Bulletin of Computer Science and Electrical Engineering
Vol. 1, No. 1, June 2020, pp. 26~32
ISSN: 2722-7324, DOI: 10.25008/bcsee.v1i1.5

Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes

Risky Novendri[1], Annisa Syafarani Callista[2], Danny Naufal Pratama[3], Chika Enggar Puspita[4]
[1,2,3,4]Department of Information Systems, Telkom University

Article Info

Article history:

Received Apr 12, 2020
Revised May 19, 2020
Accepted Jun 15, 2020

Keywords:

Sentiment analysis
Naïve Bayes
Money Heist
Youtube
Opinion Mining

ABSTRACT

Netflix has produced many TV series, one of which is Money Heist. This series has four seasons with a total of 38 episodes. The fourth season was released on April 3, 2020, which has eight episodes. The fourth season of Money Heist is 31.73 times more demand than the average series around the world. However, despite the many requests for the fourth season of the Money Heist series, there are still some negative comments made by the connoisseurs of the Money Heist series. In the YouTube comments column on the Netflix channel, there are still many who comment neutral and provide positive comments on this series. Therefore, there needs to be a method in which viewers' comments or opinions can be analyzed in order to be able to classify the opinions they make about this series by conducting sentiment analysis using the Naïve Bayes algorithm. Based on the results of research conducted, Naive Bayes can be said to be successful in conducting sentiment analysis because it achieves results of 81%, 74.83%, and 75.22% for accuracy, precision, and recall, respectively.

Corresponding Author:

Risky Novendri,
Department of Information Systems,
Telkom University
Jl. Telekomunikasi Jl. Terusan Buah Batu, Bandung, Indonesia
Email: riskynovendri@gmail.com

1. INTRODUCTION

The 21st century is a century where technology has mastered many aspects, including media and digital-based entertainment, such as Netflix. According to IDN Times, Netflix has more than 50 million subscribers. If accumulated, all subscribers can watch up to 100 million hours of content in just one day. Netflix has produced many TV series, one of which is Money Heist. This series has four seasons with a total of 38 episodes. The fourth season was released on April 3, 2020, which has eight episodes. According to Parrot Analytics, quoted by Observer, the fourth season of Money Heist is 31.73 times more than the average series around the world and beat popular series such as Game of Thrones, The Walking Dead, Brooklyn Nine-Nine, and Westworld. Demand for the fourth season during the first three days has jumped 36.6 percent compared to the third season. However, despite the many requests for the fourth season of this Money Heist series. There are still some negative comments made by connoisseurs of the Money Heist series. In the YouTube Netflix comments column, there are still many who comment neutral and give positive comments on this series. For this reason, there needs to be a method in which viewers' comments or opinions can be analyzed in order to be able to classify the opinions they make about this series by conducting sentiment analysis using the Naïve Bayes algorithm. Naïve Bayes was chosen because based on previous research, satisfactory results were obtained [1]. Naive Bayes has been used and implemented in various areas, such as economics and business, engineering, education, health, and so on [2]–[8].

Sentiment analysis is a computational-based method of analysis of opinions, sentiments and emotions [9]. Sentiment analysis is used to see the tendency of a sentiment, whether the opinion is positive, neutral, or negative. Sentiment analysis is done by processing data in the form of the text so that it is ready to be analyzed by text mining methods. Text pre-processing consists of labeling, tokenization, stemming and filtering. One algorithm that is more often used in text classification is the Naïve Bayes algorithm. Naïve Bayes algorithm has been widely used in the process of text mining because it has a simple algorithm but has a high accuracy [10]. Naïve Bayes is also used in classifying texts because it has a basic concept that combines the probability of words and categories of documents.

In this study, the data will be taken from the comments on the video trailer for the fourth season of Money Heist on YouTube which has been processed and labeled manually with positive, neutral, and negative sentiments. It aims at analyzing sentiment regarding the audience's response to the emergence of the fourth season of Money Heist based on YouTube data mining.

2. LITERATURE REVIEW

2.1 Sentiment Analysis

Sentiment analysis is a technique or method used to identify the expression of the sentiment using text and how the sentiment can be categorized as positive or negative sentiment [11]. According to Indurkhya & Damerau [9], sentiment analysis or opinion mining refers to a broad field of natural language processing, linguistic computing and text mining that aims to analyze the opinions, sentiments, evaluations, attitudes, judgments, and emotions of a speaker or writer regarding a topic, products, services, organizations, individuals, or certain other activities.

