INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Text Mining for News Forecasting on The Turnback Hoax Website

Rio Wirawan [a], Erly Krisnanik [a], Artika Arista [a,b,*]

[a] *Information systems, Universitas Pembangunan Nasional Veteran Jakarta, Cilandak, Depok, 12450, Indonesia*
[b] *Department of Information Systems, Universiti Malaya, Kuala Lumpur, 50603, Malaysia*
Corresponding author: [*]artika.arista@upnvj.ac.id

***Abstract*— News has been disseminated swiftly via the internet due to the rapid growth of information technology. The rapid spreading of news often confuses because the truth cannot be ascertained. Additionally, online social media is becoming increasingly popular, making it an excellent environment for propagating false information, including misinformation, phony reviews, advertising, rumors, political remarks, innuendo, etc. This study's specific goal is to classify data using a data mining approach model called text mining so that a system can automatically do the classification. As a result, the study will produce a dataset, which can then be used to create an application using data mining's ability to predict breaking news. An application was produced by employing data mining to forecast recent news. This study was able to classify data using a naive Bayes data mining approach model so that a system can automatically do the classification. The study produced an accuracy of 77% obtained with training data of 82%. From 994 contents, the classification of misleading content reached 33.9%, false content as many as 24.85%, imitation content was 13.48%, fake content reached 11.07%, manipulated content was 9.86%, parody content was 3.22%, satire content was 2.31%, and connection content as many as 1.31%. This study then visualizes the results using bar charts and word clouds. This work also produced datasets with the naïve Bayes method of news data and news that has been valid. Afterward, the dataset will be used in making applications to produce prototypes of computer program applications.**

***Keywords*— News; text mining; turnback hoax website; dataset; naïve bayes.**

Manuscript received 29 Jun. 2023; revised 12 Oct. 2023; accepted 3 Nov. 2023. Date of publication 31 Mar. 2024.

I. INTRODUCTION

The rapid development of information technology [1]–[6] has caused the spread of news through the internet very quickly [7]. The rapid spreading of news often confuses because the truth cannot be ascertained. Online social media is becoming increasingly popular, which makes it an excellent environment for the propagation of false information, including misinformation [8], phony reviews, false advertising, rumors, false political remarks, innuendo, etc. By disseminating information in real-time among users worldwide, online social media like Twitter and Facebook can help individuals communicate with one another [9]. Social media is a leading medium for online social interaction and information sharing due to features like simplicity of use, cheap cost, and fast speed [10].

According to a survey done in 2020 by the Indonesian Internet Service Providers Association (APJII), 73 percent of the country's inhabitants, including some members of Generation Z and millennials, who make up 54 percent of the population, are now internet users [11]. As a result of the wide variety of material that is shared on social media, uncontrolled information, such as information and documents containing hoaxes, may spread across the internet.

Fake news or false stories are circulated online or through social media [12] or other media in a way that appears to be legitimate [13] and presents real information in a fake context to support untrue assertions [14] or mislead the audience [15]. Identifying fake news on social media sites can be challenging due to the abundance of information[16]. Nowadays, social media is more widely used and popular for fake news than traditional media [10]. The distinction between social media and news media is becoming increasingly hazy because of mass media and commercialization[17]. Due to the presence of bots, this problem becomes more complicated [18]. Due to its widespread usage to mislead and seduce online users with skewed facts, online news and information quality remain major current concerns [19].

Cahya [20] states that news can take the shape of the most recent information that can draw readers and be

disseminated through media such as the internet, newspapers, glass screens, social networks, or other media. The ingredients listed below, 5W+1H, are required for a news report (what, who, when, where, why, and how). Information conveyed via the internet is news that is delivered online[21]. Most of the information conveyed can be trusted [21], [22], but not all of it can be due to the large number of people who fabricate news stories to dupe the public into believing the fabrication to be true [22].

A hoax is described as a purposefully constructed lie to undermine public trust in a company, a product, a service, or a person [23] that is recklessly disseminated online[24]. Numerous parties feel disadvantageous because of news that contradicts their beliefs due to the quick distribution of false news through social media [25]. The purpose of widely disseminated hoaxes online is to cause public fear and are disseminated by careless people. Spreading hoaxes also serves as a means of deflecting attention away from concerns intended to address a topic that is receiving a lot of attention, covering the current issues with hoaxes that left the prior issue unaddressed [26].

