AMAL INSANI foundation
The 1 st AISETH 2024

E-ISSN 2962-0226
PROSIDING SEMINAR INTERNASIONAL
Business Digitalization to Increase People's Income
(Learning from The Advancement of Digitalization in India)
Serang, Banten. June, 9. 2024
https://prosiding.amalinsani.org/

Public Sentiment Analysis on the 2024 Presidential Election Using Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) On Social Media Data

Asro[1], Nur Azizah[2], Sudaryono [3]

[1,2,3] *Universitas Raharja, Tangerang*

Coresponding Email: asro@raharja.info

ABSTRACT

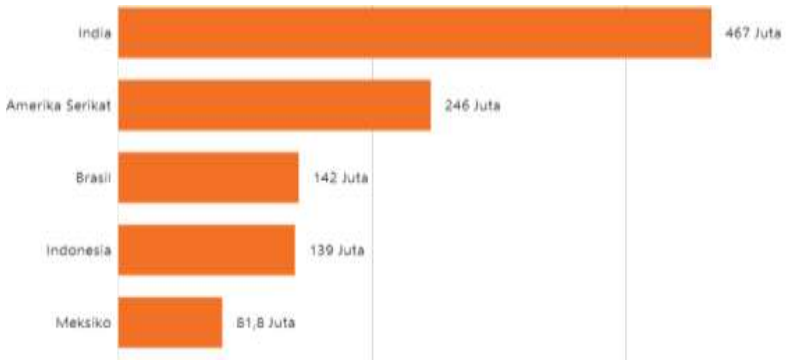
This study aims to evaluate the effectiveness of the Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) in analyzing public sentiment from YouTube comments related to the 2024 Indonesian Presidential Election. A total of 1,800 comments, collected from November 2023 to March 2024, were analyzed to test these models. The results show that SVM, with the highest accuracy of 76.33% and precision and F1-Score of 75.29% and 72.67% on the 10% test data, outperformed NBC, which recorded a highest accuracy of 72.19% under similar conditions. These findings highlight the importance of using more sophisticated methods in sentiment analysis to understand the complex and diverse dynamics of public opinion. This study provides valuable insights for stakeholders in developing effective communication strategies and offers a foundation for advancing sentiment analysis methodologies in political contexts.

Keywords: *2024 Presidential Election, Sentiment Analysis, Naive Bayes Classifier, Support Vector Machine, YouTube, Social Media.*

INTRODUCTION

In today's digital era, social media has become one of the primary communication platforms, especially for public interaction and opinion dissemination regarding political issues and elections (Manullang, Prianto, and Harani 2023). In Indonesia, internet usage has reached significant levels, with over 63 million people actively participating in the online world. According to data from Kominfo, about 95% of these internet users access social networks, making Indonesia one of the countries with very high social media activity. YouTube, as a video-based social media platform, plays an important role in political discourse by allowing two-way interactions between content creators and viewers through its comment feature, often becoming a space for controversial comments (Salsabila and Budiyanto 2023). This provides a venue not only for users to watch content but also to express opinions, support, and criticism on political issues and figures.

Figure 1.1 Visualization places Indonesia in fourth place in the world



In the context of the 2024 Presidential Election (Pilpres) in Indonesia, public opinion manifested in YouTube comments on presidential candidates reflects a dynamic and diverse political landscape. YouTube users in Indonesia, according to Databoks Katadata, reached 139 million by early 2023, placing Indonesia fourth in the world. This highlights the enormous potential of YouTube as a data source for public sentiment analysis, especially concerning the 2024 Presidential Election. This analysis is crucial to understand voter preferences, campaign dynamics, and the potential influence of public opinion on election outcomes. This study focuses on public sentiment analysis towards the three main presidential candidate pairs in the 2024 Presidential Election, utilizing text mining techniques with the application of two main methodologies: Naive Bayes

Classifier (NBC) and Support Vector Machine (SVM) (Guru et al. 2024; Rosyida, Putro, and Wahyono 2024; Saputra et al. 2019) . NBC is chosen for its effectiveness in classifying text based on sentiment polarity (positive, negative, or neutral) and its ease of implementation, making it an ideal tool for processing and understanding large unstructured text data from YouTube comments. Meanwhile, SVM is used to enhance classification accuracy through a more complex model capable of handling subtle nuances and high-dimensional text data. This technique provides advantages in identifying optimal decision boundaries between negative or positive sentiment categories. The data for this analysis is collected through crawling techniques from YouTube comments related to campaign activities and news about the three presidential candidate pairs: H. Anies Rasyid Baswedan and H. A. Muhaimin Iskandar; H. Prabowo Subianto and Gibran Rakabuming Raka; and H. Ganjar Pranowo and Prof. Dr. H. M. Mahfud MD. This sentiment analysis aims to provide insights into public sentiment, which in turn reflects the dynamics of public opinion in the context of the 2024 Presidential Election.

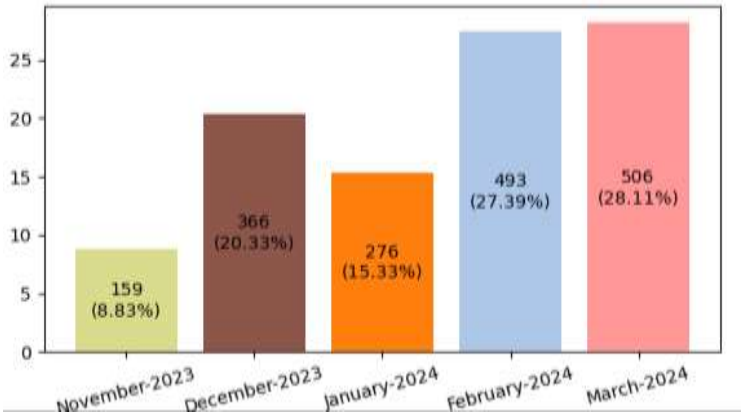
Figure 1.2 Three Pairs of Presidential Candidates for 2024

In the context of the 2024 Indonesian Presidential Election, social media, particularly YouTube, has transformed into a vital arena where citizens actively express their support, criticism, and opinions about the presidential candidates (Silalahi and Hartanto 2023). This digital interaction not only enriches public discourse but also offers opportunities to analyze untapped big data, such as YouTube comments, which can serve as a crucial information source for understanding public sentiment. Data crawling techniques have been applied in this study to collect thousands of comments from YouTube videos related to the three main presidential candidate pairs. The integration of

Google API and Collab Notebook allows the collection of a rich text corpus, which is then stored in Excel (XLSX) format to facilitate the analysis process.

The Naive Bayes Classifier (NBC) method is chosen as the primary analytical tool due to its effectiveness in classifying text based on sentiment and its ease of implementation. Additionally, Support Vector Machine (SVM) is integrated to improve classification accuracy by leveraging its capability to model complex boundaries between sentiment categories. By combining NBC and SVM, this study aims to achieve more accurate and in-depth results on public sentiment polarity. The analyzed data reflects the volume and diversity of public opinion, with a total of **1800** *YouTube comment* data collected over several months, showing the distribution of time and intensity of public discussion regarding the presidential candidates.

Figure 1.3 visualization of data collection



The bar chart displays the distribution of 1,800 YouTube comments related to the 2024 Indonesian Presidential Election, collected from November 2023 to March 2024. The data shows a gradual increase in the number of comments over time, starting from 159 comments in November 2023 (8.83%) and peaking at 506 comments in March 2024 (28.11%), indicating growing public engagement as the election approaches. Public sentiment analysis towards these candidates aims to dissect and map the existing sentiments in YouTube comments, providing insights into how public opinion is formed and changes over time. This research is expected to reveal the complex and diverse

dynamics of public sentiment and enrich our understanding of how the digital space is utilized by citizens to participate in the democratic process. Considering the importance of the applied methods, the research title is adjusted to "Public Sentiment Analysis on the 2024 Presidential Election Application of Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) Methods on Social Media Data." The objective of this research is not only to identify the diverse public sentiments towards the presidential candidates but also to offer new insights into the utilization of data analysis technology in understanding public opinion in significant political contexts such as the 2024 Presidential Election. Through innovative methodology and comprehensive data analysis, this study aims to make a significant contribution to the study of political communication and social media analysis. It highlights the importance of sentiment analysis in interpreting the dynamics of public opinion in the digital era, utilizing advanced techniques in natural language processing and machine learning. By combining NBC and SVM, this research offers a more robust approach to sentiment classification, enhancing the accuracy and depth of analysis in understanding public views and feelings towards the 2024 Presidential Election.

