JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Sentiment Analysis for Political Debates on YouTube Comments using BERT Labeling, Random Oversampling, and Multinomial Naïve Bayes

Apriandy Angdresey, Lanny Sitanayah *, and Ignatius Lucky Henokh Tangka

Department of Informatics Engineering, Universitas Katolik De La Salle, Manado, 95253, Indonesia;
e-mail : aangdresey@unikadelasalle.ac.id; lsitanayah@unikadelasalle.ac.id; 20013047@unikadelasalle.ac.id
* Corresponding Author : Lanny Sitanayah

Abstract: The 2024 Indonesian Presidential Election marked the fifth general election in the country, aimed at electing a new President and Vice President for the 2024–2029 term. Candidates competed to succeed the outgoing president, who had served two constitutional terms. A key aspect of this election was the candidate debates, where each candidate presented their vision, allowing the public to assess their policies. These debates were broadcast on platforms like YouTube, giving the public a space to comment. However, analyzing YouTube comments presents challenges due to the volume of data, language diversity, and informal expressions. Sentiment analysis, crucial for understanding public opinion, uses algorithms such as Naïve Bayes, which is based on Bayes' Theorem and assumes feature independence. Naïve Bayes is widely used in text analysis for its speed and simplicity. When applied to YouTube comments from the 2024 debates, the algorithm demonstrated its effectiveness, especially with a balanced dataset through random oversampling. It achieved 85.155% accuracy, high precision, recall, and an AUC of 96.8% on an 80:20 data split. Its fast classification time (0.000998 seconds) makes it suitable for real-time sentiment analysis, validating its use for political events. Future applications may incorporate advanced techniques like BERT for more sophisticated analysis.

Keywords: BERT; Candidate Debates; Indonesian Presidential Election; Naïve Bayes; Sentiment Analysis; Random Oversampling; YouTube Comments.

1. Introduction

Political debates have been held for a long time, and the most important and well-remembered are the Lincoln-Douglas senatorial debate in 1858, the first televised presidential debate between Kennedy and Nixon in 1960, and between Mitterand and Giscard in 1974. Nowadays, debates are one of the most important political campaigns watched by millions of viewers. In debates, candidates can express their views on key policy issues and inform voters about their policy commitments, which can foster accountability pressures that discipline the behavior of the elected candidate[1]. Voters can learn from political debates, which influence their behavior in choosing more competent individuals. Moreover, voters tend to learn more about unfamiliar candidates than better-known ones[2]. There is even a tailor-made application that can visualize political debates so they can be understood by non-expert voters eventually[3].

The 2024 Indonesian Presidential Election is the fifth general election in Indonesia aimed at electing the President and Vice President of the Republic of Indonesia. This election determines the incumbent president and vice president for the 2024–2029 term. It took place simultaneously across Indonesia on February 14, 2024. This election serves as a political contest to elect a new president to replace the previous president, who retired after serving two terms and could not run again according to the constitution. The debate between the candidates for President and Vice President of the Republic of Indonesia is a significant moment in the general election process. This debate allows each candidate to present their vision and mission and allows the public to evaluate and assess the candidates' abilities and personalities.

Received: October, 30th 2024
Revised: December, 22nd 2024
Accepted: December, 25th 2024
Published: January, 1st 2025

With technological advancements and the increasing use of social media, platforms like YouTube have become primary media for the public to watch and comment on the debate.

Comments left by YouTube users can reflect public sentiment towards the presidential and vice-presidential candidates. Sentiment analysis of these comments can provide valuable insights into public views and opinions. However, challenges in analyzing YouTube comments include the large volume of data, the diversity of languages, and the variety of contexts and expressions used by users. Sentiment analysis involves evaluating public sentiment or opinions about a product, service, or individual, including political figures and popular celebrities.

The Naïve Bayes algorithm is a simple yet effective classification algorithm often used in text analysis and natural language processing[4]. Based on Bayes' Theorem, the algorithm assumes that each feature in the data is independent, making it fast and easy to implement. In sentiment analysis, Naïve Bayes can classify text into positive, negative, or neutral categories. Despite its simplicity, Naïve Bayes has proven effective in various text analysis applications, including product reviews, social media comments, and user feedback. It is a popular choice for research involving large and complex text data. For instance, a study analyzing sentiment towards the 2019 Indonesian presidential candidates on Twitter achieved an accuracy of 80.90% using Naïve Bayes with training data for each candidate[5]. The study also compared Naïve Bayes with Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN), finding that Naïve Bayes achieved the highest accuracy of 75.58%. Furthermore, research evaluating public approval of government application services using the Naïve Bayes Classifier method, based on reviews from the Google Play Store, produced a precision of 83%, an accuracy of 83%, a recall of 100%, and an F-measure of 90.7%[6]. Additionally, an analysis of game product sentiment on the e-commerce platform Shopee, using 1,000 reviews and the TF-IDF method combined with the Naïve Bayes Classifier, showed an accuracy of 80.22%[7].

