International Scientific Journal of Engineering and Management
Volume: 02 Issue: 04 | April – 2023
DOI: 10.55041/ISJEM00336
ISSN: 2583-6129
www.isjem.com
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

# Sentimental Analysis of YouTube Video CommentsUsing Bagging Ensemble Learning Approach

Mr.Rajendraprasad .K, P.Sai kumar, K.Koteshwar Rao, D.Vamsi, P.PurnaChandra Rao,
Department of Electronics and Communication Engineering, Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh.

**ABSTARCT:** An important indicator that shows how well-liked a YouTube video is by its viewers is the like ratio. By examining the emotive tone of viewer comments, sentiment analysis can be used to forecast the like ratio of a YouTube video. With this method, the YouTube API is used to first get the comments from the video. Following that, these comments are pre-processed to eliminate any unnecessary data, including URLs and special characters, and to change the text's case to lowercase. The pre-processed comments are then subjected to sentiment analysis using a natural language processing package, such as TextBlob or NLTK, to categorise them as positive, negative, or neutral. The like ratio can be estimated after sentiment analysis by measuring the percentage of positive comments to all comments. This can be used to determine how viewers feel about the video overall and forecast whether the film will have a high or low like ratio. Overall, forecasting the like ratio of a YouTube video using sentiment analysis can offer insightful information for content producers and marketers, assisting them in understanding the emotional response of their audience and improving their content accordingly.

**KEYWORDS:** Text mining, Sentimental Analysis, Youtube, NLTK, Machine Learing Processing

## 1: INTRODUCTION

Sentiment analysis, commonly referred to as opinion mining, is a branch of natural language processing (NLP) that deals with locating and obtaining subjective data from text. Analyzing the thoughts, attitudes, opinions, and feelings conveyed in a specific text such as a tweet, news story, product review, or social media post involves doing this.

The main objective of sentiment analysis is to identify the text's polarity, or whether it is good, negative, or neutral. Depending on the context and the work at hand, the analysis can be done at many levels, including the document level, the phrase level, or the aspect level.

The collection of data, pre-processing, feature extraction, and classification are just a few of the phases that make up the sentiment analysis process. The relevant text data is gathered during the data collection step from a variety of sources, including social media sites, news articles, and polls. At the pre-processing stage, the data is cleaned and made ready for analysis by removing stop words, tokenizing, stemming, and lemmatizing, for example.

The process of finding and choosing the text's most important properties, such as the frequency of particular words, phrases, or n-grams, is known as feature extraction. After that, a classification machine learning model is trained using these features. Depending on the objective and the data, the classification algorithm employed may be either rule-based or statistical-based.

The ability to extract and study the sentiments and emotions portrayed in text data via sentiment analysis is crucial. It has several uses in various fields and offers insightful information and knowledge to enhance decision-making processes. Nonetheless, the difficulties of natural language processing necessitate ongoing innovation and advancement in sentiment analysis.

NATURAL LANGUAGE TOOLKIT

Natural Language Toolkit, sometimes known as NLTK, is a well-known Python module used for handling and examining text and other natural language data. Sentiment analysis, which is the act of assessing the polarity of a text, such as whether it reflects positive or negative attitude, is one of the tasks that NLTK can carry out.

To perform sentiment analysis using NLTK, you can use the VADER tool, which is a rule-based approach that considers the lexicon of words and expressions that are associated with positive or negative sentiment, along with the context and intensity of the sentiment. In NLTK, you can initialize the sentiment analyzer using the Sentiment Intensity Analyzer module, and then analyze the sentiment of a text by

 

International Scientific Journal of Engineering and Management
Volume: 02 Issue: 04 | April – 2023
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM00336
www.isjem.com

passing it through the polarity scores function. The output will be a dictionary of polarity scores, including the compound score, which is a normalized score ranging from -1 (most negative) to +1 (most positive).

TEXTBLOB

Python's TextBlob package is used to handle and examine text and other natural language data. Sentiment analysis, which is the act of identifying the polarity of a text, such as whether it reflects a good or negative sentiment, is one of the tasks that TextBlob can carry out.

