accuracy Rate

1. INTRODUCTION
The development of information and communication technology is inseparable from website service providers, the ability of humans to develop and utilize technology makes it develop very rapidly and significantly. In its development, things like watching and rating movies are also affected. Nowadays, people can easily watch and review the movies they have watched.

Nowadays, watching movies is one of the activities that people can do in their leisure time. With more and more movies and reviews on online platforms available both at home and abroad, it makes it easier for viewers to watch and find movies they like from the reviews available which makes movie lovers love online platforms to watch movies. Online platforms such as Netflix, Amazon Prime Video, and Youtube have become popular alternatives for movie lovers to enjoy their favorite movies. These online platforms offer various conveniences and accessibility that allow users to watch movies anytime and anywhere.

Globally, the movie industry continues to increase every year. Data from the National Association of Theater Owners shows significant growth, from 1.09 billion tickets sold in 1987 to 1.314 billion tickets in 2016 in the United States and Canada alone. This increase in viewership demonstrates the high level of public interest in movies and the increasingly important role that reviews play in influencing viewing decisions (Nurtikasari et al., 2022).

A movie is a series of images of objects in motion, then produces a series of events. The function of film is as a medium of communication, entertainment, and education supported by other elements such as music and color so as to produce something that looks realistic so that the message contained in the film can be conveyed to the audience, besides that it can also influence audience behavior.

Oppenheimer is a film directed by Christopher Nolan starring top actors that became a hot topic of conversation on social media after its release in July 2023. The movie became the talk of the IMDB website as it tackled a complex and sensitive topic. Many viewers praised the movie for the strong acting of the cast, many also praised the realistic shooting of this historical event. On the other hand, some critics felt the film was too sympathetic to Oppenheimer and not critical enough of the United States’ role in the development of the atomic bomb. The debate surrounding

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).
Oppenheimer's film shows that it has succeeded in sparking thought and discussion about one of the most important events in human history. Reviews and opinions about the film have flooded social media, the film has received over 200,000 reviews on the IMDB website, which creates valuable textual data to analyze.

Online movie reviews play an important role in helping users choose movies that suit their tastes. Movie reviews can help provide information on various aspects of the movie such as the story, acting of the cast, cinematography, and visual effects. This information can greatly help users in deciding whether the movie suits their taste or not. Online movie reviews can help audiences, especially amidst the proliferation of movie choices, to get a clearer picture of the quality and relevance of a movie.

Given the rapid growth of the movie industry and the important role of reviews in audience decision-making, this research aims to develop a movie review sentiment classification model using Naïve Bayes classification on the dataset. This model is expected to help users get more accurate information about public opinion on movies. This information is certainly very valuable for various parties, including movie producers, distributors, and the audience itself.

2. LITERATURE REVIEW

2.1. Definition of Data Mining

Data mining is the process of analyzing data from different perspectives and then summarizing it into other important information that can be used to find correlations or patterns from hundreds or thousands of fields from a large relational database (Ratna Sari Hutasuhut, 2020).

In another sense, data mining is a term used to describe the discovery of knowledge in a database using statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information (Tejawati, Septiarini A, 2023).

2.2. Definition of Text Mining

Text Mining is a process to perform recognition on text and documents (Nurtikasari et al., 2022). Text mining is the process of extracting information and knowledge from unstructured text data. This process involves various techniques, such as natural language processing (NLP), computer programming, and statistics to identify patterns, trends, and meanings in large collections of text.

2.3. Definition of Classification

Classification is the stage of grouping new data or objects into labels or classes based on certain attributes. The technique of classification is to look at variables from previously existing data groups. The purpose of classification is to predict the class of an object that was previously unknown (Setio P, Saputro D, 2020).

As for others, the definition of classification is one of the techniques in data mining that classifies data into predetermined groups. Classification is a supervised learning method, where the method requires test data into predetermined groups or classes (Budi Trisno & Agung Raharja, 2023).

In the data classification process, there are two processes carried out on the data, namely:

1. Training process
   The training process is a process where the activity is that the data that has been divided is used as training data, where the data is trained with an algorithm to be able to produce a model, which will be used to test data on test data.

2. Testing process
   The testing process is the process of testing data using a model obtained from training data with algorithms on training data. This process uses the model that has been created to predict the labels.

2.4. Definition of Analysis

Analysis is a process carried out to obtain conclusions about an activity that is being or wants to be carried out. Or, analysis is an investigation and decomposition of a problem to be able to find out the true situation and also a process in problem solving that starts with a conjecture and its truth [10].

In another sense, analysis is to describe some of its main parts and examine the parts to get the overall meaning (Azzahra Nasution & Khotimah H, 2019).