However, in sentiment analysis, particularly when dealing with large and diverse text data such as YouTube comments, using the right technique is crucial for obtaining accurate results. Bidirectional Encoder Representations from Transformers (BERT) is one of the latest natural language models developed by Google. BERT's main advantage lies in its ability to understand the bidirectional context of text, meaning it considers both the preceding and following words in a sentence. This is essential for handling YouTube comments' complex and diverse language nuances. By using BERT for data labeling, the model can better grasp the true meaning of each comment, leading to more accurate and relevant sentiment analysis results. Although BERT can improve labeling accuracy, data imbalance remains a common challenge. Certain sentiments can be more dominant in YouTube comment data, causing bias in the sentiment analysis model. To address this issue, the Random Oversampling technique is used. Random Oversampling balances the distribution of sentiment classes by duplicating samples from underrepresented classes. This approach allows the sentiment analysis model to be trained on more balanced data, resulting in more objective predictions and reducing bias towards the dominant class. Combining BERT for labeling and Random Oversampling for data balancing offers a comprehensive approach to improving the quality of sentiment analysis, especially in the context of debates between the Presidential and Vice-Presidential candidates of the Republic of Indonesia.

This study aims to develop a more effective and precise method for analyzing public sentiment toward the Republic of Indonesia's 2024 Presidential and Vice-Presidential candidate debates, focusing on comments from YouTube. By combining the Naïve Bayes algorithm for processing, BERT for data labeling, and Random Oversampling for data balancing, this research intends to address challenges related to handling large data volumes and uneven sentiment distribution. The objective is to generate deeper and more accurate sentiment analysis, offering comprehensive insights into public viewpoints and opinions. The findings from this study are anticipated to serve as valuable input for stakeholders and the community, aiding in understanding public perceptions and facilitating better-informed decision-making processes.

2. Related Works

Sentiment analysis, also known as opinion mining, analyzes people's opinions, sentiments, attitudes, and emotions expressed in written language[8]. The growth of digital media has resulted in an explosion of textual data, making sentiment analysis increasingly relevant.

Recent research has focused on improving sentiment analysis methodologies by leveraging advanced machine learning (ML) and deep learning (DL) techniques[9]. Transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT) and its variants, have significantly improved understanding of context and nuances in text, outperforming traditional approaches like Naïve Bayes classifiers. Studies like [10] and [11] highlight the effectiveness of these models in various sentiment analysis tasks.

Sentiment analysis is increasingly being adapted to specific domains such as finance, healthcare, and social media, where domain-specific models are trained on tailored datasets to enhance accuracy and relevance. In finance, sentiment analysis employs data from news articles, social media, and income statements to predict stock market movements[12]. In healthcare, applications include analyzing patient reviews and social media discussions to gauge public health sentiment[13]. Government agencies are also leveraging advances in information technology to boost transparency and the speed of public services. For instance, the Ministry of Health of the Republic of Indonesia uses social media platforms like Twitter to disseminate various information. A study [14] involves classifying tweet topics and analyzing the sentiment of comments on tweets from the Ministry of Health, demonstrating the application of sentiment analysis in enhancing public sector communication.

In the swiftly advancing domain of machine learning, a range of models and techniques are employed to address various data processing tasks. The three key components, namely BERT (Bidirectional Encoder Representations from Transformers), random sampling, and the Naïve Bayes algorithm play a vital role in natural language processing (NLP) and data analysis. BERT has transformed NLP by enabling more sophisticated text comprehension through its transformer architecture, which analyzes words in the context of all other words in a sentence simultaneously rather than sequentially[15]. Since its inception, variations and enhancements such as RoBERTa[11] and ALBERT [16] have been developed, boosting performance in tasks like question answering, sentiment analysis, and named entity recognition. A study [17] provides a comprehensive overview of BERT's influence on NLP, emphasizing its superiority in managing contextual information compared to earlier models.

Random sampling is a crucial technique in data analysis, ensuring that a subset of data accurately reflects the overall population. This method is pivotal in training machine learning models and assessing their performance. Recent advancements have aimed at optimizing random sampling methods to manage large-scale datasets more effectively. For instance, a paper [18] presented an algorithm designed to enhance the speed and accuracy of sampling in big data environments. Additionally, Zhang et al.[19] investigated an adaptive sampling technique that dynamically adjusts the sample size according to data complexity and the model's specific needs. The issue of imbalanced data affects a wide range of applications, and despite numerous sophisticated sampling techniques to address this, the simple random oversampling (ROS) method remains a robust alternative. This method does not generate new data. It only replicates the data from the underrepresented class to match the size of the dominant class, which results in reduced diversity and overfitting[20]. This is reported to improve accuracy by 3% [21]. Paper [22] compared ROS to more advanced sampling algorithms through numerical experiments on multi-label data, revealing that ROS outperforms several advanced algorithms. ROS's computational efficiency and robust accuracy provide a valuable option for handling imbalanced data. These innovations collectively improve the efficiency and effectiveness of data analysis in managing extensive datasets.

The Naïve Bayes algorithm remains popular for text classification tasks due to its simplicity and effectiveness, even though it assumes feature independence. Recent research has focused on enhancing Naïve Bayes by integrating it with other techniques to address its limitations. Hybrid models combining Naïve Bayes with deep learning approaches have improved accuracy in various applications. A study [6] using data from the Google Play Store on comments for the Avocado Betawi application reported Naïve Bayes achieving 83% accuracy, 83% precision, 100% recall, and a 90.7% F1-score. In another study[7], user reviews of a game product on the e-commerce platform Shopee found Naïve Bayes achieving 80.22% accuracy, 0.80 precision, 0.60 recall, and a 0.69 F1-score. Similarly, an analysis of starred hotels based on comments from online booking applications showed Naïve Bayes achieving 76.20% accuracy, 70.57% precision, and 99.85% recall[23]. Moreover, the study[24] analyzes using the field of natural language processing with sentiment analysis science on the TikTok platform that is being developed. The content on the TikTok platform will contain comments made by fellow users. Then these comments are collected to carry out sentiment analysis using the

classification algorithm, namely Naïve Bayes. The results of this study are accurate, measured by metric evaluation, which produces 80.95%.