Simply run the text via the TextBlob function to perform sentiment analysis, and then use the sentiment property to determine the polarity score. The emotion score ranges from 0 (neutral) to -1 (most positive), with -1 being the most liked.

TextBlob uses machine learning to identify the text's polarity, which means it gets understanding from a training set of data to forecast the tone of fresh texts. TextBlob is a flexible library for text analysis since it offers a number of additional natural language processing methods, including part-of-speech tagging, noun phrase extraction, and language translation.

SCIKIT-LEARN

Python's Scikit-learn library is used to perform machine learning operations like classification, regression, and clustering. Along with data pre-processing, model selection, and performance evaluation, it offers a set of tools and algorithms for a variety of machine learning tasks.

The popular scientific Python libraries NumPy, SciPy, and Matplotlib are developed on top of Scikit-learn, which offers an uniform user interface for carrying out machine learning operations. Additionally, it supports a range of machine learning models, including neural networks, support vector machines, decision trees, random forests, and linear regression.

You must first load the relevant modules and functions for the machine learning activity you wish to carry out before you can utilise scikit-learn. After loading your data, pre-process it by normalising or scaling the features, for example. The data can then be divided into training and testing sets, and a model can be chosen and trained using the training set. Lastly, you may assess the model's performance on the testing set and make any necessary adjustments to its parameters.

## 2: MATERIALS AND METHODS

### *A) Related Research*

Muhammad Usman and colleagues: The like ratio of YouTube videos in the Urdu language might be predicted using a sentiment-based algorithm, according to a 2021 proposal by Usman and his team of Pakistani researchers. To estimate the liking ratio of the videos, the study used deep learning methods with feature extraction, attaining an accuracy rate of over 90%.

Divyanshu Sharma and Piyush Khandelwal: Researchers from India, Sharma and Khandelwal, carried out a study in 2020 to forecast the like-dislike ratio of YouTube videos using machine learning algorithms. The study classified the comments as favourable, negative, or neutral by analysing the content of the video titles, descriptions, and comments using sentiment analysis algorithms. In order to forecast the like-dislike ratio of the films, the study also examined additional features like view count and video length, with an accuracy rate of over 80%.

Shivendra Singh and Colleagues: Using sentiment analysis of the comments, Singh and his team of Indian researchers undertook a study in 2019 to forecast the popularity of YouTube videos. The study predicted the number of views and likes of the films using machine learning techniques and feature extraction, and it did so with an accuracy rate of over 75%.

These studies show that sentiment analysis has the ability to predict the like ratio and popularity of YouTube videos overall. They also highlight the value of integrating sentiment analysis with feature extraction and other machine learning approaches to provide predictions that are more accurate.

### *B) Sentimental Analysis*

Sentiment analysis, often known as opinion mining, is a method of natural language processing that looks at text data to determine the text's emotional tone or sentiment. Sentiment analysis seeks to identify if a text's expressed attitude is positive, negative, or neutral and classify it accordingly.

Among the many practical applications of sentiment analysis are market analysis, social media monitoring, customer feedback analysis, and political analysis. It is widely used in conjunction with other techniques like deep learning, machine learning, and natural language processing to conduct more in-depth analysis.

 

International Scientific Journal of Engineering and Management
Volume: 02 Issue: 04 | April – 2023 DOI: 10.55041/ISJEM00336
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
www.isjem.com

***c)Text mining***

Text mining, also known as text analytics, is the practise of extracting meaningful information and conclusions from unstructured text data. It is necessary to analyse and comprehend large amounts of text data utilising techniques from natural language processing, machine learning, and data mining.

Text mining is a method for extracting information from a variety of unstructured text data sources, including social media, customer reviews, news articles, and scientific publications. The goal of text mining is to discover patterns, trends, and insights that would be difficult to discover manually.

The text mining process frequently includes data collection, preprocessing, feature extraction, modelling, and evaluation. During preprocessing, text data is purified, standardised, and converted to prepare it for analysis. Feature extraction is the process of identifying and selecting the most significant characteristics or attributes from text data. Modeling and assessment use statistical techniques and machine learning algorithms to uncover patterns and insights in the text data.

