Journal of Applied Science, Engineering, Technology, and Education *ISSN: 2685-0591 (online)*
Vol. 6 No. 1 (2024)
https://doi.org/10.35877/454RI.asci2754

Sentimen Classification of General Election 2024 News Titles on Detik.com Online Media Website Using Multinominal Naïve Bayes Method

Nathanael Yudhistira Pradipta* &Hari Soetanto

Faculty of Information Technology, Budi Luhur University, Ciledug, Jakarta 12260, Indonesia

Abstract

In The 2024 elections will produce a variety of political tendencies in society, with different opinions regarding its implementation and conditions. The role of news headlines is very important in influencing speculation and public responses to certain topics or issues. This study investigates the sentiment conveyed in news headlines about the 2024 Election using the Multinomial Naïve Bayes approach. Data was gathered from Detik.com, an online media platform, utilizing search terms "Pemilu 2024" and "Pemilihan Umum 2024" through web scraping methods. The data preprocessing involved converting to lowercase, tokenization, punctuation removal, stopword elimination, normalization, and stemming. Training data comprised 90% while 10% was allocated for testing. Analysis showed an accuracy of 83.59% using CountVectorizer for data transformation. Beyond sentiment classification, the research also examines how political processes shape media narratives and influence public perception. The implications highlight the impact of online media sentiment on digital democracy dynamics, providing valuable insights for political practitioners, policymakers, and media scholars. This study is not only academically significant but also offers practical insights for enhancing public understanding of the electoral process.

Keywords: 2024 Election; digital democracy; multinominal naïve bayes; sentiment

Received: 15 March 2024 Revised: 10 June 2024 Accepted: 27 June 2024

1. Introduction

The 2024 General Election has sparked diverse reactions and discussions within society. Mass media, particularly its digital forms, plays a crucial role as a readily accessible source for news coverage. However, in this dynamic environment, differing opinions and sentiments often emerge in each published news piece. The media's responsiveness is influenced by its imperative to provide unbiased news, ensuring it does not align with any particular party, which significantly impacts public perception. Moreover, contemporary society tends to react swiftly to news headlines alone, highlighting the pivotal role of news titles in media publication.

Detik.com, a widely popular media outlet known for its independence and accreditation by the Press Council, holds considerable influence over public opinion across various demographics. Media regulated by the Press Council are expected to uphold journalistic ethics and maintain credibility with the public. Independence is crucial in sentiment analysis research, particularly in the realms of politics and law, as it allows for editorial freedom and ensures objective and balanced reporting. Operating free from political pressures, specific societal groups, or commercial interests, independent media can present news that reflects diverse viewpoints and opinions. This provides researchers with an opportunity to conduct thorough sentiment analysis based on news headlines.

In the digital age, the advent of Big Data has revolutionized news gathering by offering abundant alternative data sources for generating official statistics (Pramana et al., 2017). Additionally, Detik.com serves as a platform for authentic and diverse public opinions on political and legal issues, particularly among critical and progressive youth. Therefore, this study utilizes Detik.com as a reliable and independent online media source to provide a more precise and comprehensive understanding of public views and sentiments toward current political and legal policies through

* Corresponding author.
E-mail address: nathanaelyudhistira52@gmail.com

public news channels. The journalistic freedom of independent media ensures that sentiment analysis remains unbiased by specific political or legal agendas, thereby enhancing the objectivity and significance of research findings.

Sentiment classification in text mining is crucial for discerning opinions, categorized typically as positive, negative, or neutral. The Multinomial Naïve Bayes algorithm is well-suited for this task, given its efficiency and effectiveness. By leveraging probabilities, it identifies sentiment in text by analyzing word frequencies. This algorithm stands out for its simplicity, high accuracy, and ease of implementation. Moreover, it handles irrelevant features or noise in data robustly. Therefore, Multinomial Naïve Bayes offers a dependable and efficient method for sentiment classification in news articles headlines.

2. Research Method

Sentiments are subjective expressions or attitudes conveyed through text or language, encompassing positive, negative, or neutral assessments of a specific subject or entity (*Merriam-Webster, Inc*, 2024), Referring to the definition of sentiment, the relationship with the definition of sentiment analysis is not much different. Sentiment analysis is a process for evaluating individual opinions, sentiments, evaluations, and emotions regarding a topic expressed in text or language form (Wulandari et al., 2021).The method applied in the 2024 Election News Title Sentiment Classification on the online media website Detik.com involves several main stages. These steps include data collection, data preprocessing, data transformation and extraction, training models, and evaluation. Data collection was carried out by means of web scraping from the online media website Detik.com. The data that has been obtained is then processed first to remove interference or irrelevant information so that when the data is analyzed the results are accurate. After data preprocessing is complete, sentiment labeling and data visualization based on sentiment will be carried out automatically. Data transformation and extraction was carried out using CountVectorizer to count the number of occurrences of each word in the document. After the data is transformed using CountVectorizer, a Multinominal Naive Bayes classification algorithm model is drilled on the data to measure sentiment accuracy. The process flow presented on Figure 1.

2.1. Web Scraping

The data collection technique used in this research uses the Web Scraping Algorithm. Web scraping is employed to automatically and efficiently gather data from web pages (Hafiz & Sudarmilah, 2023). Web scraping is a technique for collecting data from the internet and storing it for further analysis or use. Web data is usually retrieved via Hypertext Transfer Protocol (HTTP) or with a web browser. This process can be done manually by the user or automatically using a bot (Zhao, 2017). Web Scraping technique in This research utilizes the Visual Studio Code application along with the Python programming language and the BeautifulSoup library to facilitate the extraction of data from HTML and XML documents.