Comparative studies have been conducted to evaluate the performance of BERT, random sampling techniques, and Naïve Bayes across different tasks. BERT consistently outperforms traditional methods in NLP tasks due to its deep contextual understanding, while random sampling remains a robust method for ensuring data representativeness in various applications. Naïve Bayes, despite its simplicity, continues to be relevant, especially when combined with more advanced techniques. Studies [25], [26] provide a comparative analysis of these methods, highlighting their strengths and suitable application scenarios. Future research will likely focus on further integrating these techniques to leverage their strengths. For instance, enhancing BERT's performance with optimized random sampling methods or developing more sophisticated hybrid models involving Naïve Bayes. Additionally, addressing the computational efficiency of these models will be crucial as the scale of data continues to grow. In this paper, we implement Naïve Bayes for sentiment analysis using random oversampling with data taken from YouTube.

3. Proposed Method

The following study employs a comprehensive four-phase process to ensure robust and accurate sentiment analysis. This process includes dataset collection, preprocessing, modeling, and model evaluation, each critical to the study's success. The steps are illustrated in Figure 1, providing a clear visual representation of the methodology.

3.1. Collect Data

The initial phase involves gathering the relevant data required for the analysis. For this study, comments were extracted from YouTube videos. The data collection focused on comments from channels such as KompasTV, MetroTV, TVRI, SCTV, and TVOne, covering five debates per channel. In summary, comments from 25 videos were extracted, all about the 2024 presidential and vice-presidential debates for the public election. Data collection took place on February 25, 2024, with the intention of analyzing user polarization regarding the upcoming 2024 presidential and vice-presidential public election. The dataset comprises a total of 121,404 comments. We exclude comments that only have emoticons and no text. Hence, the remaining total is 108,867 comments, as summarized in Table 1.

Table 1. Number of Comments on Each Channel.

Channel	Number of Comments
KompasTV	20,093
TVRI	18,722
SCTV	1,701
TvOne	32,805
MetroTV	35,546
Total	108,867

3.2. Preprocessing

The data obtained in the previous stage undergoes meticulous preprocessing to ensure the comments are clean and relevant. Before further preprocessing, each comment is automatically labeled using BERT into three positive, negative, and neutral classes. We have compared the results of data balancing using the Indonesian BERT Base Sentiment Classifier [27] and Indonesian Sentiment [28]. Indonesian BERT Base Sentiment Classifier is a senti-ment-text-classification model derived from the pre-trained IndoBERT Base Model. Indonesian Sentiment is a fine-tuned version of IndoBERT Base Uncased, a BERT model trained on Indonesian text. This version has been specifically adapted to analyze the sentiment of Indonesian comments and reviews. We find that Indonesian BERT Base Sentiment Classifier gives the most typical data balancing results, and thus, we decide to utilize it.

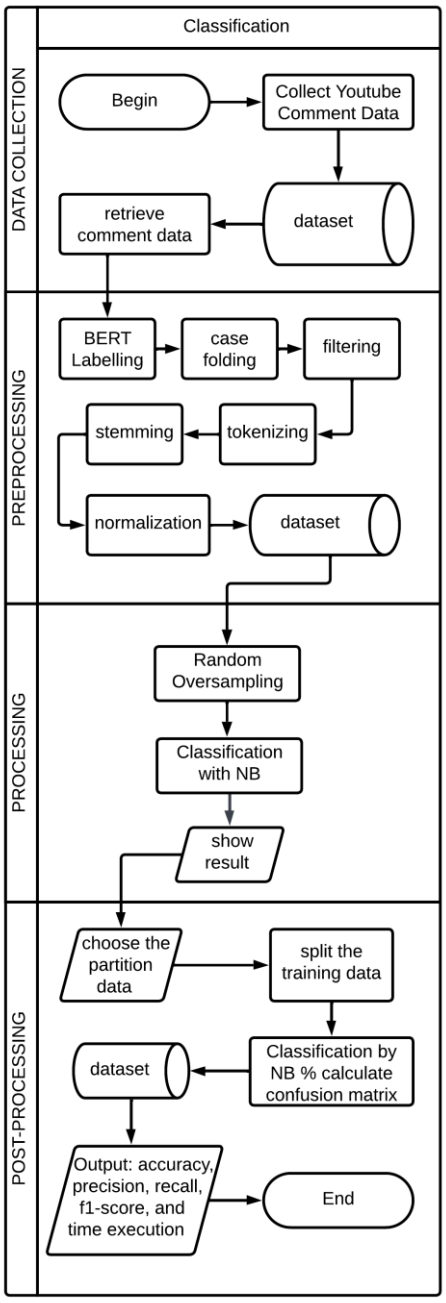

Figure 1. Proposed Method.