LITERATURE REVIEW

The literature review for this research is grounded in references collected from various internet sources, providing a robust theoretical foundation for the study. Key concepts such as data mining, sentiment analysis, Naive Bayes Classifier (NBC), and Support Vector Machine (SVM) are the primary focus (Manullang et al. 2023), (Salsabila and Budiyanto 2023). This approach allows for an in-depth examination of how these methods are applied in text data analysis, particularly from social media, to determine public sentiment on various topics.

Sentiment Analysis

Sentiment analysis is a computational process that enables the identification and categorization of opinions within text data to determine whether the expressed attitude is positive, negative, or neutral. Research by (Manullang et al. 2023) shows that this technique can be utilized to understand public sentiment regarding presidential

candidates. The categorization of sentiments as negative, positive, or neutral demonstrates how sentiment analysis can be used to evaluate public opinion in YouTube comments, highlighting the importance of this method in social and commercial research, as well as its application in political contexts.

Naive Bayes Classifier (NBC)

Naive Bayes Classifier (NBC) is an effective classification method that utilizes Bayes' Theorem with the assumption that each predictor in the model is independent. NBC has proven to be a powerful tool in text sentiment classification, as demonstrated in research applying NBC to classify sentiments on YouTube or Twitter related to presidential candidates (Salsabila and Budiyanto 2023). This research outlines the basic principles of NBC, its advantages in processing large text datasets, and its limitations in text classification contexts.

The Bayes' Theorem is generally represented by the following equation:

$$P\ (H|x) = \frac{\mathrm{P(x|H)P(H)}}{\mathrm{P(x)}}$$

Where:

- x : Data with an unknown class
- H : Hypothesis that data xxx belongs to a specific class
- $P(H|x)$: Probability of hypothesis H given data x (posterior probability)
- $P(x|H)$: Probability of data xxx given hypothesis H
- $P(x)$: Probability of data x

Naive Bayes Classifier (NBC) and Google Collab

In this research, the use of Google Collaboratory, or "Collab," has been pivotal in facilitating the implementation of NBC. Collab provides a cloud-based platform that allows Python code execution directly from the browser, making it ideal for tasks requiring intensive computation like machine learning and sentiment analysis. With Collab, researchers can easily write, run, and share data analysis code without worrying about complex development environment configurations. The integration of Collab in the research process facilitates the efficient collection of YouTube comment data through crawling techniques, where scripts are created and executed to gather data efficiently.

The use of Collab showcases how cloud computing technology can support sentiment analysis research by providing flexible and easily accessible computational resources. This enables researchers to focus on data analysis and result interpretation without being burdened by IT infrastructure issues. By combining NBC and Collab, this research successfully applies a robust sentiment analysis technique to categorize public opinions in YouTube comments related to the 2024 Presidential Election. This process yields valuable insights into public sentiment, highlighting the strengths and weaknesses of NBC in the context of large and dynamic social media text data.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning model that analyzes data and recognizes patterns, used for classification and regression. For instance, research by (Rosyida et al. 2024) explores the mechanism of SVM in handling large and complex data related to sentiment analysis of the 2024 Presidential Election, demonstrating its effectiveness and accuracy. The advantage of SVM is its ability to identify different hyperplanes to maximize the margin between classes. Therefore, this research is essential to understand the effectiveness of SVM in the case of sentiment classification of YouTube comments about politics. This study is expected to benefit political enthusiasts and subsequent research in understanding public sentiment towards presidential candidates based on YouTube comment analysis using SVM.

Data Preprocessing

Steps in data preprocessing such as case folding, cleaning, tokenizing, stopwords removal, normalization, and stemming are essential for preparing text data before analysis. Research by Pramana Yoga Saputra et al.(Saputra et al. 2019) illustrates the application of preprocessing techniques in their study on YouTube video comments related to government services, clarifying the importance of these steps in enhancing data analysis quality. These preprocessing stages are as follows:

a) Case Folding: Converting all characters in the text to lower or upper case.

b) Cleansing: Removing unnecessary words to reduce noise, such as URLs, hashtags, usernames, emails, and punctuation marks.

c) Tokenizing: Separating a sequence of words into individual tokens based on spaces or special characters.

d) Stopword Removal: Eliminating frequently occurring words that are common and less relevant to the text.

e) Stemming: Reducing words to their root form by removing prefixes or suffixes.

Measuring Effectiveness and Accuracy

Measuring the effectiveness and accuracy in classifying data using Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) is crucial to understand how well these models perform. This study compares the accuracy produced by both methods. According to Meiriza et al. [13], Naive Bayes is proven to have high accuracy and speed when used in large databases with extensive data. For instance, research by (Guru et al. 2024) compares the effectiveness of NBC and SVM in sentiment analysis of Twitter regarding the postponement of the 2024 Election, providing important insights into metrics such as accuracy, precision, recall, and F1-score.

Social Media as a Data Source

Social media, especially YouTube, is considered a rich data source for sentiment analysis. Research by(Alexander, Bria, and Witanti 2023). and (Tohidi, Perdana Herdiansyah, and Wahyudin 2024). demonstrates how social media is utilized to assess public sentiment towards presidential candidates, highlighting the relevance and potential of social media for data analysis.

Limitations in Sentiment Analysis

Sentiment analysis faces challenges, including irony, sarcasm, and changing contexts. One of the shortcomings of other studies with similar cases is the inability to adjust data in each class to be equal(Azzawagama Firdaus, Yudhana, and Riadi 2024) and (Imran, Nasirudin Karim, and Isna Ningsih 2024)on hoax news classification shows how sentiment analysis can be influenced by various factors, such as the language used and the specific context of each news piece. This emphasizes the importance of careful data preprocessing and the appropriate selection of analysis methods to overcome these limitations.

METHODS

Research Type

This research employs a quantitative approach focused on sentiment analysis to categorize YouTube comments related to the 2024 Presidential Election. The chosen data analysis methods are Naive Bayes Classifier (NBC) and Support Vector Machine (SVM), executed through Google Collaboratory (Google Collab) for ease and efficiency in the analysis process.

Requirements Analysis

To support the activities of this research, specific hardware and software specifications are required. According to the study (Robi Padri, Asro, and Indra 2023), a software and hardware requirements analysis is essential to run projects that researchers are developing using the necessary system design. This analysis utilizes a Lenovo laptop with the following specifications:

Tabel 3.1 Requirements Needs

Hardware	**Specifications**
Lenovo Thinkpad X240	
Processor	13th Gen Intel® Core™ i7-1355U 1.70 GHz
RAM	40.0 GB (39.7 GB usable)
Storage	125 SSD and 500 HDD
Monitor	12.5 inch
Software	
Operating System	Windows 11 Pro 23H2
Tools	Google Collaboratory
Programming Language	Python 3.9.13