General steps for collecting news headline data from the online media website Detik.com via web scraping include the following:

a. Understand the Website rules

   Before beginning web scraping activities, it's crucial to carefully review the privacy policy and terms of use of detik.com. Certain websites may expressly forbid the use of scraping methods or impose limitations on the frequency of data requests. Therefore, it's essential to understand and adhere to the site's regulations before initiating any scraping procedures. As part of the permission process for conducting web scraping on detik.com for non-commercial academic purposes, a formal request was sent to redaksi@detik.com from the university, ensuring compliance with ethical and legal guidelines.

b. Web Page Structure Analysis

   Understanding the structure of a web page is done using the Inspect element feature in a web browser to view the HTML code of the page.

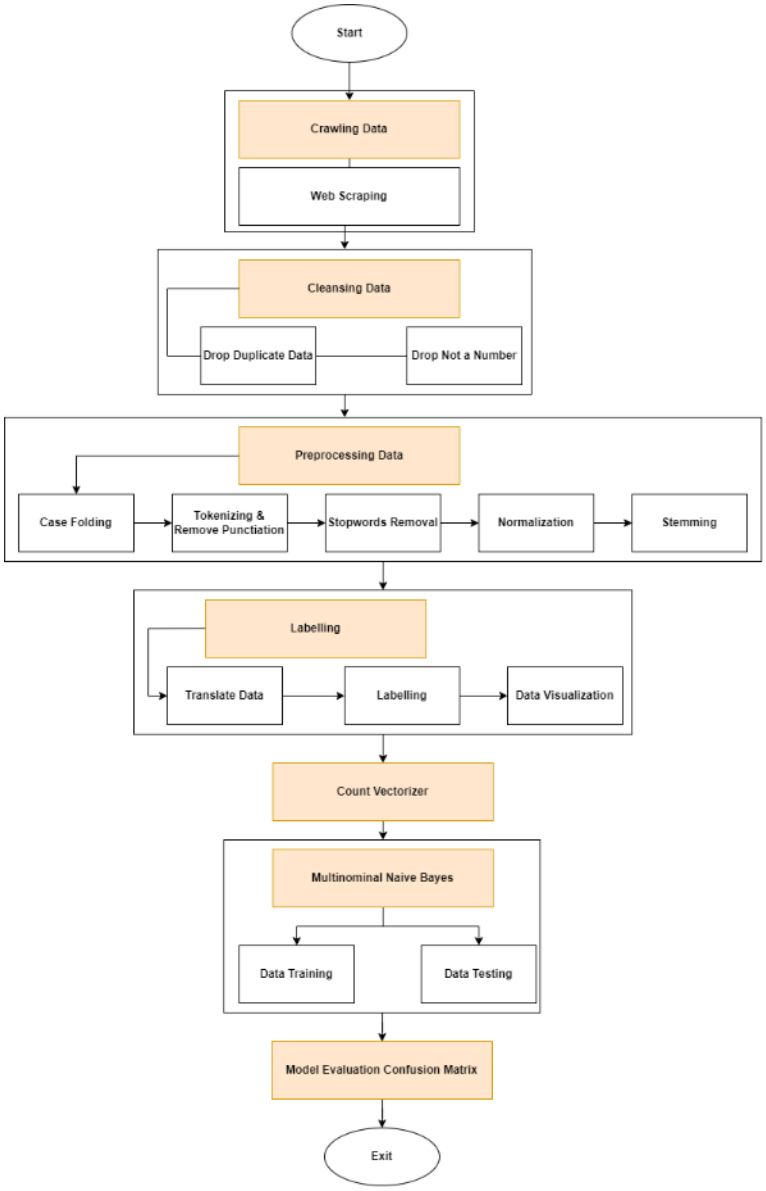

Figure 1. Method Implementation Process Flow

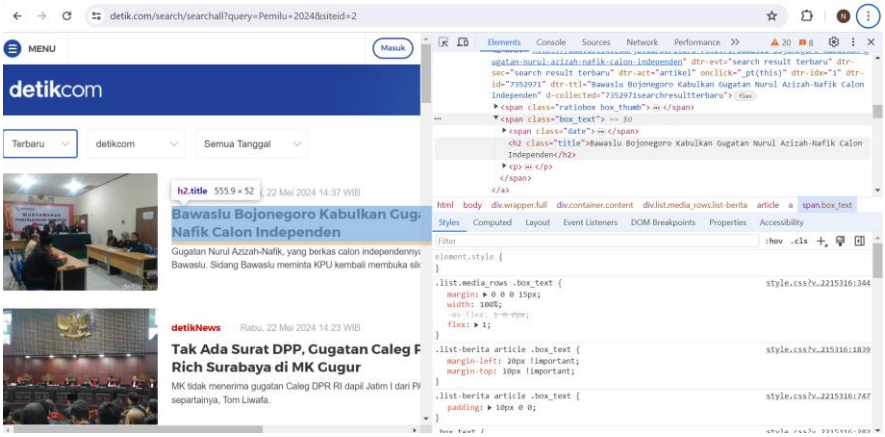

Figure 2. Detik.com Web Page Structure

c. Data Location Identification

The HTML element containing the headline on the detik.com web page is in the '<h2>' tag with class title.

d. Library

Using a web scraping library or framework for beautiful soup for the programming language used, namely Python.

e. Create Scraping Code

Create a script or program that uses the selected library or scraping tool to extract news titles from Detik.com pages. This script will make an HTTP request to the news page URL, extract the news titles from the HTML code, and save them in the desired format.

f. Saving Data

After the news titles are extracted, save the data in the appropriate format. This can be a text file, spreadsheet, or database, depending on your needs. In this research, the data will be saved in .csv file format by importing files via the library from google.colab.