This stage exclusively uses comment data collected in the prior phase. The preprocessing steps are as follows:

1. Convert comments to all lowercase: Transforming the comments to lowercase ensures consistency and ease of analysis.
2. Remove URLs: Eliminate URLs starting with "http", "www", or "https".
3. Replace escape characters: Replace escape characters such as \t, \n, and \u with spaces.
4. Replace periods with spaces.
5. Handle non-ASCII characters: Replace non-ASCII characters with their closest ASCII equivalents.
6. Remove usernames and hashtags: starting with @ or #.
7. Remove numeric digits, punctuation, and single-letter words.
8. Strip leading and trailing whitespace: Remove any leading and trailing whitespace.
9. Perform tokenization, stemming, and normalization: Apply tokenization, stemming, and normalization to prepare the text for analysis.

3.3. Processing

In the processing stage of classification, as depicted in the flowchart, several critical steps are involved in preparing and analyzing the preprocessed YouTube comment data. These steps ensure that the data is appropriately balanced and classified. Here is a detailed explanation and elaboration of the processing stage:

3.3.1. Random Oversampling

This process addresses the issue of class imbalance in the dataset. Random oversampling involves duplicating samples from the minority class to ensure that each class has an equal number of samples. This step is crucial because imbalanced datasets can lead to biased models that perform well on the majority class but poorly on the minority class. The process includes identifying the classes with fewer samples (minority classes), randomly duplicating samples from these minority classes until the number of samples in each class is balanced, and then using the balanced dataset for training the classifier.

3.3.2. Classification with Naïve Bayes (NB)

This process classifies the comments into predefined categories using the Naïve Bayes method, a probabilistic classifier based on Bayes' Theorem. Naïve Bayes assumes that the features (in this case, tokens from comments) are independent. The process begins with the balanced dataset obtained from the random oversampling step. This dataset is used to train the Naïve Bayes classifier. The training involves calculating the probabilities of each feature given a class and using these probabilities to predict the class of new, unseen comments. Specifically, the classifier calculates the likelihood of a comment belonging to each class (positive, negative, or neutral) based on the presence of certain words or tokens. Using Bayes' Theorem, the Naïve Bayes formula is as follows:

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} \tag{1}$$

Here, $P(H|X)$ is the posterior probability of hypothesis H (the class) given the evidence X (the features or tokens in the comment). $P(X|H)$ is the likelihood, which is the probability of the evidence given that the hypothesis is true. $P(H)$ is the prior probability of the hypothesis, and $P(X)$ is the probability of the evidence.

For example, suppose we want to classify a new comment, "The debate was amazing and informative," into one of the three classes: positive, negative, or neutral. The classifier looks at the words "amazing" and "informative" and calculates the probability of this comment being in each class based on the training data. If the word "amazing" is frequently associated with positive comments and "informative" also appears often in positive comments, the classifier will likely predict the comment as positive.

3.4. Post-Processing

The trained classifier is then applied to the test or new data to predict the class labels. The results are evaluated to determine the model's accuracy, precision, recall, and F1 score, providing insights into its performance and areas for improvement. In the post-processing stage of classification, as depicted in the flowchart, several steps are involved in preparing and analyzing the preprocessed YouTube comment data to ensure it is appropriately balanced and classified.

Firstly, the dataset is divided into training and testing subsets. Partitioning the dataset is crucial for both training and evaluating the model. An appropriate partitioning strategy, such as a 60-40, 70-30, or 80-20 split, is selected based on the dataset size and specific research needs. After partitioning, each subset maintains the class balance achieved in the preprocessing stage. The training data is then further divided for cross-validation purposes. Splitting the training data helps create multiple training and validation sets used to train and validate the model iteratively. Implementing k-fold cross-validation or a similar method ensures robust model performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{5}$$

Next, the Naïve Bayes classifier is applied to the test data, and the confusion matrix is calculated to assess the classification results. The trained Naïve Bayes model is used to predict class labels, which are then compared with the actual labels. The confusion matrix provides a detailed breakdown of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Performance metrics such as accuracy (using Eq. (2)), precision (Eq. (3)), recall (Eq. (4)), and F1-score (Eq. (5)) are calculated based on the confusion matrix. The time taken for the classification process is also measured to evaluate the model's efficiency. The performance metrics and execution time are summarized and presented to highlight the model's strengths and areas for improvement. The calculated metrics provide a comprehensive overview of the classifier's performance and ability to generalize to new data. This information is crucial for validating the model's effectiveness and identifying potential improvements for future iterations.

The evaluation of the research model is conducted using the confusion matrix and several evaluation criteria, including accuracy, F1-score, precision, recall, and AUC (Area Under the Curve), which is a valuable metric for assessing the performance of sentiment classifiers, ranging from 0 to 1. An AUC of 0.5 indicates random guessing (no discrimination power), whereas an AUC of 1 indicates perfect classification (ideal discrimination).

4. Results and Discussion

4.1. Experimental Setup

In this experimental setup, we utilized a CPU, specifically the AMD Ryzen 5 6600H with a 3.3GHz Base Clock and 4.5GHz Boost Clock, along with 16GB of DDR5 RAM running at 4800MHz and a 512GB NVMe PCIe Gen 4 SSD. Additionally, the results from BERT labeling and Naïve Bayes (NB) classification were obtained using TF-IDF. After implementing random oversampling, the dataset now exhibits balanced sentiments, with each category (positive, negative, and neutral) containing 39,266 entries. This balancing process resulted in 117,798 data points, compared to the original 108,867 comments, before oversampling and preprocessing. After the preprocessing steps, the total number of data points was reduced to 78,416, with 39,266 negative sentiments, 29,167 positive sentiments, and 9,983 neutral sentiments, as shown in Figure 2.