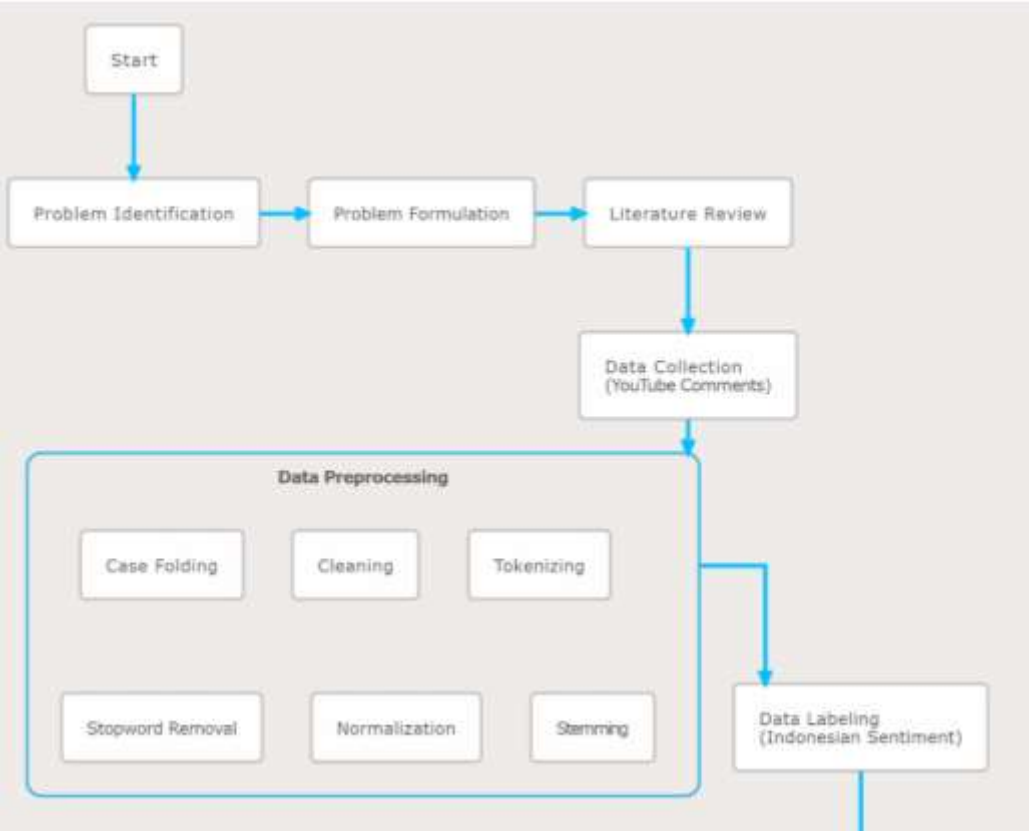
These requirements are based on the need to run data analysis processes that require high memory capacity and processing speed. The primary software used in this study is Google Collaboratory (Google Collab), a cloud-based platform that enables efficient Python code execution and data analysis. Google Collab is chosen for its ease in providing a flexible computing environment accessible from various devices, and its

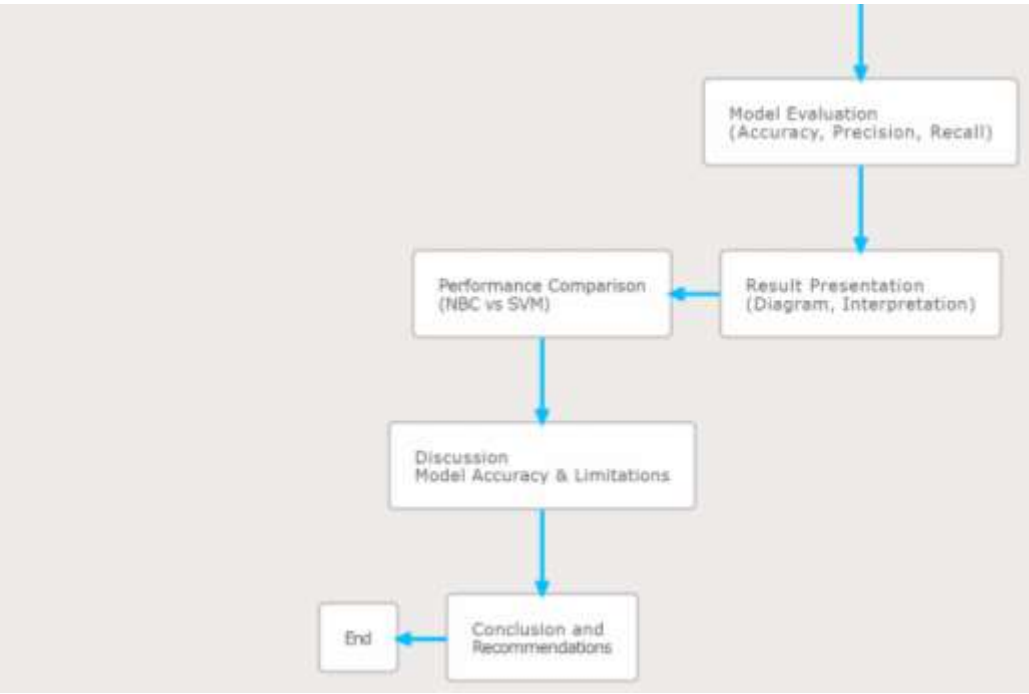
ability to handle large datasets necessary for public sentiment analysis regarding the 2024 Presidential Election.

Research Stages

The research stages are depicted in the diagram shown in Figure 3.1. According to (Ansori and Holle 2022), this research uses several machine learning algorithms to compare performance in sentiment classification analysis. The algorithms used include Support Vector Machine, Naive Bayes Classifier, and Logistic Regression. The following is Figure 3.1 about the research methods. The research process includes collecting YouTube comments related to the 2024 Presidential Election using Google Collaboratory for data crawling, followed by data preprocessing including cleaning and tokenization. Sentiment analysis is then performed using Naive Bayes Classifier and Support Vector Machine, model evaluation through accuracy, precision, and recall measurement, and results presentation and comparison of both methods' performance. The research concludes with conclusions and recommendations based on the analysis results.

Figure 3.1 Research Stages



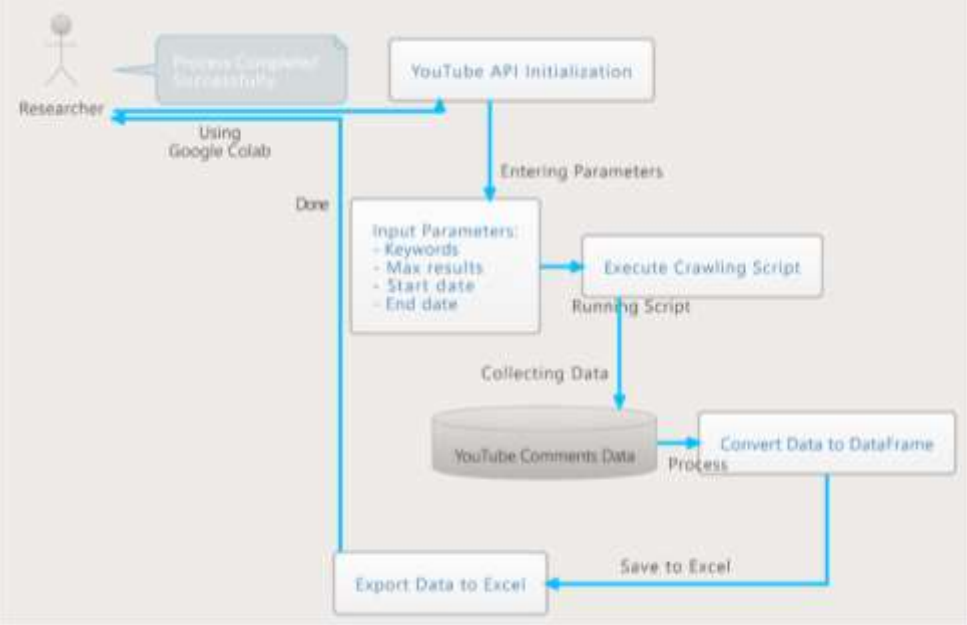


Data Collection

The research begins with collecting YouTube comments related to the 2024 Presidential Election. The researcher uses Google Collaboratory (Collab) to run Python scripts for data crawling. Examples of collected data include comments on campaign videos from three presidential candidate pairs, encompassing various opinions, support, and criticism from the public. This process yields a dataset containing thousands of comments as raw data for analysis.

Data Collection Process Explanation:

Figure 3.2 Crawling data



Data Collection: Data was collected using Python scripts run on Google Collaboratory. These scripts utilized the YouTube Data API to retrieve comments from videos related to the three presidential candidate pairs. This process ensures that the dataset includes a variety of public opinions expressed during the campaign period.