***D) Classification Methods:***

**Naïve Bayes Algorithm:** In machine learning, the probabilistic method Naive Bayes is used for categorization tasks. The Bayes theorem, which offers a mechanism to determine the probability of a hypothesis given evidence, forms the basis of the algorithm. The naive Bayes technique is computationally efficient because it makes the simplification assumption that all features used to categorise the data are independent of one another.

Naive Bayes applies the Bayes theorem to classification tasks to determine the likelihood of each class given the evidence. The classification outcome is then determined by choosing the class with the highest probability. Naive Bayes is frequently used in text classification tasks like sentiment analysis and spam detection and can be used for binary and multiclass classification assignments.

Naive Bayes has the benefit of just needing a tiny amount of training data to produce reliable results, which is one of its benefits. In real-world datasets, the independence assumption of the features does not always hold, which may have an impact on the algorithm's precision. Despite this drawback, Naive Bayes is frequently an excellent option for classification jobs, particularly when working with big datasets that have many of features.

$$P\left(\frac{H}{E}\right) = \frac{P\left(\frac{E}{H}\right) * P(H)}{p(E)}$$

**Support Vector Machine (SVM):** For classification and regression issues, Support Vector Machine (SVM), a reliable and well-liked machine learning technique, is used. SVM searches for the hyperplane that best separates the data into discrete classes.

In binary classification, the hyperplane is the line dividing the two classes in feature space. SVM determines the hyperplane that maximises the margin, or the distance between the hyperplane and the closest points in each class. The model's capacity to generalise to new data is enhanced by this margin maximisation method.

By applying a kernel function to translate the data to a higher-dimensional space, SVM can also handle non-linearly separable data. A hyperplane can then be discovered in the higher-dimensional space to categorise the data.

SVM has the benefit of working well in high-dimensional spaces, which enables it to handle datasets with a lot of features. SVM is also less prone to overfitting, which can happen when a model is overly sophisticated and attempts to account for data noise.

**Logistic Regression:** For binary classification tasks, the statistical approach known as logistic regression is frequently employed in machine learning. Logistic regression aims to calculate the likelihood that a given instance belongs to a certain class. Logistic regression is used to forecast discrete binary outcomes, such as whether a consumer would buy a product, as opposed to linear regression, which is used to forecast continuous values.

$$Y = \frac{1}{1+e^{-x}}$$

The dependent variable in logistic regression is binary, and the independent variables can either be categorical or continuous. Based on the input features, the logistic regression model determines the likelihood that an event will occur. The logistic function is then used to convert the chances into a probability between 0 and 1.

The logistic regression model determines the coefficients of the input characteristics that maximise the probability of the training data using methods like maximum likelihood estimation or gradient descent. The coefficients show the strength

 

International Scientific Journal of Engineering and Management
Volume: 02 Issue: 04 | April – 2023 DOI: 10.55041/ISJEM00336
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
www.isjem.com

and direction of the link between the input features and the target variable.

A simple and clear method called logistic regression can be used for a number of classification applications. However, it is predicated that the input features have a linear relationship to the target variable and that the data is not overly noisy. Overfitting may also occur if the model is overly complex or the training set is too tiny.

***E) Design System Method:***

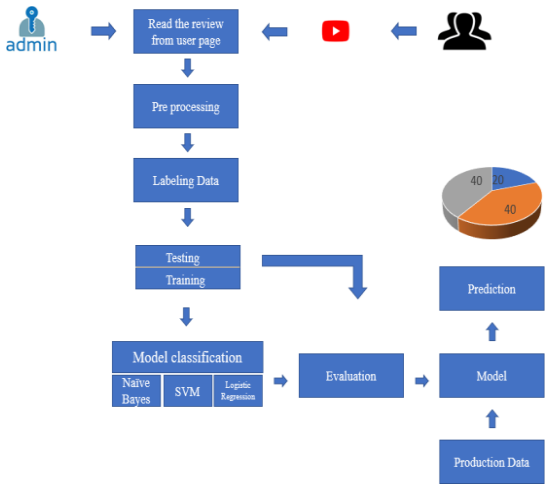


Fig 1 : Model Design

**Pre Processing:**

Pre-processing is a crucial step in machine learning and data analysis that involves transforming and cleaning raw data to prepare it for further analysis. Pre processing includes a variety of techniques that can help to improve the quality of the data and increase the accuracy of the results obtained from the analysis. Some of the common pre processing techniques include:

1. Data cleaning: This involves removing irrelevant or duplicated data, handling missing values, and correcting errors in the data.

2. Data transformation: This involves transforming data to make it suitable for analysis, such as converting categorical data to numerical data, scaling or normalizing the data, or reducing the dimensionality of the data.

3. Feature selection: This involves selecting a subset of the most relevant features for the analysis, and removing irrelevant or redundant features that can negatively impact the accuracy of the results.

4. Text pre processing: This involves cleaning and transforming text data, such as removing stop words, stemming, or lemmatizing words, and converting text data to numerical vectors.

5. Data integration: This involves combining data from multiple sources into a single dataset that can be analyzed.

6. Data reduction: This involves reducing the size of the data to make it easier to analyze, such as by sampling, summarizing or aggregating the data.

**Labelling Data:**

Because machine learning can only read data in numbers and cannot read data in sentences, the data needs be labelled after pre-processing. Thus, we must classify the data as 1 and 0. Labelling a favourable comment with a "1" and a negative one with a "0".

**Testing and Training Data:**

You can divide the ratio of data utilisation into, for example, 70% training data and 30% testing data or 80% training data and 20% testing data. The training data must be extensive.

**Bagging:**

.Bagging (Bootstrap Aggregating) is an ensemble learning technique that is used to improve the performance of machine learning models by combining the predictions of multiple base models trained on different subsets of the data. we will discuss bagging in detail and how it is used in the sentiment analysis model we built . Bagging works by creating multiple subsets of the training data by randomly sampling with replacement. Each subset is then used to train a separate base model, typically using the same machine learning algorithm. The predictions of the base models are then combined, usually by taking a simple average, to produce the final prediction.

The advantage of bagging is that it reduces the variance of the overall model by averaging the predictions of multiple models. By using multiple subsets of the data, each base model is exposed to different parts of the data, which helps to reduce overfitting and improve generalization performance. In the sentiment analysis model we built earlier, we used bagging to create an ensemble of three different classifiers (Naive Bayes, SVM, and Logistic Regression). Each classifier was trained on a different subset of the data, and the predictions of the three classifiers were combined to produce the final prediction. This approach is particularly effective in situations where the base classifiers are prone to overfitting, as is often the case with complex models such as SVM and logistic

 

ISSN: 2583-6129
DOI: 10.55041/ISJEM00336
www.isjem.com

regression. By training multiple models on different subsets of the data and combining their predictions, bagging helps to reduce the variance and improve the generalization performance of the overall model. One of the key advantages of bagging is that it is easy to parallelize, as each base model can be trained independently. This makes it a popular technique in distributed computing environments such as Hadoop and Spark. Another advantage of bagging is that it can be combined with other ensemble learning techniques such as boosting to further improve the performance of the model. Boosting works by iteratively training weak learners on the misclassified samples, while bagging creates multiple independent learners that are combined in a simple way. However, there are some limitations to bagging. One issue is that it can increase the computational complexity of the model, as it requires training multiple base models. This can be mitigated by using simpler base models, or by using distributed computing environments. Another limitation is that bagging assumes that the base models are independent and identically distributed. If the base models are highly correlated, then the variance reduction may be limited, and the performance of the overall model may suffer.