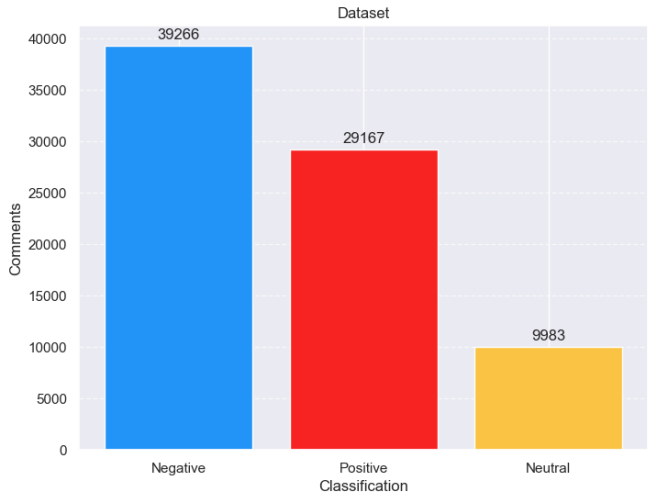

Figure 2. Dataset Distribution.

4.2. Experimental Results

This subsection presents experimental results obtained using the previously described setup. Figure 3 is a bar graph depicting sentiment classification results from a dataset with a random oversampling process. The graph displays the number of comments in three sentiment categories – negative, positive, and neutral – each with an equal number of comments, specifically 39,266. This indicates that the dataset has been processed using the random oversampling technique to achieve class balance. Such balance ensures that the machine learning model trained on this dataset will not be biased towards any particular sentiment category, enhancing its overall performance. The graph demonstrates a balanced distribution post-oversampling, which is crucial for improving model performance in sentiment classification by mitigating the bias that may arise from initial data imbalance.

Figure 4 presents a confusion matrix depicting the performance of the sentiment classification model in categorizing comments into three classes: negative, neutral, and positive. This matrix illustrates the percentage of model predictions compared to the actual labels. Rows represent the actual labels of the data, while columns represent the model predictions. For the negative category, 92.379% of comments that were actually negative were correctly classified as negative, 2.498% were incorrectly classified as neutral, and 5.122% were incorrectly classified as positive. In the neutral category, 11.357% of comments that were actually neutral were incorrectly classified as negative, 80.588% were correctly classified as neutral, and 8.055% were incorrectly classified as positive. Regarding the positive category, 10.072% of comments that were actually positive were incorrectly classified as negative, 5.220% were incorrectly classified as neutral, and 84.707% were correctly classified as positive.

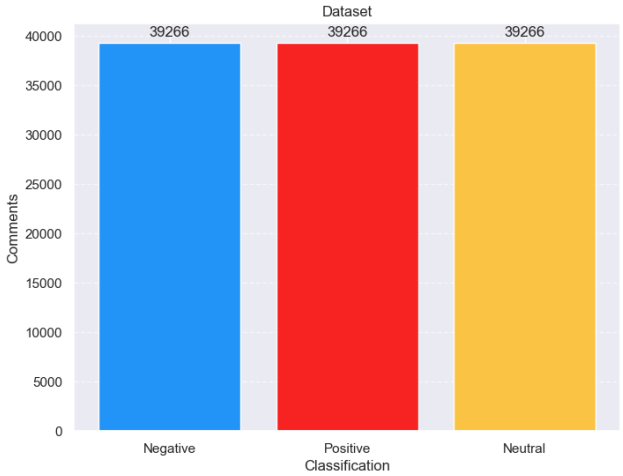

Figure 3. Dataset Distribution after Random Oversampling.

This confusion matrix indicates strong overall performance by the classification model, particularly in identifying negative and positive comments. The model achieves a high accuracy rate for negative comments (92.379%) and demonstrates good performance for positive comments (84.707%). However, there is room for improvement in classifying neutral comments. While the model shows proficiency in identifying neutral sentiments (80.588%), there is a notable error rate where neutral comments are misclassified as negative (11.357%). Errors between positive comments and other categories are relatively lower than those involving neutral comments.

Furthermore, we conducted multiple tests by partitioning the training data into three distinct subsets: 60% for data training and 40% for data testing (60:40), 70% for data training and 30% for data testing (70:30), and 80% for data training and 20% for data testing (80:20). Each partitioning scheme was evaluated over 10 iterations, and the summarized results are presented in the tables below. Table 2 presents the evaluation results for the 60:40 data partition, which yielded an accuracy of 84.17% and an Area Under the Curve (AUC) of 92.6%. Table 3 reports the evaluation outcomes for the 70:30 data partition, achieving an accuracy of 84.693% and an AUC of 96.4%. Moreover, Table 4 displays the evaluation metrics for the 80:20 data partition scheme, showing an accuracy of 85.155% and an AUC of 96.8%. This

approach comprehensively assesses the sentiment classification model across different training and testing data splits, demonstrating robust performance metrics across varying data partition ratios. The larger the partition in the training data, the better the accuracy value, and vice versa.