2.5. Definition of Naïve Bayes

* Corresponding author
Naïve Bayes is one of the algorithms in the classification method, which utilizes probability and statistical methods. This algorithm utilizes the probability theorem, namely Bayes' theorem, by predicting future probabilities or possibilities based on previous information (Khotimah & Utami, 2022).

Another definition, Naïve Bayes is one of the popular algorithms used in data mining purposes because of its ease of use and fast processing, easy to implement with a fairly simple structure with a high level of effectiveness (Yunefri & Ersan Fadrial, 2021). The Naïve Bayes algorithm calculates the probability of entering a particular character sample into class h (posterior), namely the probability of the appearance of class x multiplied by the appearance of the sample character in class c (likelihood). The general form of the Naïve Bayes algorithm is shown in the equation:

\[
P(c|x) = \frac{P(x|c)P(c)}{P(x)}
\]

Description:
- \(P(c|x)\): Posterior probability of class (c, target) given predictor (x, attribute)
- \(P(c)\): Is the prior probability of the class.
- \(P(x|c)\): Is the likelihood which is the probability of a particular class predictor.
- \(P(x)\): Is the prior probability of the predictor.

2.6. Oppenheimer

Oppenheimer, directed by Christopher Nolan and released in 2023, tells the story of J. Robert Oppenheimer, an American theoretical physicist who led the Manhattan Project to develop the atomic bomb during World War II. The movie explores the moral dilemmas and dire consequences of developing nuclear weapons, as well as Oppenheimer's contribution in changing the course of history.

Oppenheimer's film has received widespread acclaim for its stunning cinematography, strong acting, and sensitive portrayal of complex historical events. Oppenheimer has also been nominated for numerous awards, including Best Picture at the 2024 Academy Awards.

However, despite the praise, the movie has also been criticized for some historical inconsistencies and controversial interpretations of Oppenheimer's characters. Despite its flaws, Oppenheimer remains a powerful and provocative movie that triggers reflections on the responsibility of scientists, the destructive potential of technology, and the ethics of war.

3. METHOD

This research focuses on sentiment classification of Oppenheimer movie reviews using Naïve Bayes algorithm. Qualitative data in the form of movie reviews and numerical ratings were collected from Kaggle.com, a trusted platform for public data. This secondary data was chosen because it allows in-depth exploration of audience reasoning, opinions, and experiences that cannot be captured with quantitative data.

The dataset consists of two attributes: a text attribute containing audience reviews and a rating attribute containing a numerical value of 1-10 reflecting their sentiments. To process the data, movie reviews will be cleaned of noise, converted into lowercase and base word forms, and removed from stop words. Next, the TF-IDF (Term Frequency-Inverse Document Frequency) method is applied to convert the reviews into numerical representations and the numerical ratings are converted into binary sentiment labels (positive or negative). The Naïve Bayes algorithm is then trained on the processed data and evaluated on a subset of the data not used for training.

The main objective of this research is to produce an accurate sentiment classification model to predict the sentiment of new movie reviews. The results of the sentiment analysis are expected to provide a comprehensive picture of public opinion towards Oppenheimer's film, allowing him to understand how the Naïve Bayes algorithm can be used in sentiment classification of movie reviews.

4. RESULT

This chapter will explain the steps of dataset processing carried out by researchers. Where the process starts from cutting the dataset, until the results are obtained.

4.1. Dataset

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
The Oppenheimer movie review data in the research made by the author comes from the kaggle.com website. Kaggle.com website provides datasets containing movie reviews from various sources, one of which is the IMDB reviews website. The amount of data used by the author in this study is 10,000 data cut from 84,000 data loaded from the IMDB Movie Reviews website. It is estimated that 10,000 data is enough to adequately represent the population of Oppenheimer movie reviews. Data sources for Oppenheimer movie reviews can be movie criticism sites, personal blogs, social media, or other online platforms.

<table>
<thead>
<tr>
<th>No</th>
<th>text</th>
<th>rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Definitely worth it!!!harshitguptaeleventh23 Jul…</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Stunning in every.aspectmrbubblesiyag21 July 20…</td>
<td>8</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>3227</td>
<td>I learned a lot about the basics if this eventarnaudibiliotti27 July 2023</td>
<td>9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>9999</td>
<td>Multi Oscar Pictureinfo-5806121 July 2023Amazing movie, the…</td>
<td>1</td>
</tr>
<tr>
<td>10000</td>
<td>Awesometiffveronica24 July 2023The movie is very well-wr…</td>
<td>9</td>
</tr>
</tbody>
</table>

Next, data processing is carried out from the previously obtained dataset, which starts from the labeling process until obtaining accuracy results. The following are the steps.