These stages are often used sequentially in text data preprocessing to prepare text for further analysis, such as text classification, clustering, or sentiment analysis.

2.2. Data Sample Population

The study population consists of data collected by scraping news pages from the Detik.com website. The research utilized a sample of news titles containing real-time search keywords "2024 Election" and "2024 General Election". Through web scraping techniques, a dataset comprising 5057 relevant news titles related to the upcoming general election was gathered from website Detik.com.

Table 1. Sample Data

No	Headline
1	Air Mata Puan Maharani di Penghujung Rakernas PDIP
2	Megawati: Posisi Politik PDIP Selalu Diputuskan dalam Kongres Partai
3	Rakernas PDIP Rekomendasikan Megawati Jadi Ketum Lagi
4	Ketua AKD Trenggalek Daftar Bacawabup PDIP, Ini Janjinya
5	Wanti-wanti Said Abdullah ke Kader PDIP Jatim agar Tak Jual Kursi ke Bacalon
6	Golkar Persilakan RP dan Rudal Incar Rekomendasi di Pilwalkot Makassar
7	Megawati Ngaku Disebut Ratu Preman: Aku Nggak Pernah Mau Percaya
8	PDIP Lebih Sreg Usung Khofifah daripada Bikin Poros Baru di Pilgub Jatim
9	Ganjar Nilai Hasil Rakernas ke-V Gambarkan Sikap Politik PDIP ke Depan
10	KPU Kupang Lantik 153 Anggota PPS Pilkada 2024, Ada Kaum Disabilitas
…	…
…	…
…	…
5055	Update Rekapitulasi Suara di KPU Sumut: 27 Selesai, 4 Belum Masuk
5056	Saksi Ganjar-Mahfud Tolak Rekapitulasi Suara Pilpres Tingkat Provinsi di Bali
5057	Hasil Pleno KPU Jabar di 14 Daerah, Prabowo-Gibran Unggul

2.3. Cleaning Data

Data cleaning involves examining data to detect and rectify errors and inconsistencies found within the dataset. This stage (Widiari et al., 2020), also known as data auditing, focuses on identifying and correcting various anomalies that may impact the quality of the data.

In this research, the data cleaning process is conducted to ensure the cleanliness and accuracy of the data collected through web scraping methods. The steps involve eliminating duplicate entries and removing Not a Number (NaN) or empty data. Removing duplicates is essential to prevent unnecessary replication of information that could impact the final analysis. Simultaneously, eliminating NaN values enhances data quality by excluding invalid or irrelevant entries, thereby maintaining the integrity of the dataset for subsequent analysis. Implementing these data cleaning

procedures aims to yield more precise and dependable research outcomes while reducing the potential for bias or errors during analysis.

Table 2. Data Cleansing Results

Data	Duplicate	Nor a Number (NaN
5057	0	0

2.4. Preprocessing Data

Data preprocessing plays a crucial role in the sentiment analysis phase as it significantly enhances the quality and accuracy of the analytical results. Preprocessing serves as the primary step in data processing, aimed at transforming input data into a suitable format that is ready for further processing (Setyohadi et al., 2017). In this research, data preprocessing involves several key steps to clean and standardize text obtained through web scraping methods. These steps encompass case folding, tokenization and removing punctuation, eliminating stopwords, normalization, and stemming.

2.4.1. Case Folding

This stage involves changing all the letters in the text to lowercase or uppercase. This is done to ensure consistency in text processing (Ratino et al., 2020), so that the same words with different spellings are not considered different words.

Table 3. Case Folding Results

No	Before Case Folding	After Case Folding
1	Air Mata Puan Maharani di Penghujung Rakernas PDIP	air mata puan maharani di penghujung rakernas pdip
2	Megawati: Posisi Politik PDIP Selalu Diputuskan dalam Kongres Partai	megawati: posisi politik pdip selalu diputuskan dalam kongres partai
3	Rakernas PDIP Rekomendasikan Megawati Jadi Ketum Lagi	rakernas pdip rekomendasikan megawati jadi ketum lagi
4	Ketua AKD Trenggalek Daftar Bacawabup PDIP, Ini Janjinya	ketua akd trenggalek daftar bacawabup pdip, ini janjinya
5	Wanti-wanti Said Abdullah ke Kader PDIP Jatim agar Tak Jual Kursi ke Bacalon	wanti-wanti said Abdullah ke kader pdip jatim agar tak jual kursi ke bacalon
6	Golkar Persilakan RP dan Rudal Incar Rekomendasi di Pilwalkot Makassar	golkar persilakan rp dan rudal incar rekomendasi di pilwalkot makassar
7	Megawati Ngaku Disebut Ratu Preman: Aku Nggak Pernah Mau Percaya	megawati ngaku disebut ratu preman: aku nggak pernah mau percaya
8	PDIP Lebih Sreg Usung Khofifah daripada Bikin Poros Baru di Pilgub Jatim	pdip lebih sreg usung Khofifah daripada bikin poros baru di pilgub jatim
9	Ganjar Nilai Hasil Rakernas ke-V Gambarkan Sikap Politik PDIP ke Depan	ganjar nilai hasil rakernas ke-v gambarkan sikap politik kedepan
10	KPU Kupang Lantik 153 Anggota PPS Pilkada 2024, Ada Kaum Disabilitas	kpu kupang lantik 153 anggota pps pilkada 2024, ada kaum disabilitas
…	…	…
…	…	…
…	…	…
5055	Update Rekapitulasi Suara di KPU Sumut: 27 Selesai, 4 Belum Masuk	update rekapitulasi suara di kpu sumut: 27 selesai, 4 belum masuk
5056	Saksi Ganjar-Mahfud Tolak Rekapitulasi Suara Pilpres Tingkat Provinsi di Bali	saksi ganjar-mahfud tolak rekapitulasi suara pilpres tingkat provinsi di bali