Steps:

a) YouTube API Initialization: Using google Api client. discovery to access comment data.

b) Input Parameters: Entering search parameters such as keywords, maximum results, start date, and end date.

c) Execution of Crawling Script: Running the Python script to crawl data from YouTube based on the entered parameters.

d) Collecting YouTube Comments: The script retrieves comments from videos matching the search parameters.

e) Conversion to DataFrame: The collected data is converted into a DataFrame format for easy analysis.

f) Export to Excel: The converted data is exported to Excel format for documentation and further analysis.

Data Preprocessing

Following data collection, the next critical step is data preprocessing. Preprocessing aims to clean the data of unnecessary or disruptive elements such as URLs, emojis, and to normalize and prepare the text for sentiment analysis. According to (Fadlila Nurwanda et al. 2023) , the data preprocessing level involves case folding, tokenizing, removing stopwords, and lemmatizing. These tasks are performed using the Python programming language, Google Collaboratory software, and libraries such as pandas, Sastrawi, and nltk.

Preprocessing Steps:

a) Case Folding: Converting all text to lowercase to ensure uniformity and prevent differences between uppercase and lowercase letters.
b) Cleaning & Tokenizing: Removing unnecessary URLs, emojis, special characters, and breaking the text into a list of words or 'tokens'.
c) Stopwords Removal: Eliminating common words that do not contribute to sentiment analysis, such as prepositions and conjunctions, using a 'stopword.txt' list.
d) Normalization: Standardizing non-standard words to their correct form according to a list contained in 'normalisasi.xlsx'.
e) Stemming: Reducing words to their root form using the Sastrawi library.

Preprocessing is critical for cleaning data of elements that could disrupt the analysis. This enhances the quality and accuracy of results when applying Naive Bayes Classifier (NBC) or Support Vector Machine (SVM) to analyze public sentiment regarding the 2024 Presidential Election.

Tabel 3.1 Sample Preprocessing data

Preprocessing Step	Before Example	After Example
Case Folding	"Saya akan MEMILIH di Pilpres 2024!"	"saya akan memilih di pilpres 2024!"
Cleaning & Tokenizing	"Dengarkan debat kandidat Pilpres di radio kita! 😀 #DebatPilpres"	["Dengarkan", "debat", "kandidat", "Pilpres", "di", "radio", "kita"]
Stopwords Removal	["saya", "akan", "memilih", "di", "pilpres", "2024"]	["pilih", "pilpres"]

Normalization	"ga sabar nunggu hasil pilpres, pdhl udh vote kemaren."	"tidak sabar menunggu hasil pilpres, padahal sudah vote kemarin."
Stemming	"Pemilihannya akan sangat menentukan masa depan negara."	["pilih", "akan", "tentu", "masa", "depan", "negara"]

After data has been cleaned and normalized through preprocessing, we proceed to feature extraction using the TF-IDF method. This method will generate a numerical representation of the text used to train the classification model. TF-IDF helps assess the importance of a word relative to the document within the dataset based on its frequency in the document and across the dataset. This process is crucial for identifying key words that significantly influence sentiment. According to (Madjid, Ratnawati, and Rahayudi 2023), the application of TF-IDF extends beyond weighting to classification by implementing algorithms such as Support Vector Machine (SVM) and Naive Bayes Classifier (NBC). These algorithms evaluate the effectiveness of key words in determining sentiment. Subsequently, the data undergoes a testing process, including validation with 10-fold cross-validation to ensure the classification models perform effectively under various conditions and are reliable for accurate sentiment analysis.

Data Labeling

After preprocessing, the next step is 'Labeling Data,' a crucial phase in preparing data for sentiment analysis. At this stage, each processed comment or text is labeled with a sentiment based on predefined criteria. Labeling is done using an Indonesian sentiment lexicon, a collection of words annotated with positive, negative, or neutral sentiment values.

Data Labeling Process:

a) Sentiment Lexicon Integration: We use a sentiment lexicon specifically developed for the Indonesian language. This lexicon contains a list of words with sentiment values based on their influence on emotions or opinions.

b) Applying Lexicon to Data: For each word in the preprocessed text, the system matches these words with entries in the lexicon. If a word is found in the lexicon, its sentiment value is used to calculate the overall sentiment score of the text.

c) Determining Sentiment Label: Based on the total score from the aggregated word values, the text is labeled as 'Positive,' 'Negative,' or 'Neutral.' This allows us to classify comments or text into appropriate sentiment categories.

The purpose of data labeling is to produce a dataset ready for further analysis. By labeling data, we can perform sentiment classification using machine learning methods like Naive Bayes Classifier and Support Vector Machine.

Sentiment Analysis Using NBC & SVM

Naive Bayes Classifier (NBC): NBC is a probability-based classification technique effective in sentiment analysis, especially when dealing with large datasets like social media comments. In the context of the 2024 Presidential Election, NBC is used to automatically categorize comments into positive, negative, or neutral based on word frequency in those comments. The model is trained with a labeled dataset to understand the context of words in sentences that indicate sentiment. For example, words like "successful" and "support" can be associated with positive sentiment, while "disappointed" and "criticize" may indicate negative sentiment.

According to (Bustomi et al. 2023), the Naive Bayes algorithm is highly efficient and capable of combining evidence from data, making it suitable for predicting future outcomes based on past historical data.

Support Vector Machine (SVM): SVM, on the other hand, is a model that operates by finding a hyperplane in the feature space that maximizes the margin between two sentiment categories. This technique is very effective in handling cases where the separation between categories is not linear. In the context of sentiment analysis for the 2024 Presidential Election, SVM will classify comments into sentiment categories based on the feature vectors generated from the text, considering the context and complexity of the language used in the comments.

According to (Firdaus et al. 2024)., initial testing of data using SVM and Naive Bayes methods shows high accuracy in text classification tasks. and according to (Saifullah Fattah and Indah Ratnasari 2023) describe SVM as a classification algorithm that finds the best hyperplane to separate data into different classes with maximum margin.

Explanation of Hyperplane Selection:

a) **Left Panel**: Exploring Hyperplanes Multiple Green Lines: This panel shows several potential lines as hyperplanes to separate two data classes (blue circles and red squares). These lines try various positions to find the best separation.

b) **Right Panel**: Optimal Hyperplane Selection Solid Black Line: This is the optimal hyperplane that successfully separates the two classes with the largest margin. The margin is the distance between the hyperplane and the closest data points from both classes, and the goal is to maximize it. Dashed Lines: These lines mark the margin boundaries. No data points are within this margin, helping to reduce classification errors. Support Vectors: The data points that define the margin (the nearest red squares and blue circles) are called support vectors. Their positions directly influence the definition and orientation of the hyperplane.

According to (Geni, Yulianti, and Sensuse 2023) [23] and (Damayanti and Lhaksmana 2024), the Support Vector Machine (SVM) classifier is a novel and effective implementation of statistical learning theory for text mining, offering high accuracy. (Haryanto et al. 2019). note that the Naive Bayes Classifier's assumption of independence among class attributes simplifies probability calculation. According to (Alfonso and Bhisetya Rarasati n.d.)highlight SVM's high accuracy in text mining tasks.