4.2. Labeling

After that, the researcher handled the labeling on the data, changing it by modifying the rating column into a column containing sentiment labels. The parameter used by researchers to determine the label is based on the scale in the rating column that is already available in the dataset. Here are the results of the data that has been labeled in the table 2.

<table>
<thead>
<tr>
<th>No</th>
<th>text</th>
<th>sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Definitely worth it!!!harshitguptaeleventh23 Jul…</td>
<td>negative</td>
</tr>
<tr>
<td>2</td>
<td>Stunning in every.aspectmrbubblesiyag21 July 20…</td>
<td>positive</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>3216</td>
<td>As a Nolan fan this was boringssolidsnake22 July 2…</td>
<td>negative</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10000</td>
<td>Awesometiffveronica24 July 2023The movie is very well-wr…</td>
<td>positive</td>
</tr>
</tbody>
</table>

4.3. Preprocessing

Preprocessing is a stage of changing the initial text or original text as input by applying some basic routines to change or remove textual elements that are not useful in further processing (Najjichah et al., 2019). The steps in preprocessing activities are as follows:

a. Cleaning, which removes characters in the text which can be URLs, hastags, mentions, and other symbols to produce original review data (Normawati & Prayogi, 2021)

* Corresponding author
b. Tokenizing, is a process used to cut data that was originally a sentence form into words to see the tokens or values of each word (Syahril Dwi Prasetyo et al., 2023).

c. Filtering is the process of filtering and removing irrelevant or useless data from the text to be able to parse noise and improve the quality of text data.

d. Stemming is the process of changing words from the filtering results of removing affixes on words in documents so as to produce the basic word (Normawati & Prayogi, 2021)

e. Transform Cases is a process to convert all capital letters into lowercase letters or vice versa. The goal is to maintain consistency in the use of letters to proceed to the next stage, thus avoiding potential errors in the tokenizing process (Rani Yunita, 2023).

f. Remove stopword is the process of removing words that do not have significant meaning (Zuhri et al., 2020). Examples are "this", "or", “are” etc.

![Fig. 1 Preprocessing Data](image)

**Table 3 Data Clean**

<table>
<thead>
<tr>
<th>No</th>
<th>text</th>
<th>sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>defenit worth juli liter amaz thi movi fantast direct cinemat sc…</td>
<td>negative</td>
</tr>
<tr>
<td>2</td>
<td>stun everi juli littl previou christoph nolan everi singl acto…</td>
<td>positive</td>
</tr>
<tr>
<td>3216</td>
<td>nolan fan thi juli bore movi mostli polit court scenesthei japan b…</td>
<td>negative</td>
</tr>
<tr>
<td>10000</td>
<td>juli movi veri wellwritten cast job veri understand movi hour excit…</td>
<td>positive</td>
</tr>
</tbody>
</table>

4.4. Feature Extraction

Feature extraction is a process used to obtain unique values. Used to recognize an object based on a special histogram that the object has. The goal is to perform calculations and comparisons that can be used to classify an image based on the characteristics of its histogram (Nurtikasari et al., 2022).

TF-IDF is a way of weighting text data that is used to improve classification accuracy in machine learning. This method works by assigning a weight value to each word in a document based on how often the word appears, and how rarely it appears in the entire document set.

4.5. Split Data

Data Splitting is the process of dividing data in a dataset into two different subsets, namely training data and testing data, which are used for training and testing machine learning models. The goal is to measure how strong the model's performance is on data that has never been seen during the training process (Putra et al., 2024).

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
Training data is a collection of data used to train machine learning models which contains data that has been labeled. The label indicates the category or value that the model wants to predict, then, the model will learn the pattern in the training data and then use the pattern to make predictions on new data. Meanwhile, test data is a collection of data that is used to evaluate the performance of the machine learning model that has been previously created. The model will predict the labels for the data in the test data, and the predictions will be compared with the actual labels to calculate evaluation metrics such as accuracy, precision, and recall.

Fig. 2 Data Splitting Process

The figure above shows a simple process in RapidMiner that involves three main operators, namely: read csv, split data, and store. Data is read from a csv file and then split into two parts and stored using two store operators, for training data and test data.

<table>
<thead>
<tr>
<th>No</th>
<th>text</th>
<th>sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>movi plai write thscodyfleuri juli result import …</td>
<td>neutral</td>
</tr>
<tr>
<td>2</td>
<td>dazzlingli direct biopic caution juli fine audienc mani understand the…</td>
<td>positive</td>
</tr>
<tr>
<td></td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>995</td>
<td>masterpiec understand forriveradisaac juli watch opp…</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>7999</td>
<td>tiw juli todai watch nolan oppenheim goodto start cillian…</td>
<td>neutral</td>
</tr>
</tbody>
</table>

The table contains the test data used to evaluate the model. Each row in the table represents one data point or instance. The “text” column contains the review text, and the “sentiment” column contains the corresponding sentiment label (positive, negative, neutral).