No	Before Case Folding	After Case Folding
5057	Hasil Pleno KPU Jabar di 14 Daerah, Prabowo-Gibran Unggul	hasil pleno kpu jabar di 14 daerah, Prabowo-gibran unggul

2.4.2. Tokenizing and Remove Punctuation

The tokenization process is the process of dividing a sentence into a number of independent words. breaking text into smaller units, called tokens (Ratino et al., 2020). Tokens can be specific words, phrases, or characters. This separation allows further analysis of the text, such as calculating word frequencies or identifying sentence structures. In this research Tokenizing was carried out simultaneously by removing special characters, URLs, links from text, numbers, punctuation characters and white space. The tokenizing and removing punctuation stage is carried out after case folding is complete.

Table 4. Tokenizing and Punctuation Removal Results

No	Before Tokenizing and Remove Puctuation	After Tokenizing and Remove Puctuation
1	air mata puan maharani di penghujung rakernas pdip	[‘air’][‘mata’][‘puan’] [‘di’][‘penghujung’][‘rakernas’] [‘pdip’]
2	megawati: posisi politik pdip selalu diputuskan dalam kongres partai	[‘megawati’][‘posisi’][‘politik’] [‘pdip’][‘selalu’][‘diputuskan’] [‘dalam’][‘kongres’][‘partai’]
3	rakernas pdip rekomendasikan megawati jadi ketum lagi	[‘rakernas’][‘pdip’][‘rekomendasikan’] [megawati’][‘jadi’][‘ketum’] [‘lagi’]
4	ketua akd trenggalek daftar bacawabup pdip, ini janjinya	[‘ketua’][‘akd’][‘trenggalek’] [‘daftar’][‘bacawabup’][‘pdip’] [‘ini’][‘janjinya’]
5	wanti-wanti said Abdullah ke kader pdip jatim agar tak jual kursi ke bacalon	[‘wantiwanti’][‘said’][‘abdullah’] [‘ke’][‘kader’][‘pdip’] [‘jatim’][‘agar’][‘tak’] [‘jual’][‘kursi’][‘ke’][‘bacalon’]
6	golkar persilakan rp dan rudal incar rekomendasi di pilwalkot makassar	[‘golkar’][‘persilakan’][‘rp’][‘dan’] [‘rudal’][‘incar’][‘rekomendasi’][‘di’] [‘pilwalkot’][‘makassar’]
7	megawati ngaku disebut ratu preman: aku nggak pernah mau percaya	[‘megawati’][‘ngaku’][‘disebut’] [‘ratu’][‘preman’][‘aku’][‘nggak’] [‘pernah’][‘mau’][‘percaya’]
8	pdip lebih sreg usung Khofifah daripada bikin poros baru di pilgub jatim	[‘pdip’][‘lebih’][‘sreg’][‘khofifah’] [‘daripada’][‘bikin’][‘poros’][‘baru’] [‘di’][‘pilgub’][‘jatim’]
9	ganjar nilai hasil rakernas ke-v gambarkan sikap politik kedepan	[‘ganjar’][‘nilai’][‘hasil’][‘kev’] [‘gambarkan’][‘sikap’][‘politik’] [‘kedepan’]
10	kpu kupang lantik 153 anggota pps pilkada 2024, ada kaum disabilitas	[‘kpu’][‘kupang’][‘lantik’][‘anggota’] [‘pps’][‘pilkada’][‘ada’][‘kaum’] [‘disabilitas’]
…	…	…
…	…	…
…	…	…
5055	update rekapitulasi suara di kpu sumut: 27 selesai, 4 belum masuk	[‘update’][‘rekapitulasi’][‘suara’][‘di’] [‘kpu’][‘sumut’][‘selesai’][‘belum’] [‘masuk’]
5056	saksi ganjar-mahfud tolak rekapitulasi suara pilpres tingkat provinsi di bali	[‘saksi’][‘ganjarmahfud’][‘tolak’] [‘rekapitulasi’][‘suara’][‘pilpres’] [‘tingkat’][‘provinsi’][‘di’][‘bali’]
5057	hasil pleno kpu jabar di 14 daerah, Prabowo-gibran unggul	[‘hasil’][‘pleno’][‘kpu’[‘jabar’][‘di’] [‘daerah’][‘prabowogibran’][‘unggul’]

2.4.3. Stopwords Removal

Stopword removal entails eliminating frequently used words that do not contribute significant meaning in a text or document context (Guswandari & Cahyono, 2022). Common English stopwords include “the”, “is”, “and”, “in”, among others. This process aims to enhance the relevance of textual analysis by reducing irrelevant terms and noise. In the study, Indonesian stopwords from the NLTK library are employed for stopwords removal, followed by the elimination of additional irrelevant words. Subsequently, a function is applied to remove stopwords from each token.