Model Evaluation Using Confusion Matrix

The Confusion Matrix is a common tool in machine learning for evaluating classification model performance. For public sentiment analysis towards the 2024 Presidential Election using Naive Bayes Classifier (NBC) and Support Vector Machine (SVM), according to (Setiawan and Dewi 2023)*Multinomial Naïve Bayes is one of a series of text classification processes owned by Naïve Bayes by presenting a calculation method to determine the frequency of each word that appears*. dan Support Vector Machine (SVM) we can set up a Confusion Matrix with three sentiment categories: positive, negative, and neutral. Below is a detailed description of how the Confusion Matrix is configured and how evaluation metrics are calculated:

Confusion Matrix Configuration:

Type	Predicted Positive	Predicted Negative	Predicted Neutral
Actual Positive	TP_pos	FN_pos_neg	FN_pos_net
Actual Negative	FP_neg_pos	TN_neg	FN_neg_net
Actual Neutral	FP_net_pos	FP_net_neg	TN_net

Definitions and Formulas:

a) True Positive (TP): The count of actual positives correctly predicted as positive.

b) True Negative (TN): The count of actual negatives correctly predicted as negative.

c) False Positive (FP): The count of negatives incorrectly predicted as positive.

d) False Negative (FN): The count of positives incorrectly predicted as negative

Each element of the matrix is a succinct representation of the model's prediction outcomes, allowing for a straightforward calculation of performance metrics:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

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

$$Recal = \frac{TP}{TP + FN}$$

$$F1\ SCORE = 2\ X \left(\frac{Precision\ X\ Recall}{Precision + Recall}\right)$$

RESULTS AND DISCUSSION

Data Description

This study analyzes comment data from YouTube videos concerning the three presidential candidate pairs in the 2024 Indonesian Presidential Election. Data was collected through a data crawling process on Google Collaboratory, which facilitates the efficient execution of Python scripts in the cloud. The dataset contains approximately 1,800 comments, providing a diverse spectrum of public sentiments including support, criticism, and neutral viewpoints towards the candidates. This comprehensive data composition is essential for understanding the broad public opinion expressed on social media regarding these political figures.

Characteristics of the Data:

a) Support: Comments that show support for one of the candidate pairs, reflecting either satisfaction with past policies or positive expectations for proposed programs.
b) Criticism: Comments that express dissatisfaction or criticism towards the candidates, often referring to perceived past failures or skepticism about the effectiveness of proposed plans.
c) Neutral: Comments that neither support nor criticize but discuss related issues or request further information.

Labeling Data with Indonesian Sentiment Lexicon

The method utilized for automatic labeling is Lexicon Based, which uses dictionaries as sources of language or vocabulary. This approach assigns a sentiment label to each opinion, enabling automatic classification of sentences into positive, negative, and neutral categories (Amal 2023). The process involves employing sentiment lexicons and sentiment analysis functions to ensure accurate categorization.

Creating Sentiment Lexicons:

Two lexicons were imported from Excel files: one containing words with positive connotations and the other with negative connotations. These were converted into dictionaries in Python to facilitate lookups and analysis during sentiment evaluation.

Code for Indonesian Sentiment Lexicon:

```
# @title Lebeling Data Sentimen Lexicon
lexicon_positive =
pd.read_excel('/content/drive/MyDrive/P2024/sempro/kamusID/kamus_positive.xlsx')
lexicon_positive_dict = {}
for index, row in lexicon_positive.iterrows():
  if row[0] not in lexicon_positive_dict:
    lexicon_positive_dict[row[0]] = row[1]
lexicon_negative =
pd.read_excel('/content/drive/MyDrive/P2024/sempro/kamusID/kamus_negative.xlsx')
lexicon_negative_dict = {}
for index, row in lexicon_negative.iterrows():
  if row[0] not in lexicon_negative_dict:
    lexicon_negative_dict[row[0]] = row[1]
def sentiment_analysis_lexicon_indonesia(text):
  score = 0
  for word in text:
    if (word in lexicon_positive_dict):
      score = score + lexicon_positive_dict[word]
  for word in text:
    if (word in lexicon_negative_dict):
      score = score + lexicon_negative_dict[word]
```

```
    sentimen=''
    if (score > 0):
        sentimen = 'Positive'
    elif (score < 0):
        sentimen = 'Negative'
    else:
        sentimen = 'Neutral'
    return score, sentimen
results = df['TextStemming'].apply(sentiment_analysis_lexicon_indonesia)
results = list(zip(*results))
df['Polarity'] = results[0]
df['Sentimen'] = results[1]
#data['sentimen'] = results[1]
df[['TextStemming','Polarity','Sentimen']]
```

The Lexicon Based method uses dictionaries to automatically assign sentiment labels to text, categorizing opinions into positive, negative, and neutral sentiments. This involves importing lexicons with positive and negative connotations from Excel files, converting them into Python dictionaries for efficient lookup and analysis. Each word in a text is checked against these dictionaries; words found in the positive dictionary increase the sentiment score, while those in the negative dictionary decrease it. Based on the final score, the text is classified as positive, negative, or neutral. This process aids in accurately categorizing large datasets by sentiment, providing a foundational analysis for further sentiment evaluation and visualization.

Figure 4.1: Result Sentiment analysis

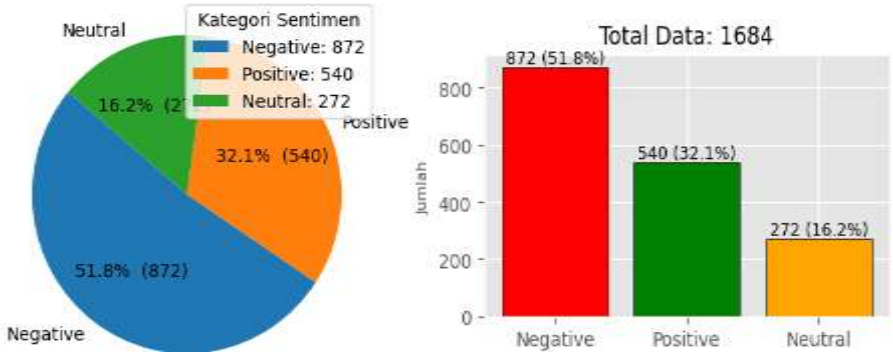


Table 4.1: Sentiment Analysis Category Results

Sentiment	Count
Negative	872
Neutral	272
Positive	540

Sentiment Analysis Classification

NBC is used for its effectiveness in processing large datasets and its ability to classify text based on probability. This model is suitable for sentiment analysis because it can quickly categorize comments based on the frequency of words associated with positive, negative, or neutral sentiments. According to research by (Christian, Wibowo, and Lyawati 2024), "The results show that splitting the data into 90% for training and 10% for testing produced the best results in the experiments," and according to research by (Anam et al. 2023), "The proportion of data for data testing and training is 20% and 80% respectively. The training data will be used to train the algorithm in determining the appropriate model." Based on research by Kurnia Ardiansyah Lubis et al. [33], where testing was conducted by splitting test data and training data in a 70% training and 30% testing ratio, the data train and test ratios used in this study are:

Table 4.2: Data Train and Data Test Ratios for NBC

Data Train	Data Test
90%	10%
80%	20%
70%	30%

SVM is implemented for its ability to find the optimal hyperplane that separates data categories with the largest margin. This method is effective for datasets with complex boundaries between sentiment categories. SVM uses kernels to transform data and find effective decision boundaries, even in cases where categories are not linearly separable. The same data train and test ratios used for NBC are applied to SVM.

Table 4.3: Data Train and Data Test Ratios for SVM

Data Train	Data Test
90%	10%
80%	20%
70%	30%

Sentiment Analysis Results

Analyzing the processed data provides significant insights into public sentiment distribution towards the presidential candidates:

a) **Sentiment Distribution:** Visualizations including bar and pie charts vividly depict the proportion of positive, negative, and neutral sentiments. These charts offer a clear visual representation of the public's reactions to each candidate.
b) **Model Evaluation:** The models are evaluated using confusion matrices which detail True Positives, False Positives, True Negatives, and False Negatives. Metrics such as accuracy, precision, recall, and F1-Score are utilized to quantitatively assess the performance of each model.

Visual Representation of Sentiment Words through Word Clouds

To further enhance our analysis, we employ Word Cloud visualizations, which provide a graphic representation of word usage frequency across different sentiment categories. These visualizations allow us to quickly identify and understand the most frequent and impactful words within each sentiment category, thereby giving us deeper insights into the themes and nuances present in the public commentary.