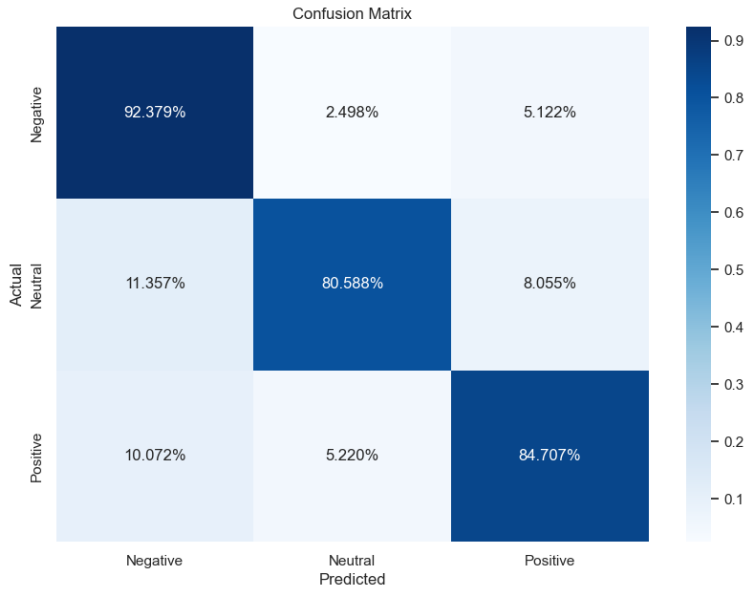

Figure 4. Confusion Matrix of Sentiment Classification Model Performance.

Table 2. Evaluation Results of the 60:40 Data Partition Scheme.

Iter.	Accuracy	F1-score	Precision	Recall	AUC
1	84.174%	0.841	0.848	0.842	0.963
2	84.370%	0.843	0.849	0.844	0.963
3	84.457%	0.844	0.851	0.845	0.963
4	84.119%	0.841	0.848	0.841	0.962
5	84.202%	0.842	0.849	0.842	0.962
6	83.809%	0.838	0.844	0.838	0.961
7	84.119%	0.841	0.848	0.841	0.961
8	84.058%	0.840	0.849	0.843	0.962
9	84.319%	0.843	0.849	0.843	0.962
10	84.177%	0.841	0.847	0.842	0.962
Rate	84.170%	0.841	0.848	0.842	0.962

Table 3. Evaluation Results of the 70:30 Data Partition Scheme.

Iter.	Accuracy	F1-score	Precision	Recall	AUC
1	84.720%	0.847	0.852	0.847	0.964
2	84.960%	0.849	0.854	0.850	0.965
3	84.813%	0.848	0.853	0.848	0.964
4	84.748%	0.847	0.853	0.847	0.964
5	84.247%	0.842	0.848	0.842	0.962
6	83.870%	0.848	0.854	0.849	0.964
7	84.720%	0.847	0.853	0.847	0.964
8	84.539%	0.845	0.850	0.845	0.963
9	84.757%	0.847	0.853	0.848	0.964
10	84.556%	0.845	0.850	0.846	0.993
Rate	84.693%	0.847	0.849	0.848	0.964

Table 4. Evaluation Results of the 80:20 Data Partition Scheme.

Iter.	Accuracy	F1-score	Precision	Recall	AUC
1	84.936%	0.849	0.853	0.849	0.965
2	85.093%	0.851	0.855	0.851	0.966
3	85.556%	0.855	0.860	0.856	0.967
4	85.123%	0.851	0.856	0.851	0.965
5	84.733%	0.847	0.852	0.847	0.964
6	85.038%	0.850	0.855	0.850	0.966
7	85.208%	0.852	0.856	0.852	0.966
8	85.314%	0.853	0.857	0.853	0.966
9	85.081%	0.851	0.856	0.851	0.966
10	85.471%	0.855	0.860	0.855	0.996
Rate	85.155%	0.849	0.850	0.851	0.968

Figure 5 compares various evaluation metrics for the classification model using three different data partition schemes. This graph evaluates the model's consistency and performance using accuracy, F1-score, precision, recall, and AUC (Area Under the Curve). The results demonstrate that the classification model performs consistently across all three data partition schemes, with all evaluation metrics showing stable values and slight improvements. Accuracy, F1-score, precision, and recall all range from 0.84 to 0.85, indicating balanced performance in sentiment analysis across the three data partition trials. The high AUC values, ranging from 0.98 to 0.99, indicate that the model can distinguish between classes. Therefore, the model exhibits stable and robust performance across multiple trials, which is crucial for its validity and reliability.

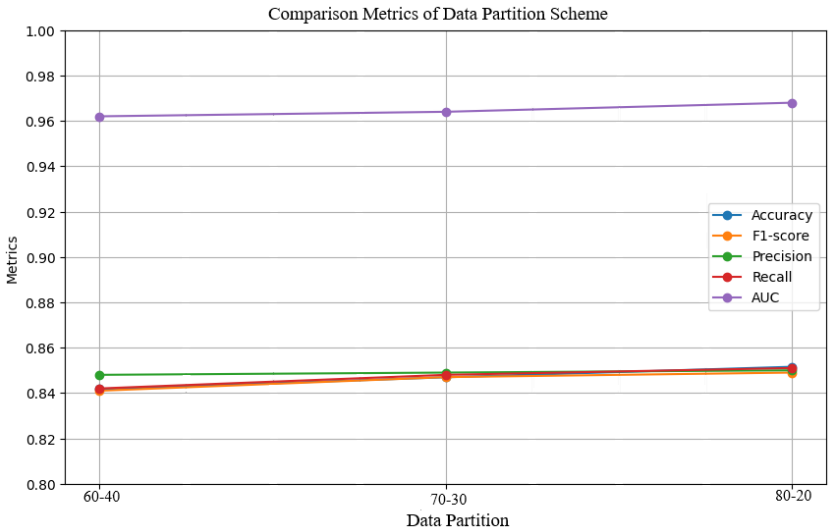

Figure 5. Comparison of Metrics Across Three Data Partition Schemes.