Table 5. Stopwords Removal Results

Before Stopwords Removal				After Stopwords Removal	
Word	Attribute Name	Total Occurences	Document Occurences	Word	Attribute Name
ada	ada	>1	1	adian	adian
adalah	adalah	>1	1	adu	adu
adian	adian	>1	1	ahli	ahli
adu	adu	>1	1	aksi	aksi
ahli	ahli	>1	1	anies	anies
aksi	aksi	>1	1	bawaslu	bawaslu
anies	anies	>1	1	beda	beda
bawaslu	bawaslu	>1	1	…	…
beda	beda	>1	1	…	…
di	di	>1	1	…	…
dan	dan	>1	1	…	…
ini	ini	>1	1	…	…
ke	ke	>1	1	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

2.4.4. Normalization

Normalization is a process in text that aims to correct misspelled words or abbreviated words. For example, the word “kamu” can be spelled as “lu, lo”, and various other commonly used variations(Limbong et al., 2022). In this research, word normalization is carried out using normalized words which are saved into a CSV file. Apart from that, short word normalization was also carried out using the dictionary provided.

Table 6. Normalization Results

Word	Before Normalization	After Normalization
gue	gue	saya
lu	lu	kamu
nggak	nggak	tidak
tak	tak	tidak
yg	yg	yang

2.4.5. Stemming

Stemming involves reducing word inflections to their basic or uniform forms to enhance Information Retrieval performance (Agusta, 2009). For instance, words such as “berlari” and “lari-lari” are transformed into their base form “lari”. This process reduces word variations that could potentially disrupt analysis, ensuring more efficient and consistent processing. In this study, stemming was conducted using the Sastrawi library. The normalized word list was adjusted to their basic forms using a stemming algorithm

2.4.6. Preprocessing Results

Table 7. Preprocessing Data Results

No	Preprocessing Data Results
1	‘air’, ‘mata’, ‘puan’, ‘maharani’, ‘hujung’, ‘rakernas’, ‘pdip’
2	‘megawati’, ‘posisi’, ‘politik’, ‘pdip’, ‘putus’, ‘kongres’, ‘partai’

No	Preprocessing Data Results
3	‘rakernas’, ‘pdip’, ‘rekomendasi’, ‘megawati’, ‘tum’
4	‘ketua’, ‘akd’, ‘trenggalek’, ‘daftar’, ‘bacawabup’, ‘pdip’, ‘janji’
5	‘wantiwanti’, ‘said’, ‘abdullah’, ‘kader’, ‘pdip’, ‘jatim’, ‘jual’, ‘kursi’, ‘bacalon’
6	‘golkar’, ‘sila’, ‘rp’, ‘rudal’, ‘incar’, ‘rekomendasi’, ‘pilwalkot’, ‘makassar’
7	‘megawati’, ‘ngaku’, ‘ratu’, ‘preman’, ‘nggak’, ‘percaya’
8	‘pdip’, ‘sreg’, ‘usung’, ‘khofifah’, ‘bikin’, ‘poros’, ‘pilgub’, ‘jatim’
9	‘ganjar’, ‘nilai’, ‘hasil’, ‘rakernas’, ‘kev’, ‘gambar’, ‘sikap’, ‘politik’, ‘pdip’
10	‘kpu’, ‘kupang’, ‘lantik’, ‘anggota’, ‘pps’, ‘pilkada’, ‘kaum’, ‘disabilitas’
…	…
…	…
…	…
5055	‘update’, ‘rekapitulasi’, ‘suara’, ‘kpu’, ‘sumut’, ‘selesai’, ‘masuk’
5056	‘saksi’, ‘ganjarmahfud’, ‘tolak’, ‘rekapitulasi’, ‘suara’, ‘pilpres’, ‘tingkat’, ‘provinsi’, ‘bali’
5057	‘hasil’, ‘pleno’, ‘kpu’, ‘jabar’, ‘daerah’, ‘prabowogibran’, ‘unggul’

2.5. Translate Data to English Language

After the data preprocessing process is carried out, the next step is to translate the data into English using the Deep Translator Library. This is necessary because labeling will be done automatically using TextBlob. TextBlob can only process text in English. This translation process allows the use of TextBlob for the data labeling process, making it easier to analyze and further process it in English.