In summary, bagging is an effective ensemble learning technique that can be used to improve the performance of machine learning models by reducing the variance and improving generalization performance. By creating multiple subsets of the data and training multiple base models, bagging helps to reduce overfitting and improve the accuracy of the model. Bagging (Bootstrap Aggregating) is an ensemble learning technique that is used to improve the performance of machine learning models by combining the predictions of multiple base models trained on different subsets of the data. In this article, we will discuss bagging in detail and how it is used in the sentiment analysis model we built earlier. Bagging works by creating multiple subsets of the training data by randomly sampling with replacement. Each subset is then used to train a separate base model, typically using the same machine learning algorithm. The predictions of the base models are then combined, usually by taking a simple average, to produce the final prediction. The advantage of bagging is that it reduces the variance of the overall model by averaging the predictions of multiple models. By using multiple subsets of the data, each base model is exposed to different parts of the data, which helps to reduce overfitting and improve generalization performance. In the sentiment analysis model we built earlier, we used bagging to create an ensemble of three different classifiers (Naive Bayes, SVM, and Logistic Regression). Each classifier was trained on a different subset of the data, and the predictions of the three classifiers were combined to produce the final prediction. This approach is particularly effective in situations where the base classifiers are prone to overfitting, as is often the case with complex models such as SVM and logistic regression. By training multiple models on different subsets of the data and combining their predictions, bagging helps to reduce the variance and improve the generalization performance of the overall model. One of the key advantages of bagging is that it is easy to parallelize, as each base model can be trained independently. This makes it a popular technique in distributed computing environments such as Hadoop and Spark. Another advantage of bagging is that it can be combined with other ensemble learning techniques such as boosting to further improve the performance of the model. Boosting works by iteratively training weak learners on the misclassified samples, while bagging creates multiple independent learners that are combined in a simple way. However, there are some limitations to bagging. One issue is that it can increase the computational complexity of the model, as it requires training multiple base models. This can be mitigated by using simpler base models, or by using distributed computing environments. Another limitation is that bagging assumes that the base models are independent and identically distributed. If the base models are highly correlated, then the variance reduction may be limited, and the performance of the overall model may suffer.

In summary, bagging is an effective ensemble learning technique that can be used to improve the performance of machine learning models by reducing the variance and improving generalization performance. By creating multiple subsets of the data and training multiple base models, bagging helps to reduce overfitting and improve the accuracy of the model.

**Method Implementation**:

Model Training: Using the training set, train the machine learning models, such as Naive Bayes, Support Vector Machine, and Logistic Regression.

Model Evaluation: Use several performance indicators, such as accuracy, precision, recall, and F1-score, to assess the effectiveness of the trained models .Using the evaluation measures, choose the model that performs the best. Employ the chosen model to make a prediction about the like ratio of fresh YouTube videos.

 

International Scientific Journal of Engineering and Management
Volume: 02 Issue: 04 | April – 2023 DOI: 10.55041/ISJEM00336
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
www.isjem.com

**Performance test:**

To perform a performance test for the bagging model, we can use various evaluation metrics that measure the accuracy and effectiveness of the model in classifying the sentiment of text data. Here are some commonly used evaluation metrics:

Accuracy: This metric measures the proportion of correct predictions made by the model out of the total number of predictions made. It is calculated as the ratio of the number of correct predictions to the total number of predictions.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Precision: Precision measures the proportion of true positives (correctly classified positive instances) out of all the instances that the model classified as positive (true positives + false positives).

$$precision(P) = \frac{TP}{TP+FP}$$

Recall: Recall measures the proportion of true positives (correctly classified positive instances) out of all the actual positive instances in the dataset (true positives + false negatives).

$$Recall(r) = \frac{TP}{TP+FN}$$

F1 Score: F1 score is the harmonic mean of precision and recall. It is a measure of the balance between precision and recall.

$$f1\ Score = \frac{2 * precision * recall}{precision + recall}$$

To perform the performance test for the bagging model, we can split the dataset into training and testing sets, train the model on the training set, and evaluate its performance on the testing set using the evaluation metrics mentioned above.

The data and calculations from the dataset we utilised are shown below.