Sentiment analysis is used to understand comments made by internet users and explain how a product or brand is received by them [12]. Internet users write down their experiences, opinions, and everything that concerns them based on how they feel, and these feelings can be positive, neutral, or negative feelings that can be expressed in a fairly complex manner [13].

2.2 Text Mining

Text mining is a text analysis where data sources can be obtained from documents, and the purpose is to find words that can represent the contents of documents so that analysis can be done relatedness, interrelation, and class between documents (Lesmeister, 2015). Text mining involves the processing of documents into text categorization, information extraction, and word extraction. This method is used to extract information from data sources through the identification and exploration of interesting data [9].

Text mining is a technique that can be used to deal with classification, clustering, information extraction, and information retrieval problems [14]. Text mining in sentiment analysis is able to identify emotionally about a statement [15]. Text mining usually refers to several combinations of relevance, novelty, and interestingness. The process of working on text mining is widely adopted from data mining research. However, the difference is the pattern used by text mining which is taken from a collection of unstructured natural languages, while data mining patterns are taken from a structured database. Typical text mining processes include text categorization, text clustering, concept/entity extraction, granular taxonomic production, sentiment analysis, document inference, and entity-relationship modeling, namely learning the relationship between entities [8].

2.3 Naïve Bayes

Naïve Bayes algorithm is an algorithm used to find the highest probability value to classify test data in the most appropriate category. This algorithm uses probability and statistical methods that were first discovered by British scientist Thomas Bayes, which predicts future opportunities based on experience so that it is known as the Bayes Theorem. Naïve Bayes Classifier is one of the most popular algorithms used for data mining purposes because of its ease of use [16] and fast processing time, easy to implement with a fairly simple structure and high level of effectiveness [17].

Naïve Bayes calculates the probability of a class based on its attributes and determines the class that has the highest probability. Naïve Bayes classifies classes based on simple probabilities by

assuming that each attribute in the data is mutually exclusive. In the probability model, each class k and the number of attributes can be written as in the equation below:

$$P = (y_1|x_1, x_2, \cdots, x_n) \quad (1)$$

Naïve Bayes calculation is the probability of the appearance of document X_a in the class category of Y_k $P(x_a|y_k)$, multiplied by the probability of the class category $P(y_k)$. From the results, the distribution of the documents $P(x_a)$ will occur. So, we get the Naïve Bayes calculation formula written in the equation:

$$P(y_k|x_a) = \frac{P(y_k)P(x_a|y_k)}{P(x_a)} \quad (2)$$

Then the optimal class selection process is carried out, then the largest opportunity value of each class probability is selected. Then the formula is obtained to choose the largest value as in the following equation:

$$y(x_i) = \arg\max P(y) \prod_{i=1}^{a} P(x_1|y) \quad (3)$$

Weighting a class attribute can increase the effect of prediction. By calculating the weight of attributes to class, then the basis for classification accuracy is not only probabilities but also the weight of each class attribute [10].

3. RESEARCH METHOD

3.1 Crawling Dataset

In this study, the authors implemented a crawling technique on Youtube about Money Heist films using web testing automation with Selenium WebDriver and Python as shown in Figure 1.

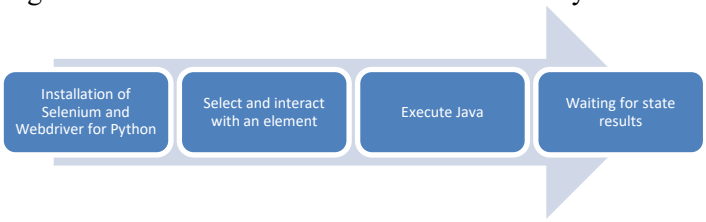

Figure 1. The steps of crawling dataset

The explanation of the above steps is as follows:

1. Installing Selenium and WebDriver for Python

Previously what had to be done was to install WebDriver which served as a link between SWD and the web browser so that it could be controlled through the programming language. After WD is installed, we need to install the SWD with the Python programming language by managing the "pip" package to install the package named Selenium.

2. Select and interact with an element

Here we connect an SWD element with the YouTube link that we will create as a dataset. Once connected we will be able to retrieve existing comment data on YouTube through the python application that will be analyzed.

3. Execution of Java

At this stage we execute Java on the page that is being opened by calling the execute_script function via the Web Driver and specifying the Java snippet code that we want to execute, we also retrieve the variable value on the page we are executing.

4. Waiting for state results

At this stage, we can do the testing process through a function that returns true if the expected state is reached, and false if the state is not reached and the program continues to wait. If the program runs false more than 10 times, then we have to "run" the program again to get the value true.