Table 8. Data Translation Results From Data Preprocessing

No	Preprocessing Data Output	Translate
1	‘air’, ‘mata’, ‘puan’, ‘maharani’, ‘hujung’, ‘rakernas’, ‘pdip’	‘water’, ‘eye’, ‘“ma’am”‘, ‘empress’, ‘end’, ‘National Working Meeting’, ‘pdip’
2	‘megawati’, ‘posisi’, ‘politik’, ‘pdip’, ‘putus’, ‘kongres’, ‘partai’	‘Megawati’, ‘position’, ‘political’, ‘pdip’, ‘separated’, ‘congress’, ‘party’
3	‘rakernas’, ‘pdip’, ‘rekomendasi’, ‘megawati’, ‘tum’	‘National Working Meeting’, ‘pdip’, ‘recommendation’, ‘Megawati’, ‘tum’
4	‘ketua’, ‘akd’, ‘trenggalek’, ‘daftar’, ‘bacawabup’, ‘pdip’, ‘janji’	‘chairman’, ‘akd’, ‘trenggalek’, ‘list’, ‘vice regent candidate’, ‘pdip’, ‘promise’
5	‘wantiwanti’, ‘said’, ‘abdullah’, ‘kader’, ‘pdip’, ‘jatim’, ‘jual’, ‘kursi’, ‘bacalon’	‘warning’, ‘said’, ‘Abdullah’, ‘cadre’, ‘pdip’, ‘East Java’, ‘sell’, ‘chair’, ‘Bacalon’
6	‘golkar’, ‘sila’, ‘rp’, ‘rudal’, ‘incar’, ‘rekomendasi’, ‘pilwalkot’, ‘makassar’	‘Golkar’, ‘please’, ‘rp’, ‘missile’, ‘aim’, ‘recommendation’, ‘election’, ‘Makassar’
7	‘megawati’, ‘ngaku’, ‘ratu’, ‘preman’, ‘nggak’, ‘percaya’	‘Megawati’, ‘confess’, ‘queen’, ‘thug’, ‘No’, ‘believe’
8	‘pdip’, ‘sreg’, ‘usung’, ‘khofifah’, ‘bikin’, ‘poros’, ‘pilgub’, ‘jatim’	‘pdip’, ‘comfortable’, ‘carry’, ‘Khofifah’, ‘make’, ‘axis’, ‘gubernatorial election’, ‘East Java’
9	‘ganjar’, ‘nilai’, ‘hasil’, ‘rakernas’, ‘kev’, ‘gambar’, ‘sikap’, ‘politik’, ‘pdip’	‘reward’, ‘mark’, ‘results’, ‘National Working Meeting’, ‘kev’, ‘picture’, ‘attitude’, ‘political’, ‘pdip’
10	‘kpu’, ‘kupang’, ‘lantik’, ‘anggota’, ‘pps’, ‘pilkada’, ‘kaum’, ‘disabilitas’	‘kpu’, ‘kupang’, ‘appoint’, ‘member’, ‘pps’, ‘regional elections’, ‘race’, ‘disability’
…	…	…
…	…	…
…	…	…
5055	‘update’, ‘rekapitulasi’, ‘suara’, ‘kpu’, ‘sumut’, ‘selesai’, ‘masuk’	‘updates’, ‘recapitulation’, ‘voice’, ‘kpu’, ‘North Sumatra’, ‘finished’, ‘enter’
5056	‘saksi’, ‘ganjarmahfud’, ‘tolak’, ‘rekapitulasi’, ‘suara’, ‘pilpres’, ‘tingkat’, ‘provinsi’, ‘bali’	‘witness’, ‘Ganjarmahfud’, ‘reject’, ‘recapitulation’, ‘voice’, ‘presidential election’, ‘level’, ‘province’, ‘bal’
5057	‘hasil’, ‘pleno’, ‘kpu’, ‘jabar’, ‘daerah’, ‘prabowogibran’, ‘unggul’	‘results’, ‘plenary’, ‘kpu’, ‘West Java’, ‘area’, ‘Prabowogibran’, ‘superior’

2.6. Labelling

Labeling is the process of assigning categories or tags to data or objects, with the aim of identifying or classifying the data so that it can be analyzed or processed further. In the context of machine learning, labeling is generally used in supervised learning where the model learns from data that has been labeled. For example, in text classification, each document is given a label such as "positive," "negative," or "neutral" for sentiment analysis, or in topic classification such as "sports" or "politics". The parameters of the three sentiment class labels are:

a. Neutral: This news provides clear and descriptive information without using words that can trigger readers' emotions.

Figure 3. Example of a News Headline with Neutral Sentiment

b. Positive: This news includes statements and sentences that provide support or are constructive to the topic discussed, avoiding the use of excessive or overly emotional terms.

Figure 4. Example of a News Headline with Positive Sentiment

c. Negative: This news includes statements, sentences, words and terms that contain criticism and criticism of the subject being discussed, but in a neutral form and do not incite the reader's emotions(McQuail, 1992).

Figure 5. Example of a News Headline with Negative Sentiment

The data labeling process carried out in this research uses sentiment analysis using the TextBlob library. This process aims to determine the polarity (positive, neutral, or negative) of each news headline in English that has gone through the translation process before the labeling stage is carried out.

After translating the data from Indonesian to English, it will enter the labeling stage automatically using the TextBlob library. Sentiment labeling produces 3 classes, namely positive, neutral and negative with the following distribution:

Table 9. Sentiment Distribution

Class	Count
Positive	1437
Neutral	2979
Negative	641
Count : 5057	

The distribution of the labeling data shows a significant inequality between the number of samples in each sentiment class in the DataFrame. The Positive class has 1437 samples, while the Negative class only has 641 samples, and the Neutral class has the highest number of samples, 2979.

2.6.1. Data Visualization

The results of data visualization in the form of a word cloud from labeling results using TextBlob show the distribution of words that reflect the sentiment in the text. The words that frequently appear in this word cloud provide an idea of the focus or main theme of the text being analyzed. The size of a word in a word cloud indicates the frequency with which it appears in the dataset, with larger words indicating higher frequency. This visualization helps quickly identify common patterns in the data, such as dominant positive, negative, or neutral words, enabling deeper analysis of the sentiment contained in the labeled text.

Figure 6. Positive Sentiment WordCloud

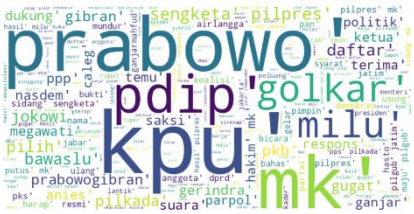

Figure 7. Neutral Sentiment WordCloud

Figure 8. Negative Sentiment WordCloud

2.7. Count Vectorizer

CountVectorizer is a transformation and feature extraction method that converts text into a numerical representation by counting the frequency of occurrence of each word in a document. It is one of the basic techniques in natural language processing (NLP) used to convert raw text data into features that can be used by machine learning models.