4.6. Cross Validation

Cross validation is a technique that can get a value or can validate the accuracy of the model based on a particular dataset used to perform prediction or classification activities (Homepage et al., 2022). Cross validation aims to estimate model performance more accurately and objectively, in this study researchers divided the dataset into 5 equal subsets (fold). Cross validation is used to evaluate the performance of the movie sentiment classification model.

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
The evaluation results show that the model achieved an average accuracy of 96.34%, an average precision of 96.39%, and an average recall of 96.33% as shown in Figure below.

4.7. Classification

Next, the researcher performs the data classification process. The Naïve Bayes model is trained using training data where, at this stage, the model learns the relationship between the features in the data and their corresponding labels. The Naïve Bayes algorithm assumes that the occurrence of each feature is independent, and calculates the probability of each class based on the observed features.

Then, the trained Naïve Bayes model is used to classify the test data. At this stage, the model predicts the label for each test data based on the observed features. Label prediction is done by selecting the class that has the highest probability based on the Naïve Bayes calculation. The process is described in the figure below.

Based on the test results of the calcification, it shows that the model trained on the training data is able to classify the test data with an accuracy rate of 97.45%, precision rate of 97.46% and recall rate of 97.45%. These results show that the Naïve Bayes model is able to identify positive, negative and neutral sentiments in Oppenheimer movie reviews.

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
with a high degree of accuracy. As for the accuracy results of the above classification process can be seen in the figure below.

![Classification Accuracy Result](image)

Fig. 6 Classification Accuracy Result

5. DISCUSSION

5.1. Model Performance Evaluation

The results show that the sentiment classification model built is able to achieve an accuracy rate of 97.45% in predicting the sentiment of Oppenheimer movie reviews. The high precision and Recall values (97.46% and 97.45%) indicate that the model is not only accurate in identifying positive and negative sentiments, but also able to minimize classification errors.

These results show that the developed model is able to capture the patterns and characteristics present in Oppenheimer’s movie review data. The use of TF-IDF features in the feature extraction process also proved effective in giving proper weight to relevant words, thus improving the performance of the model in classifying sentiments.

5.2. Research Limitation

Although the results show good model performance, there are some limitations that need to be considered. One of the limitations is the assumption of independence between features underlying the Naive Bayes algorithm. In reality, words in a sentence often have interdependent relationships. In addition, this research only focuses on one type of film, namely Oppenheimer, so generalization of the research results to other films needs to be done carefully.

5.3. Implication and Suggestion

The results of this study have broad implications, especially in the field of sentiment analysis and system recommendations. The sentiment classification model developed can be used to understand people’s sentiments towards a movie or other topics, suggest relevant movies to users based on their preferences, and assist movie producers in developing products that suit market tastes.

For future research, some suggestions include using other more complex algorithms such as Support Vector Machine (SVM), Random Forest, or deep learning to see if it can improve the performance of the model, combining with techniques such as word embedding or dependency parsing to capture better word context, and also testing the model on a larger dataset and covering various movie genres to see the generalization of the model.

6. CONCLUSION

This study examines the use of the Naive Bayes algorithm to classify the sentiment of Oppenheimer movie reviews. The test results show that Naïve Bayes produces a high accuracy value, with an accuracy rate of 97.45%. This shows that the Naïve Bayes model is able to identify positive, negative and neutral sentiments in Oppenheimer movie reviews with a good level of precision. The ability of Naïve Bayes to handle text data is one of the key success factors in sentiment classification. This algorithm is able to identify patterns and relationships between words in the review, so that it can predict patterns and relationships between words in the review, so that it can predict sentiment precisely. In addition, the use of the TF-IDF method allows the model to focus on the most important words in the review thus improving its ability to classify sentiment.

To conclude, this study opens the gates to many exciting opportunities for further research, encouraging researchers to explore new horizons in sentiment analysis of movie reviews. Here are some suggestions that can be considered, namely, exploring other classification algorithms using algorithms other than Naïve Bayes. State-of-the-art algorithms such as Support Vector Machine (SVM) and Convolutional Neural Network (CNN) can also be tested to compare their

* Corresponding author
performance in movie review sentiment classification. Then also by enriching the dataset and scope of analysis, by incorporating reviews from various social media platforms, online forums, and review websites to increase the generalizability of the model. Also explore more complex aspects of sentiment by conducting more analysis.

7. REFERENCES


* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.