3.2 Sentiment Analysis

In this study, the authors also conducted a labeling process for sentiment analysis. This labeling process is done manually by using 3 labels or using a sentiment word dictionary consisting of:

1. The word positive sentiment

The word positive sentiment here is taken according to good-spoken comments and does not contain ridicule of the film.

2. Negative sentiment words

We take the word negative sentiment according to comments that are offensive or contain satirical words about this film.

3. The word neutral sentiment

The word neutral sentiment here is a comment that is not offensive or discussed outside of this film.

3.3 Preprocessing

Based on the irregularity of the text data structure, the process of retrieval of information or text mining requires several initial stages which in essence are preparing so that the text can be changed to be more structured. One implementation of text mining is the text preprocessing stage. Text preprocessing is the stage where the application selects data to be processed in each document or sentence. In Figure 2, this preprocessing process includes:

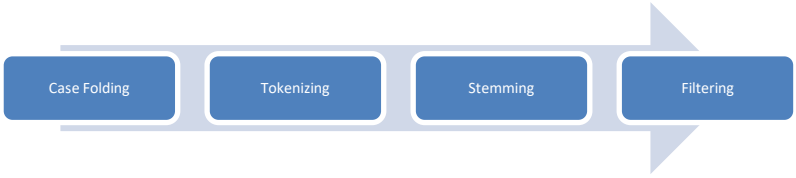

Figure 2. Preprocessing steps

The following is an explanation of the pre-processes above:

1. Case Folding

Not all text documents are consistent in the use of capital letters. Therefore, the role of Case Folding is needed in converting the entire text in a document into a standard form (usually lowercase or lowercase). Case Folding is changing all the letters in a document into lowercase letters. Only the letters 'a' to 'z' are accepted. Characters other than letters are omitted and are considered a delimiter.

2. Tokenizing

The tokenizing stage is the stage of cutting the input string based on each word that makes it up. Tokenisasi outline breaks a group of characters in a text into word units, how to distinguish certain characters that can be treated as word separators or not.

3. Stemming

The stemming technique is needed in addition to reducing the number of different indexes of a document, also to do a grouping of other words that have similar basic words and meanings but have different forms or forms because they get different affixes. The stemming process in Indonesian texts is different from stemming in English texts. In English texts, the only process required is the process of removing suffixes. While in the Indonesian-language text all affixes both suffixes and prefixes are also omitted.

4. Filtering

The filtering stage is the stage of taking important words from the token results. Can use the Stoplist algorithm (discarding less important words) or Wordlist (save important words). Stoplist / Stopword are non-descriptive words that can be discarded in a bag-of-words approach.

3.4 Feature Extraction

In this study, the authors also performed the Feature Extraction Phase which is about TF-IDF. TF-IDF is used to measure how relevant each word in a document is in several documents if sentiment documents are the same as sentences.

3.5 Processing

In this study, the authors also performed the Feature Extraction Phase which is about Learning Model. Learning Model is a processing phase which is carried out before getting the accuracy results on the evaluation metrics.

3.6 Apply Model and Performance Calculation

At this stage, we apply the processed model to produce the accuracy and data we need. This stage will present the final results regarding the performance of Naïve Bayes for sentiment analysis.

4. RESULTS AND DISCUSSION

4.1 Dataset

In this study using the crawling dataset sourced from trailer Money Heist session 4 on YouTube (https://www.youtube.com/watch?v=p_PJbmrX4uk). The sample of 998 comments data was collected, the attributes that were not used were deleted and only comments were kept. The 5 (five) sample datasets used in this study are presented in Table 1. In the table, the labels 0, 1, 2 indicate positive, negative, and neutral sentiments, respectively.

Table 1. Sample Dataset crawled from Youtube

No	Comment	Sentiment
1	We need a never-ending season of money heist until all the characters have kids and will be a generation haha	0
2	The best thing what Netflix gave is Money Heist	0
4	Dunno why Arturito is still alive tho man is so freaking annoying..he turns my tummy in a bad way	1
5	The only bad thing about this season is that it ends up with double suspense	1
3	I told him to watch professor	2

The dataset has been preprocessed and has been manually labeled. The next phase is preprocessing, feature extraction, implemented in Naive Bayes and its performance is measured using a confusion matrix.

4.2 Model Evaluation

After the preprocessing phase, feature extraction uses TF-IDF, the next phase is an implementation using Naive Bayes and its performance is measured using a confusion matrix. The dataset is then divided into two, with a portion of 80% for training models and the remaining 20% for testing models. In the testing phase of the model, the Naïve Bayes algorithm is measured using a confusion matrix. In Figure 3 the following explains the four divisions in the confusion matrix.