Count Vectorizer Formula:

$$\text{Frequency(term, doc)} = \text{the number of occurences the term in the doc} \quad (1)$$

Vocabulary: {‘air’: 64, ‘mata’: 2443, ‘puan’: 3306, ‘maharani’: 2353, ‘hujung’: 1436, ‘rakernas’: 3370, ‘pdip’: 3002, ‘megawati’: 2470, ‘posisi’: 3200,… }

Vector Array:

[[0 0 0 . . . 0 0 0]

[0 0 0 . . . 0 0 0]

[0 0 0 . . . 0 0 0]

. . .

[0 0 0 . . . 0 0 0]

[0 0 0 . . . 0 0 0]

[0 0 0 . . . 0 0 0]]

The values in the matrix indicate the frequency of occurrence of these words in each document. This technique is effective for converting text into a format that machine learning models can understand, such as in sentiment classification, where the occurrence of certain words can be an important feature in distinguishing between positive, negative, and neutral classes. CountVectorizer helps in preparing data for the machine learning process by retaining the structural information of the original text and ignoring irrelevant or common words (stop words), thereby increasing the accuracy and interpretability of the resulting model.

2.8. Multinominal Naïve Bayes Classifier

Naïve Bayes is a classification algorithm that is quite effective (Hadna et al., 2016), including in text sentiment analysis. This algorithm is based on Bayes' Theorem and assumes independence between each feature used in classification (Watratan et al., 2020), although in practice, the features may not be truly independent. Even though it is simple, Naïve Bayes often provides good results in sentiment classification because of its ability to overcome the problem of text data that has high and relatively small dimensions.

In this research, the Multinominal Naïve Bayes Classifier formula is applied in calculating conditional probabilities. The simple formula used is:

$$P(Word \mid Class) = \frac{the\ number\ of\ occurrences\ of\ the\ word\ in\ the\ class+1}{all\ the\ words\ in\ the\ class+all\ the\ words\ in\ the\ dictionary} \quad (2)$$

In the context of classification using Multinominal Naïve Bayes, conditional probability refers to the probability that a feature (in this case, words in a document) appears in a class. Each data must be given a label first before training is carried out,

This formula calculates the probability that a class exists, using Laplace smoothing to avoid zero probability values. This formula is applied iteratively for each word in the sample, then the posterior probability for each class is updated by adding the log likelihood of each word. Next, in this research the data was divided into 90% test data and 10% training data

3. Analysis Result

Classification Report training data using the Multinominal Naïve Bayes method includes the presented on Table 10.

Table 10. Classification Report

	Precision	Recall	F1-Score	Support
Negative	0.58	0.57	0.58	54
Neutral	0.89	0.88	0.89	317
Positive	0.80	0.84	0.82	136
Accuracy		0.84		506
Macro AVG	0.76	0.77	0.76	506
Weighted AVG	0.84	0.84	0.84	506

3.1. Confusion Matrix

In principle, the confusion matrix provides an overview of how well a classification system or model makes predictions on test data by comparing the predicted results with the actual values. This is represented in the form of a matrix table that describes the performance of the model on test data with known true values. The following is an illustration of a confusion matrix which includes 6 different combinations of predicted values and actual values.

		Predicted Class		
		Positive	Neutral	Negative
True Class	Positive	31 True Positive (TP)	17 False Neutral (FNt)	6 False Negative (FN)
	Neutral	17 False Positive (FP)	278 True Neutral (TNt)	22 False Negative (FN)
	Negative	5 False Positive (FP)	16 False Neutral (FNt)	114 True Negative (TN)

Figure 9. Confusion Matrix Training Data Results

There are 6 terms as representations of the results of the classification process in the confusion matrix. The six terms are True Positive (TP) for positive data that is predicted correctly. True Neutral (TNt) for neutral data that is predicted

correctly. True Negative (TN) for negative data that is predicted correctly. False Positive (FP) for negative data that is incorrectly predicted as positive by the model or known as type I error. False Negative (FN) for positive data that is incorrectly predicted as negative by the model or known as type II error. And finally is False Neutral (FNt).

Accuracy reflects the extent to which the model can perform correct classification or how close the model predictions are to the actual values.

$$accuracy = \frac{TP + TNt + TN}{TP + TNt + TN + FP + FNt + FN} \quad (3)$$

$$accuracy = \frac{31 + 278 + 114}{31 + 278 + 114 + 22 + 33 + 28}$$

$$accuracy = 0{,}83596838 = 0{,}84$$

$$accuracy = 0{,}84 * 100\%$$

$$accuracy = 84\%$$

From the results of the Confusion Matrix classification above, by calculating the accuracy value you can answer how effective the data training method is using the Multinominal Naïve Bayes classification algorithm.

4. Discussion

This literature review contains several journal references that compare previous research with current research. For example, (Permadi, 2020) conducted research entitled “Sentiment Analysis Using the Naïve Bayes Algorithm on Restaurant Reviews in Singapore”, applying the Naïve Bayes algorithm directly to the entire dataset without TF-IDF, Count Vectorizer, or similar data transformations, achieving an accuracy of 73.33 % (Permadi, 2020) In this research, CountVectorizer data transformation is applied which calculates the frequency of occurrence of each word in a document which is quite effective in converting text into a format that can be understood by machine learning models. As in sentiment classification, where the occurrence of words certain features can be important features in differentiating between classes positive, negative, and neutral. Another research conducted (Arsi & Waluyo, 2021) entitled “Sentiment Analysis of Discourse on Moving the Indonesian Capital City Using the Support Vector Machine (SVM) Algorithm” uses SVM for classification, where data preprocessing involves manual labeling by two annotators. In this research, labeling was carried out automatically using TextBlob so that the data was not very accurate and was often subjective. An annotator or expert can identify nuances and context that automated methods might miss. Finally, (Putra et al., 2021) conducted research entitled “Analysis of Community Sentiment towards PPKM Policy on Twitter Social Media Using the SVM Algorithm”, the accuracy of this analysis was low, even though the data collection between classes was balanced in terms of quantity.