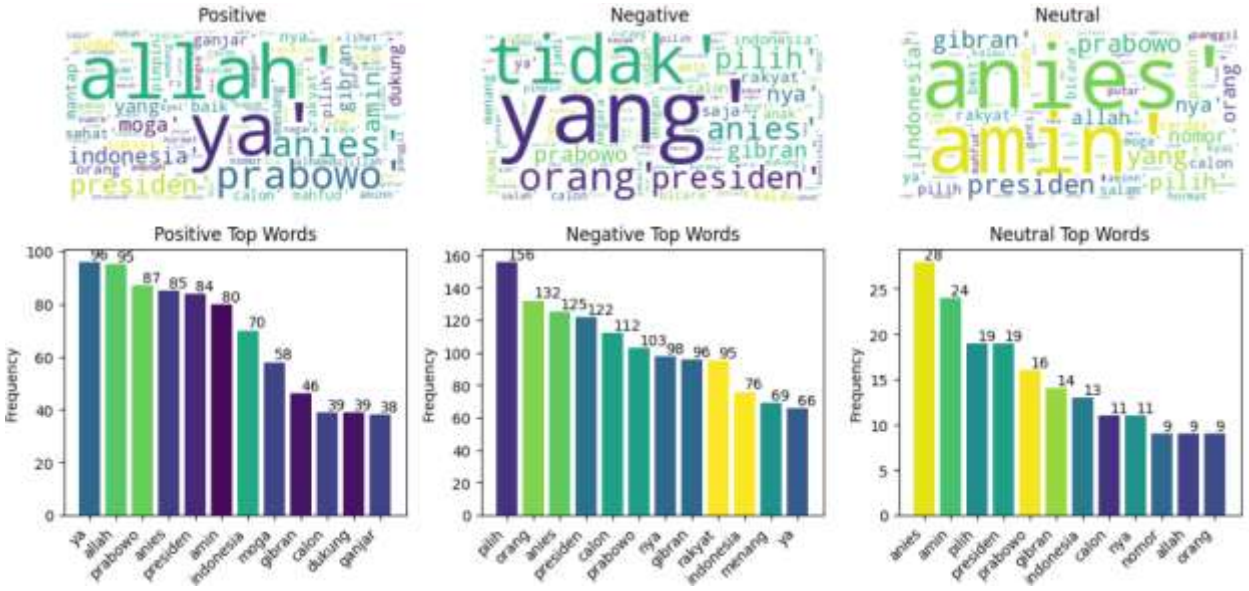
Generating Word Clouds

The following Python code snippet illustrates how we generate WordClouds for each sentiment category using the text data classified as positive, negative, and neutral:

```
import pandas as pd
import numpy as np
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer
import random
from nltk.corpus import stopwords
import nltk
# Pastikan stopwords telah diunduh dan siap digunakan
nltk.download('stopwords')
# Menggunakan stopword dari nltk untuk bahasa Indonesia
stop_words_indo = set(stopwords.words('indonesian'))
custom_stopwords = pd.read_csv('/content/drive/MyDrive/P2024/sempro/kamusID/stopword.txt',
header=None)
custom_stopwords = set(custom_stopwords[0].tolist())
# Gabungan semua stopwords menjadi list untuk digunakan di CountVectorizer
all_stopwords = list(stop_words_indo.union(custom_stopwords))
# Setup for visualization
fig, axes = plt.subplots(2, 3, figsize=(12, 5))  #
for i, sentiment in enumerate(['Positive', 'Negative', 'Neutral']):
    temp_df = df[df['Sentimen'] == sentiment]
    words = ' '.join(temp_df['TextStemming'])
    # Generate the word cloud with Indonesian stopwords
    wordcloud = WordCloud(stopwords=all_stopwords, background_color='white', max_words=100,
collocations=False).generate(words)
    axes[0, i].imshow(wordcloud, interpolation="bilinear")
    axes[0, i].set_title(f"{sentiment} ")
    axes[0, i].axis('off')
    # Bar chart for top words using Indonesian stopwords
    vectorizer = CountVectorizer(stop_words=all_stopwords, max_features=12)
    count_data = vectorizer.fit_transform(temp_df['TextStemming'])
    sum_words = count_data.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word, idx in vectorizer.vocabulary_.items()]
    words_freq = sorted(words_freq, key=lambda x: x[1], reverse=True)
    words, freqs = zip(*words_freq)
    colors = [plt.cm.viridis(random.random()) for _ in words]  # Random color for each bar
    bars = axes[1, i].bar(range(len(words)), freqs, color=colors)
    axes[1, i].set_title(f"{sentiment} Top Words")
    axes[1, i].set_xticks(range(len(words)))
    axes[1, i].set_xticklabels(words, rotation=45, ha="right")
    axes[1, i].set_ylabel('Frequency')
    for bar in bars:
        yval = bar.get_height()
        axes[1, i].text(bar.get_x() + bar.get_width()/2, yval, f'{yval}', va='bottom')
plt.tight_layout()
plt.show()
```

This code is instrumental in visualizing the frequency of words within texts classified by sentiment, helping identify dominant words or phrases in each category, which provides quick visual insights into the themes and sentiments prevalent in the analyzed data. This further aids in understanding the emotional and topical elements discussed in the comments related to the presidential candidates.

Figure 4.2 Word cloud and frequency visualization



The WordCloud and frequency bar charts effectively highlight the most prominent words associated with each sentiment category in the dataset of YouTube comments on the Indonesian Presidential Election.

a) **Positive Sentiment**: Features optimistic and supportive terms like "baik" and "presiden," suggesting a positive regard towards certain candidates.

b) **Negative Sentiment**: Dominated by terms such as "tidak" and "pilih," indicating criticism or opposition related to the election or candidates.

c) **Neutral Sentiment**: Contains more factual or neutral terms such as "calon" and "Indonesia," which are typically used in general discussions about the candidates without expressing clear sentiment.

These visualizations succinctly encapsulate the main themes and sentiments expressed by the public, providing a quick visual reference to understand the prevailing opinions in the electoral discourse.

Naive Bayes Classifier (NBC)

NBC Results and Visualization: 10% Data Test: Accuracy 72%. Confusion matrix and classification results show strength in identifying negative sentiment but weaknesses in positive and neutral sentiments. 20% Data Test: Accuracy 72%. Consistent performance in identifying negative sentiment with high recall but still weak in positive and neutral sentiments. 30% Data Test: Accuracy 69%. Strength in negative sentiment identification with a decline in positive identification.

10% Data Test:

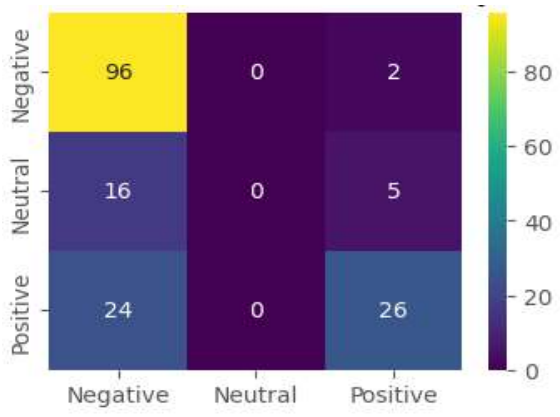
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{26 + 96}{26 + 2 + 96 + 24} = 0.72$$

$$Precision = \frac{TP}{TP + FP} = \frac{26}{26 + 2} = 0.929$$

$$Recal = \frac{TP}{TP + FN} = \frac{26}{26 + 24} = 0.52$$

$$F1\ SCORE = 2\ X\ \left(\frac{Precision\ X\ Recall}{Precision + Recall}\right) = 2\ X\ \left(\frac{0.929\ X\ 0.52}{0.929 + 0.52}\right) = 0.667$$