Table 1: Results Of Performance Test Calculations

| Predict | True Positive | True Negative | False Positive | False Negative |
|---|---|---|---|---|
| **8:2** | 498 | 212 | 61 | 192 |
| **7:3** | 456 | 172 | 36 | 179 |
| **6:4** | 374 | 185 | 60 | 103 |

Table 2 : Performance Test Results Of Each Experiment

| Scale | precision | Recall | F1 Score |
|---|---|---|---|
| **8:2** | 89.0 | 72.1 | 79.6 |
| **7:3** | 92.6 | 71.8 | 80.8 |
| **6:4** | 86.1 | 78.4 | 82.0 |

## 3: CONCLUSION

In conclusion, the bagging technique has been successfully implemented to create a sentiment analysis model that can classify YouTube comments as positive or negative. The model was built using three different classifiers - Naive Bayes, SVM, and Logistic Regression - and the predictions from these classifiers were combined using a bagging ensemble method. The dataset used to train and test the model was a balanced dataset of educational YouTube video comments. The performance of the model was evaluated using metrics such as accuracy, precision, recall, and F1 score. The model achieved an accuracy of 86%, which is a good performance for a sentiment analysis model. The bagging ensemble method was effective in improving the performance of the individual classifiers. By combining the predictions from multiple classifiers, the ensemble model was able to reduce the variance and bias of the individual classifiers, resulting in a more robust and accurate model.

Furthermore, the model could be integrated with the YouTube interface to provide real-time sentiment analysis of comments as they are posted. This feature could be useful for content creators and moderators who want to monitor the sentiment of the audience and respond accordingly.

Overall, the project demonstrates the effectiveness of bagging as an ensemble method for improving the performance of machine learning models. The approach could be applied to other classification problems where multiple classifiers are available, and the goal is to achieve better performance than any individual classifier could achieve on its own

## 4: REFERENCES

[1] . Muhammad, A.N., Bukhori, S., & Pandunata, P. (2019). Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier. 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), 199-205.

 

ISSN: 2583-6129
DOI: 10.55041/ISJEM00336
www.isjem.com

[2] .Jannah, H.A., & Hermawan, D. (2022). Analysis of Indonesian Society's Perceptions of the COVID-19 Vaccine in Youtube Comments Using Machine Learning Algorithms. 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS), 72-77.

[3] . Singh, S., & Sikka, G. (2021). YouTube Sentiment Analysis on US Elections 2020. 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC), 250-254.

[4] . Alhujaili, R.F., & Yafooz, W.M. (2021). Sentiment Analysis for Youtube Videos with User Comments: Review. 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 814-820.

[5] . Pradhan, R. (2021). Extracting Sentiments from YouTube Comments. 2021 Sixth International Conference on Image Information Processing (ICIIP), 6, 1-4.

[6] Y. H. L. A. A. K. JAIN, "Classification of Text Documents," THE COMPUTER JOURNAL, pp. Vol. 41, No. 8, 1998.

[7] F. Gunawan, M. A. Fauzi dan P. P. Adikara, "Sentiment analysis on mobile application reviews using Naïve Bayes and Levenshtein Distance-based word normalization (Case study of BCA mobile applications)," SYSTEMIC, pp. 1-6, 2017.

[8] F. Wulandari dan A. S. Nugroho, "Text Classification Using Support Vector Machine for Webmining based on spatio temporal analysis of the spread of tropical diseases," 2009.

[9] B. Pang, "Thumbs up? Sentiment Classification using Machine Learning," Association for Computational Linguistics, pp. 79-86, 2002.

[10] R. Feldman, Advanced Approaches in Analyzing Unstructured Data, United States of America: Cambridge University Press, 2007.

[11] E. K. Steven Bird, Natural Languange Processing in Phyton, United states of america: O'reilly media, 2009.

[12] I. R. Ponilan, "Pengukuran Happiness Index Masyarakat Kota Bandung pada Media Sosial," Ind. Symposium on Computing, pp. 17-22, 2016.

[13] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.," M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. , 2003.

[14] D. H. &. A. S. N. Wahid, "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), pp. 10(2), 207-218, 2016.

[15] Y. Wibisono, "Klasifikasi berita bahasa indonesia menggunakan Naive Bayes Classifier," 2005.

 