Although the method applied in this research achieved good results, there are shortcomings that can still be corrected to achieve maximum results. As in the example that has been explained, manual labeling is carried out using language experts and other methods that can increase the accuracy of the results.

5. Conclusion

Based on research conducted on sentiment analysis in the headlines of the 2024 General Election news on the Detik.com website, using the Multinomial Naïve Bayes classification algorithm approach with CountVectorizer for data transformation and extraction, it can be concluded that the designed system works as expected. This system makes it easier to classify news titles on the Detik.com site. The application of sentiment analysis through data preprocessing processes such as data cleaning, case folding, tokenization, removal of stopwords, normalization and stemming, as well as data training using Multinomial Naïve Bayes, has been proven to make analysis easier to measure the accuracy of sentiment results. The accuracy level achieved reached 83.59%, showing good results in measuring sentiment regarding the 2024 General Election.

The results of the analysis show that neutral sentiment dominates 58% of election news titles on the online media website Detik.com. The results of this analysis can be a benchmark that Detik.com media, with its independence or media under the auspices of the press council, tends to publish news that is neutral in nature.

References

Agusta, L. (2009). Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia. *Konferensi Nasional Sistem dan Informatika*, 196–201.

Arsi, P., & Waluyo, R. (2021). Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM). *Jurnal Teknologi Informasi dan Ilmu Komputer*, 8(1), 147. https://doi.org/10.25126/jtiik.0813944

Guswandari, A., & Cahyono, R. P. (2022). Penerapan Sentimen Analis Menggunakan Metode Naïve Bayes dan SVM. *Jurnal Ilmu Data*, 2, 1–16.

Hadna, N. M. S., Santosa, P. I., & Winarno, W. W. (2016). Studi literatur tentang perbandingan metode untuk proses analisis sentimen di Twitter. *Seminar Nasional Teknolologi Informasi dan Komunikasi*, *2016*, 57-64.

Hafiz, Y. A., & Sudarmilah, E. (2023). Implementasi Web Scraping Pada Portal Berita Online. *Inisiasi*, 55–60. https://doi.org/10.59344/inisiasi.v12i1.120

Limbong, J. J. A., Sembiring, I., & Hartomo, K. D. (2022). Analisis Klasifikasi Sentimen Ulasan pada E-Commerce Shopee Berbasis Word Cloud dengan Metode Naive Bayes dan K-Nearest Neighbor. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 9(2), 347. https://doi.org/10.25126/jtiik.2022924960

McQuail, D. (1992). *Media performance : mass communication and the public interest*. Sage Publications.

Merriam-Webster, Inc. (2024, April 25). Merriam-Webster.com

Permadi, V. A. (2020). Analisis Sentimen Menggunakan Algoritma Naive Bayes Terhadap Review Restoran di Singapura. *Jurnal Buana Informatika*, 11(2), 141–151. https://doi.org/10.24002/jbi.v11i2.3769

Pramana, S., Yuniarto, B., Kurniawan, R., Yordani, R., Lee, J., Amin, I., Satyaning, P. P. N. L. P., Riyadi, Y., Hasyyati, A. N., & Indriani, R. (2017). Big data for government policy: Potential implementations of bigdata for official statistics in Indonesia. *2017 International Workshop on Big Data and Information Security (IWBIS)*, 17–21. https://doi.org/10.1109/IWBIS.2017.8275097

Putra, A., Haeirudin, D., Khairunnisa, H., & Latifah, R. (2021). Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Pada Media Sosial Twitter Menggunakan Algoritma SVM. *Prosiding Semnastek*.

Ratino, R. R., Hafidz, N. H. H., Anggraeni, S. A. A., & Gata, W. G. G. (2020). Sentimen Analisis Informasi Covid-19 Menggunakan Support Vector Machine Dan Naïve Bayes. *Jurnal Penelitian Ilmu dan Teknologi Komputer*, 12(2), 1–11.

Setyohadi, D. B., Kristiawan, F. A., & Ernawati, E. (2017). Perbaikan Performansi Klasifikasi dengan Preprocessing Iterative Partitioning Filter Algorithm. *Telematika*, 14(1), 12–20. https://doi.org/10.31315/telematika.v14i01.1960

Watratan, A. F., Puspita B, A., & Moeis, D. (2020). Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. *Journal of Applied Computer Science and Technology*, 1(1), 7–14. https://doi.org/10.52158/jacost.v1i1.9

Widiari, N. P. A., Suarjaya, I. M. A. D., & Githa, D. P. (2020). Teknik Data Cleaning Menggunakan Snowflake untuk Studi Kasus Objek Pariwisata di Bali. *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, 137. https://doi.org/10.24843/JIM.2020.v08.i02.p07

Wulandari, D. A., Saedudin, Rd. R., & Andreswari, R. (2021). Analisis Sentimen Media Sosial Twitter Terhadap Reaksi Masyarakat Pada Ruu Cipta Kerja Menggunakan Metode Klasifikasi Algoritma Naive Bayes. *E-Proceeding of Engineering*, 8, 9007–9016.

Zhao, B. (2017). *Web Scraping. In Encyclopedia of Big Data* (pp. 1–3). Springer International Publishing. https://doi.org/10.1007/978-3-319-32001-4_483-1