Gambar 4.3 NBC Visualisasi Confusion Matrix Pada Data Test 10%



20% Data Tes

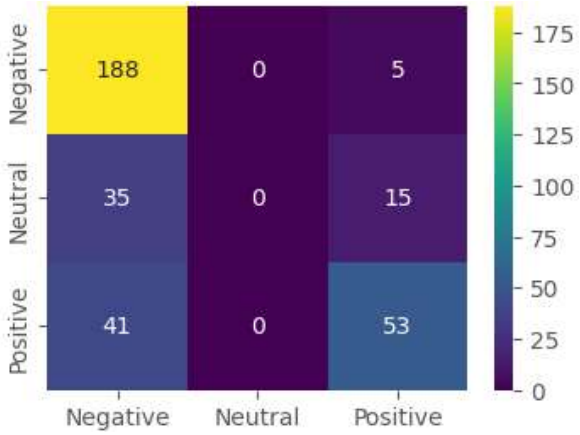
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{53 + 188}{53 + 5 + 188 + 41} = 0.72$$

$$Precision = \frac{TP}{TP + FP} = \frac{53}{53 + 5} = 0.914$$

$$Recal = \frac{TP}{TP + FN} = \frac{53}{53 + 41} = 0.564$$

$$F1\ SCORE = 2\ X\ \left(\frac{Precision\ X\ Recall}{Precision + Recall}\right) = 2\ X\ \left(\frac{0.914\ X\ 0.564}{0.914 + 0.564}\right) = 0.694$$

Gambar 4.4 NBC Visualisasi Confusion Matrix Pada Data Test 20%



30% Data Test:

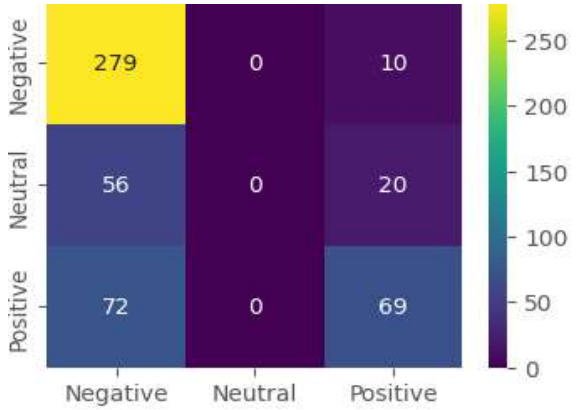
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{69 + 279}{69 + 10 + 279 + 72} = 0.69$$

$$Precision = \frac{TP}{TP + FP} = \frac{69}{69 + 10} = 0.873$$

$$Recal = \frac{TP}{TP + FN} = \frac{69}{69 + 72} = 0.49$$

$$F1\ SCORE = 2\ X\ \left(\frac{Precision\ X\ Recall}{Precision + Recall}\right) = 2\ X\ \left(\frac{0.873\ X\ 0.49}{0.873 + 0.49}\right) = 0.624$$

Gambar 4.5 NBC Visualisasi Confusion Matrix Pada Data Test 30%



Support Vector Machine (SVM)

SVM Results and Visualization:

10% Data Test: Accuracy 76%. Better at handling positive sentiment with 77% precision and 72% recall. 20% Data Test: Accuracy 75%. Improvement in handling positive sentiment, with 76% recall. 30% Data Test: Accuracy 73%. Stable overall performance, with strength in negative and positive.

10% Data Test:

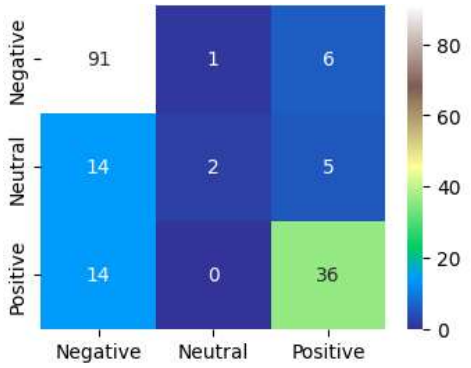
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{36 + 91}{36 + 6 + 91 + 14} = 0.76$$

$$Precision = \frac{TP}{TP + FP} = \frac{69}{36 + 6} = 0.857$$

$$Recal = \frac{TP}{TP + FN} = \frac{36}{36 + 14} = 0.72$$

$$F1\ SCORE = 2\ X\ \left(\frac{Precision\ X\ Recall}{Precision + Recall}\right) = 2\ X\ \left(\frac{0.857\ X\ 0.72}{0.857 + 0.72}\right) = 0.782$$

Gambar 4.5 SVM Visualisasi Confusion Matrix Pada Data Test 10%



20% Data Test:

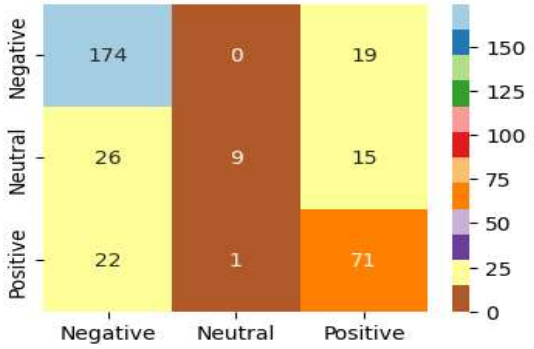
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{71 + 174}{71 + 19 + 174 + 22} = 0.75$$

$$Precision = \frac{TP}{TP + FP} = \frac{71}{71 + 19} = 0.789$$

$$Recal = \frac{TP}{TP + FN} = \frac{71}{71 + 22} = 0.763$$

$$F1\ SCORE = 2\ X\ \left(\frac{Precision\ X\ Recall}{Precision + Recall}\right) = 2\ X\ \left(\frac{0.789\ X\ 0.763}{0.789 + 0.763}\right) = 0.776$$

Gambar 4.5 SVM Visualisasi Confusion Matrix Pada Data Test 20%



30% Data Test:

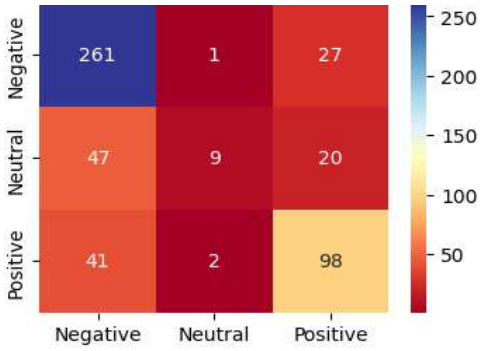
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{98 + 261}{98 + 27 + 261 + 41} = 0.73$$

$$Precision = \frac{TP}{TP + FP} = \frac{98}{98 + 27} = 0.784$$

$$Recal = \frac{TP}{TP + FN} = \frac{98}{98 + 41} = 0.705$$

$$F1\ SCORE = 2\ X\ \left(\frac{Precision\ X\ Recall}{Precision + Recall}\right) = 2\ X\ \left(\frac{0.784\ X\ 0.705}{0.784 + 0.705}\right) = 0.742$$

Gambar 4.5 SVM Visualisasi Confusion Matrix Pada Data Test 30%



CONCLUSION

This study has evaluated the effectiveness of Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) in analyzing public sentiment from YouTube comments related to the 2024 Indonesian Presidential Election. Using a dataset of **1.800** comments collected from November 2023 to March 2024, this study reveals that:

a) **Support Vector Machine (SVM)** recorded a higher accuracy rate compared to NBC, with the **highest** accuracy at 10% test data being 76.33%. For 20% and 30% test data, the accuracy was 75.37% and 72.73%, respectively. The highest precision and F1-Score recorded were 75.29% and 72.67%, confirming its ability to address complex sentiment classification issues.

b) **Naive Bayes Classifier (NBC)**, while offering ease of implementation, recorded lower accuracy, peaking at 72.19% for 10% test data, and 71.51% and 68.77% for 20% and 30% test data, respectively.

This study presents findings with significant implications for various stakeholders:

a) **Communication Strategy**: These findings are highly valuable for campaign teams and policymakers. By understanding the dominant public sentiments, they can tailor their

messages and communication strategies to better resonate with voters. For instance, if negative sentiment towards a specific issue or candidate is high, the campaign team can focus efforts on addressing those concerns.

b) **Sentiment Analysis Methodology Development**: This study highlights the need for more sophisticated methodologies in sentiment analysis, especially in handling complex and diverse languages as found in YouTube comments. Integrating deep learning technologies, such as more advanced neural networks or transformer-based models, could enhance the accuracy of sentiment classification by understanding broader contexts and more nuanced linguistic subtleties.

c) **Academic Contribution**: Academically, this research contributes to the existing literature by demonstrating the effective application of sentiment analysis methods to large and unstructured social media data. It opens opportunities for further research in optimizing algorithms for better accuracy and efficiency and applying them in other election contexts or similar application domains.