Table 5 shows the accuracy comparison with and without Random Oversampling. The accuracy with Random Oversampling is better because addressing class imbalance helps the model to gain a better understanding of the minority class, thus leading to improved performance and accuracy on the test set.

Table 5. Accuracy Comparison with and without Random Oversampling.

Data Partition	Without Random Oversampling	With Random Oversampling
60:40	67.10%	84.170%
70:30	67.30%	84.693%
80:20	67.66%	85.155%

In addition, we conducted a single test by inputting comments into the model to classify new comment data with the newly trained learning model. The results of this single test can be seen in Figure 6. This figure displays both the classification results and the execution time, which was recorded at 0.000998 ~ 0.001 seconds. This rapid execution time demonstrates the efficiency of the model in processing and classifying new data, highlighting its potential for real-time sentiment analysis applications.

Figure 6. Single Test Result.

5. Conclusions

Based on the development process and testing results, the following conclusions about the Naïve Bayes algorithm can be drawn. First, applying the Naïve Bayes algorithm in the context of the 2024-2029 Presidential and Vice-Presidential debates in Indonesia, assessing the algorithm's effectiveness in analyzing and classifying the debate results. In a series of tests using a complete debate dataset, the Naïve Bayes algorithm was evaluated over 10 iterations. The highest accuracy achieved by Naïve Bayes was 85.424% on an 80:20 data partition. The highest precision and recall values for Naïve Bayes were obtained from the complete debate dataset on an 80:20 data partition, with precision reaching 0.86 and recall of 0.855.

Recommendations for future application development include The experimental results indicating that applying the Naïve Bayes algorithm for sentiment classification on YouTube comments related to the 2024-2029 Indonesian Presidential and Vice Presidential debates is effective and reliable. The data, preprocessed and balanced through random oversampling, ensured equal representation of positive, negative, and neutral sentiments, with each category containing 39,266 comments. This balanced dataset improved the model's performance, as evidenced by the high accuracy, precision, recall, and F1 score across different data partition schemes (60:40, 70:30, and 80:20) with 10 iterations. The highest accuracy achieved by Naïve Bayes was 85.155%, and an AUC of 96.8% on an 80:20 data partition. The highest precision and recall values for Naïve Bayes were obtained from the complete debate dataset on an 80:20 data partition, with precision reaching 0.86 and recall of 0.855.

Additionally, the rapid execution time of 0.000998 seconds for classifying new comments highlights the model's efficiency, making it suitable for real-time sentiment analysis applications. The consistent performance across multiple trials and partitioning strategies validates the model's effectiveness and underscores its potential for accurately analyzing and classifying sentiment in large-scale datasets. These findings demonstrate the Naïve Bayes algorithm's capability to provide valuable public opinion insights during significant political events. For future work, incorporating a feature that allows crawling based on inputted YouTube video links containing new debates from Presidential and Vice-Presidential candidates is recommended.

Author Contributions: Conceptualization: Apriandy Angdresey and Lanny Sitanayah; methodology: Apriandy Angdresey and Ignatius Lucky Henokh Tangka; software: Ignatius Lucky Henokh Tangka; validation: Apriandy Angdresey, Lanny Sitanayah, and Ignatius Lucky Henokh Tangka; formal analysis: Apriandy Angdresey and Ignatius Lucky Henokh Tangka; investigation: Ignatius Lucky Henokh Tangka; resources: Ignatius Lucky Henokh Tangka; data curation: Ignatius Lucky Henokh Tangka; writing—original draft preparation: Apriandy Angdresey, Lanny Sitanayah, Ignatius Lucky Henokh Tangka; writing—review and editing: Apriandy Angdresey and Lanny Sitanayah; visualization: Ignatius Lucky Henokh Tangka; supervision: Apriandy Angdresey and Lanny Sitanayah; project administration: Apriandy Angdresey and Lanny Sitanayah; funding acquisition: None.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] K. Bidwell, K. Casey, and R. Glennerster, "Debates: Voting and Expenditure Responses to Political Communication," *J. Polit. Econ.*, vol. 128, no. 8, 2020.

[2] T. M. Holbrook, "Political Learning from Presidential Debates," *Polit. Behav.*, vol. 21, no. 1, pp. 67–89, 1999, doi: 10.1023/A:1023348513570.

[3] L. South, M. Schwab, N. Beauchamp, L. Wang, J. Wihbey, and M. A. Borkin, "DebateVis: Visualizing Political Debates for Non-Expert Users," in *2020 IEEE Visualization Conference (VIS)*, Oct. 2020, pp. 241–245. doi: 10.1109/VIS47514.2020.00055.

[4] P. P. Surya and B. Subbulakshmi, "Sentimental Analysis using Naive Bayes Classifier," in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, Mar. 2019, pp. 1–5. doi: 10.1109/ViTECoN.2019.8899618.