Confusion Matrix		*Modeled Values: x_m*	
		True	*False*
Actual Values: x	*True*	*TP*	*FN (Type II error)*
	False	*FP (Type I error)*	*TN*

Figure 3. Confusion Matrix

The confusion matrix obtained from the evaluation metrics, namely the score of accuracy, precision, and recall.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (4)$$

$$Precision = (TP)/(TP + FP) \quad (5)$$

$$Recall = (TP)/(TP + FN) \quad (6)$$

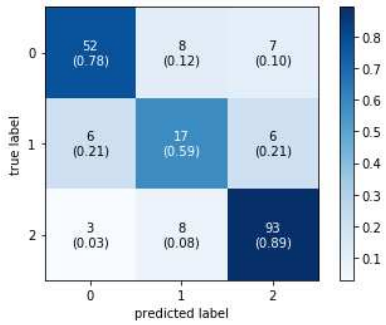

Figure 4. The results of the confusion matrix

Figure 4 shows the results of the confusion matrix. The confusion matrix obtained from the evaluation metrics, namely the value of accuracy, precision, and recall, are presented in Table 2. The results for the accuracy, precision, and recall scores were 81%, 74.83%, and 75.22%, respectively.

Table 2. The result of evaluation metrics

Accuracy	Precision	Recall
81%	74.83%	75.22%

5. CONCLUSION

Sentiment analysis is a computational-based method of analysis of opinions, sentiments, and emotions. Sentiment analysis is used to see the tendency of a sentiment, whether the opinion is positive, neutral, or negative. In this study sentiment analysis of the Money Heist session 4 movie trailer on YouTube. From the results of sentiment analysis, it is found that comments tend to be many positive ones so it can be concluded that the film is in demand by the audience. Based on the results of research conducted, Naive Bayes can be said to be successful in conducting sentiment analysis because it achieves results of 81%, 74.83%, and 75.22% for accuracy, precision, and recall, respectively.

REFERENCES

[1] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Jan. 2008.
[2] L. C. Huang, S. Y. Hsu, and E. Lin, "A comparison of classification methods for predicting chronic fatigue syndrome based on genetic data," *J. Transl. Med.*, vol. 7, p. 81, 2009.
[3] M. Wibowo, S. Sulaiman, and S. M. Shamsuddin, "Comparison of Prediction Methods for Air Pollution Data in Malaysia and Singapore," *Int. J. Innov. Comput.*, vol. 8, no. 3, pp. 65–71, 2018.
[4] E. Sutoyo and A. Almaarif, "Educational Data Mining for Predicting Student Graduation Using the Naïve Bayes Classifier Algorithm," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 95–101, 2020.
[5] A. Aninditya, M. A. Hasibuan, and E. Sutoyo, "Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 2019, pp. 112–117.
[6] E. Sutoyo and A. Almaarif, "Twitter Sentiment Analysis of The Relocation of Indonesia's Capital City," *Bull. Electr. Eng. Informatics*, vol. 9, no. 04, pp. 1620–1630, 2020.
[7] N. Altrabsheh, M. M. Gaber, and M. Cocea, "SA-E: Sentiment analysis for education," *Front. Artif. Intell. Appl.*, vol. 255, pp. 353–362, 2013.
[8] P. Bhargavi and S. Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 9, no. 8, pp. 117–122, 2009.
[9] N. Indurkhya and F. J. Damerau, *Handbook of natural language processing, second edition*. 2010.
[10] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, vol. 3, no. 22, pp. 41–46.
[11] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP 2003*, 2003.
[12] I. P. Cvijikj and F. Michahelles, "Understanding social media marketing: A case study on topics, categories and sentiment on a Facebook brand page," in *Proceedings of the 15th International*

Academic MindTrek Conference: Envisioning Future Media Environments, MindTrek 2011, 2011.
[13] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of Facebook statuses using Naive Bayes Classifier for language learning," in *IISA 2013 - 4th International Conference on Information, Intelligence, Systems and Applications*, 2013.
[14] M. W. Berry and J. Kogan, *Text Mining: Applications and Theory*. 2010.
[15] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis," *HP Lab. Tech. Rep.*, 2011.
[16] M. Hall, "A decision tree-based attribute weighting filter for Naive Bayes," in *Research and Development in Intelligent Systems XXIII - Proceedings of AI 2006, the 26th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 2007.
[17] S. Taheri and M. Mammadov, "Learning the naive bayes classifier with optimization models," *Int. J. Appl. Math. Comput. Sci.*, 2013.