REFERENCES

Alexander, Nyongki, Radja Bria, and Arita Witanti. 2023. *ANALISIS SENTIMEN MASYARAKAT INDONESIA MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE TENTANG PILPRES 2024*. Vol. 7.

Alfonso, Michael, and Dionisia Bhisetya Rarasati. n.d. "JISA (Jurnal Informatika Dan Sains) Sentiment Analysis of 2024 Presidential Candidates Election Using SVM Algorithm."

Anam, M. Khairul, Triyani Arita Fitri, Agustin Agustin, Lusiana Lusiana, Muhammad Bambang Firdaus, and Agus Tri Nurhuda. 2023. "Sentiment Analysis for Online Learning Using The Lexicon-Based Method and The Support Vector Machine Algorithm." *ILKOM Jurnal Ilmiah* 15(2):290–302. doi: 10.33096/ilkom.v15i2.1590.290-302.

Ansori, Yusuf, and Khadijah Fahmi Hayati Holle. 2022. "Perbandingan Metode Machine Learning Dalam Analisis Sentimen Twitter." *Jurnal Sistem Dan Teknologi Informasi (JustIN)* 10(4):429. doi: 10.26418/justin.v10i4.51784.

Azzawagama Firdaus, Asno, Anton Yudhana, and Imam Riadi. 2024. "Prediction of Indonesian Presidential Election Results Using Sentiment Analysis with Naïve Bayes Method." doi: 10.30865/mib.v8i1.7007.

Bustomi, Yosep, Anwar Nugraha, Christina Juliane, and Sri Rahayu. 2023. "Data Mining Selection of Prospective Government Employees with Employment Agreements Using Naive Bayes Classifier." *Sinkron* 8(1):1–8. doi: 10.33395/sinkron.v8i1.11968.

Christian, Yefta, Tony Wibowo, and Mercy Lyawati. 2024. "Sentiment Analysis by Using Naïve Bayes Classification and Support Vector Machine, Study Case Sea Bank." *Sinkron* 9(1):258–75. doi: 10.33395/sinkron.v9i1.13141.

Damayanti, Lisyana, and Kemas Muslim Lhaksmana. 2024. "Sentiment Analysis of the 2024 Indonesia Presidential Election on Twitter." *Jurnal Dan Penelitian Teknik Informatika* 8(2). doi: 10.33395/v8i2.13379.

Fadlila Nurwanda, Winita Sulandari, Yuliana Susanti, and Zakya Reyhana. 2023. "Comparative Analysis Of Performance Levels Of Svm And Naïve Bayes Algorithm For Lifestyle Classification On Twitter Social Media." *INTERNATIONAL CONFERENCE ON DIGITAL ADVANCE TOURISM, MANAGEMENT AND TECHNOLOGY* 1(1):215–30. doi: 10.56910/ictmt.v1i1.65.

Firdaus, Asno Azzawagama, Anton Yudhana, Imam Riadi, and Mahsun. 2024. "Indonesian Presidential Election Sentiment: Dataset of Response Public before 2024." *Data in Brief* 52. doi: 10.1016/j.dib.2023.109993.

Geni, Lenggo, Evi Yulianti, and Dana Indra Sensuse. 2023. "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models." *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika (JITEKI)* 9(3):746–57. doi: 10.26555/jiteki.v9i3.26490.

Guru, Peran, Dalam Mengembangk, An Karakter, Peserta Didik Smkn, Jakarta Timur, Maryam Sulaeman, Muhammad Bachrun', Ulum Romadhoni, Frencis Matheos Sarimole, and Wahyu Septian. 2024. "Analisis Sentimen Masyarakat Terhadap Isu Penundaan Pemilu 2024 Pada Twitter Dengan Metode Naive Bayes Dan Support Vector Machine." *Jurnal Sains Dan Teknologi* 5(3):890–99. doi: 10.55338/saintek.v5i1.2789.

Haryanto, Budi, Yova Ruldeviyani, Fathur Rohman, T. N. Julius Dimas, Ruth Magdalena, and F. Muhamad Yasil. 2019. "Facebook Analysis of Community Sentiment on 2019 Indonesian Presidential Candidates from Facebook Opinion Data." Pp. 715–22 in *Procedia Computer Science*. Vol. 161. Elsevier B.V.

Imam, Muhamad, and Syafii Informatika. 2023. *SENTIMEN ANALISIS PADA MEDIA SOSIAL MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER (NBC)*. Vol. 3.

Imran, Bahtiar, Muh Nasirudin Karim, and Nur Isna Ningsih. 2024. *KLASIFIKASI BERITA HOAX TERKAIT PEMILIHAN UMUM PRESIDEN REPUBLIK INDONESIA TAHUN 2024 MENGGUNAKAN NAÏVE BAYES DAN SVM*.

Madjid, Marchenda Fayza, Dian Eka Ratnawati, and Bayu Rahayudi. 2023. "Sentiment Analysis on App Reviews Using Support Vector Machine and Naïve Bayes Classification." *Sinkron* 8(1):556–62. doi: 10.33395/sinkron.v8i1.12161.

Mantik, Jurnal, Imam Rasyidin Muqsith Rizqi Prasetyo, and Aziz Musthafa. 2023. *Comparison between Naive Bayes Method and Support Vector Machine in Sentiment Analysis of the Relocation of the Indonesian Capital*. Vol. 7. Online.

Manullang, Oktaviami, Cahyo Prianto, and Nisa Hanum Harani. 2023. *Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based Dan Random Forest*.

Robi Padri, Abdul, Asro Asro, and Indra Indra. 2023. "Classification of Traffic Congestion in Indonesia Using the Naive Bayes Classification Method." *Journal of World Science* 2(6):877–88. doi: 10.58344/jws.v2i6.285.

Rosyida, Tamara, Harjono P. Putro, and Herry Wahyono. 2024. "ANALISIS SENTIMEN TERHADAP PILPRES 2024 BERDASARKAN OPINI DARI TWITTER MENGGUNAKAN NAÏVE BAYES DAN SVM." *Teknokris*.

Saifullah Fattah, Fardhan, and Chanifah Indah Ratnasari. 2023. *Sentiment Analysis of Indonesian Presidential Candidate 2024 on Facebook*. Vol. 7. Fardhan Saifullah Fattah.

Salsabila, Fadila, and Utomo Budiyanto. 2023. *IMPLEMENTASI NAÏVE BAYES CLASSIFIER TERKAIT PENCALONAN GANJAR PRANOWO SEBAGAI CALON PRESIDEN 2024 DI TWITTER*. Vol. 2.

Saputra, Pramana Yoga, Dian Hanifudin Subhi, Fahmi Zain, and Afif Winatama. 2019. "IMPLEMENTASI SENTIMEN ANALISIS KOMENTAR CHANNEL VIDEO PELAYANAN PEMERINTAH DI YOUTUBE MENGGUNAKAN ALGORITMA NAÏVE BAYES." *Jurnal Informatika Polinema* 5.

Setiawan, Rudi, and Fitria Dewi. 2023. "Analysis of Twitter User Sentiment on Presidential Candidate Anies Baswedan Using Naïve Bayes Algorithm." *ILKOM Jurnal Ilmiah* 15(3):473–87. doi: 10.33096/ilkom.v15i3.1775.473-487.

Silalahi, William, and Adi Hartanto. 2023. "Klasifikasi Sentimen Support Vector Machine Berbasis Optimasi Menyambut Pemilu 2024." *JRST (Jurnal Riset Sains Dan Teknologi)* 7(2):245. doi: 10.30595/jrst.v7i2.18133.

Tohidi, Edi, Reza Perdana Herdiansyah, and Edi Wahyudin. 2024. *ANALISA SENTIMEN KOMENTAR VIDEO YOUTUBE DI CHANNEL TVONENEWS TENTANG CALON PRESIDEN PRABOWO SUBIANTO MENGGUNAKAN SUPPORT VECTOR MACHINE*. Vol. 8.