[5] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Oct. 2019, pp. 1–5. doi: 10.1109/ICIC47613.2019.8985884.

[6] W. Purbaratri, H. D. Purnomo, D. Manongga, I. Setyawan, and H. Hendry, "Sentiment Analysis of e-Government Service Using the Naive Bayes Algorithm," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 23, no. 2, pp. 441–452, Mar. 2024, doi: 10.30812/matrik.v23i2.3272.

[7] R. Kosasih and A. Alberto, "Sentiment analysis of game product on shopee using the TF-IDF method and naive bayes classifier," *Ilk. J. Ilm.*, vol. 13, no. 2, pp. 101–109, Aug. 2021, doi: 10.33096/ilkom.v13i2.721.101-109.

[8] H. Huang, A. A. Zavareh, and M. B. Mustafa, "Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions," *IEEE Access*, vol. 11, pp. 90367–90382, 2023, doi: 10.1109/ACCESS.2023.3307308.

[9] D. R. I. M. D. R. I. M. Setiadi, D. Marutho, and N. A. Setiyanto, "Comprehensive Exploration of Machine and Deep Learning Classification Methods for Aspect-Based Sentiment Analysis with Latent Dirichlet Allocation Topic Modeling," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 12–22, May 2024, doi: 10.62411/faith.2024-3.

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[11] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv*. Jul. 26, 2019.

[12] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, pp. 1–19, Feb. 2009, doi: 10.1145/1462198.1462204.

[13] C. R. Machuca, C. Gallardo, and R. M. Toasa, "Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach," *J. Phys. Conf. Ser.*, vol. 1828, no. 1, p. 012104, Feb. 2021, doi: 10.1088/1742-6596/1828/1/012104.

[14] A. Angdresey, I. Y. Kairupan, and K. G. Emor, "Classification and Sentiment Analysis on Tweets of the Ministry of Health Republic of Indonesia," in *2022 Seventh International Conference on Informatics and Computing (ICIC)*, Dec. 2022, pp. 1–6. doi: 10.1109/ICIC56845.2022.10007008.

[15] A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *J. Big Data*, vol. 10, no. 1, p. 110, Jun. 2023, doi: 10.1186/s40537-023-00781-w.

[16] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations," in *Proceeding of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[17] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 842–866, Dec. 2020, doi: 10.1162/tacl_a_00349.

[18] P. Nishad and S. Sankar, "Efficient Random Sampling Statistical Method to Improve Big Data Compression Ratio and Pattern Matching Techniques for Compressed Data," *Int. J. Comput. Sci. Inf. Secur.*, vol. 14, no. 6, pp. 179–184, May 2016.

[19] Z. Zhang, Y. Tang, P. Zhang, P. Zhang, D. Zhang, and P. Wang, "An Adaptive Drilling Sampling Method and Evaluation Model for Large-Scale Streaming Data," in *Web Information Systems Engineering (WISE)*, Springer, 2023, pp. 813–825. doi: 10.1007/978-981-99-7254-8_63.

[20] P. Sundarreson and S. Kumarapathirage, "SentiGEN: Synthetic Data Generator for Sentiment Analysis," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 461–477, Apr. 2024, doi: 10.62411/jcta.10480.

[21] K. K. Yusuf, E. Ogbuju, T. Abiodun, and F. Oladipo, "A Technical Review of the State-of-the-Art Methods in Aspect-Based Sentiment Analysis," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 287–298, Feb. 2024, doi: 10.62411/jcta.9999.

[22] F. Kamalov, H.-H. Leung, and A. K. Cherukuri, "Keep it simple: random oversampling for imbalanced data," in *2023 Advances in Science and Engineering Technology International Conferences (ASET)*, Feb. 2023, pp. 1–4. doi: 10.1109/ASET56582.2023.10180891.

[23] J. P. Matrutty, A. M. Adrian, and A. Angdresey, "Sentiment Analysis of Visitor Reviews on Star Hotels in Manado City," *J. Inf. Technol. Comput. Sci.*, vol. 8, no. 1, pp. 21–32, Apr. 2023, doi: 10.25126/jitecs.202381403.

[24] Y. Y. Lase, A. R. Lubis, F. Elyza, and S. A. Syafli, "Mental Health Sentiment Analysis on Social Media TikTok with the Naïve Bayes Algorithm," in *2023 6th International Conference of Computer and Informatics Engineering (IC2IE)*, Sep. 2023, pp. 186–191. doi: 10.1109/IC2IE60547.2023.10331126.

[25] P. J. B. Pajila, B. G. Sheena, A. Gayathri, J. Aswini, M. Nalini, and S. S. R, "A Comprehensive Survey on Naive Bayes Algorithm: Advantages, Limitations and Applications," in *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, Sep. 2023, pp. 1228–1234. doi: 10.1109/ICOSEC58147.2023.10276274.

[26] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey," *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: 10.3390/info10040150.

[27] M. H. Lidero, "mdhugol/indonesia-bert-sentiment-classification," *Hugging Face*. https://huggingface.co/mdhugol/indonesia-bert-sentiment-classification (accessed Mar. 15, 2024).
[28] T. D. Purnomo, "taufiqdp/indonesian-sentiment," *Hugging Face*. https://huggingface.co/taufiqdp/indonesian-sentiment (accessed Mar. 15, 2024).