The Indonesian Anti-Defamation Society, or MAFINDO, is a group that is particularly active in identifying news content containing hoaxes and informing the public about the actual events [27]. The internet community that is a part of the website turnbackhoax.id has tasked identifying content containing hoaxes [28]. The site can be accessed by internet users every day. News that begins with [false] indicates that the news contains inaccurate content or information that cannot be accounted for as true. News that begins with the title [clarification] supposing that the news has been clarified.

The method of identifying or categorizing news on the turnbackhoax.The id website is currently done manually; therefore, if the amount of information increases, it will be challenging to process. However, Mafindo's actions will be utilized as data for text mining to estimate the degree to which the news is considered a hoax, as this data pertains to news clarification from January to May 2021. We used text mining and the naive Bayes algorithm to create patterns from fake news data from mafindo's website https://turnbackhoax.id.

The Naive Bayes algorithm is a well-known, effective, and efficient classification technique in machine learning [29], [30]. Because it produces very good results, the classifier is utilized more frequently than sophisticated classification algorithms [31]. The algorithm has been applied numerous times to perform sentiment analysis for information retrieval systems. It has been used to identify trend titles of Indonesian-language journals, categorize fake news on Twitter, analyze public opinion regarding the "new normal" campaign on Twitter, and assess reviews of hotels. LinkedIn also uses naïve Bayes to identify name spam [32]. The Naïve Bayes algorithm is widely recognized for its high classification accuracy. In addition to its high speed and accuracy, Naïve Bayes is an easy-to-understand, straightforward algorithm [33]. Because of its simplicity, the Naive Bayes theorem has been included into the methods of numerous scholars. Effective computational techniques for deciding which features are essential for Naive Bayes classification [34].

A popular and easy-to-use classifier, Naïve Bayes (NB) is a robust machine learning approach. It is a very skilled probabilistic classifier with mathematical solid foundations. Therefore, NB is among the top classifiers. This can be attributed to numerous factors, which are summed up as follows: (1) Because of its training time's order O(N) with the dataset, NB can make predictions more quickly than other classification algorithms. (2) it can be taught with little input training datasets and work with larger ones. (3) the ease of use and simplicity of implementation combined with the capability of real-time training for new things, (4) No domain knowledge or adjustment parameters are needed to implement this classifier. (5) It manages data that is both discrete and continuous. (6) NB is less susceptible to missing data, (7) NB can manage a large amount of noise in the dataset, (8) Because its functions rely on an approximation of low-order probabilities that are taken from the training data, NB is an incremental learning approach. As new training data are acquired, they can be promptly updated. (9) If the Naive Bayes conditional independence assumption is true, it will converge faster than discriminative models such as logistic regression. (10) NB is adequate for real-time applications, and (11) NB can be applied to binary and multiclass classification issues.

As a result of mafindo's clarification of the news, it is anticipated that the training data will be more accurate and pertinent when using its data. This study aims to classify data using a data mining approach model so that a system can automatically do the classification. As a result, the study will produce a dataset, which can then be used to create an application due to data mining's ability to predict breaking news. An application is produced from data mining to forecast recent news. This is done using a data mining approach model so that the system can automatically handle classification.

II. Material And Method

Data mining is the process of extracting or understanding interesting patterns [35]in data contained in documents, or it can be understood as a series of operations to delve into or discover patterns in data that has been gathered as an added benefit of a dataset in the form of knowledge that has not previously been known manually. Finding links, including patterns, linkages, and effects that may offer clues to relevant knowledge, is the basic goal of the data mining process. Text mining is one of the specialized subfields of data mining [36]. Text mining is a new technique for semi-automatically identifying useful information from unstructured text data by indexing the text's words[37]. Text mining is a technique that is frequently used in literary studies to organize textual data and identify correlations and latent semantic patterns in the data [38]. The text mining technique can be widely applied for academic study to comprehend and examine user viewpoints [39].

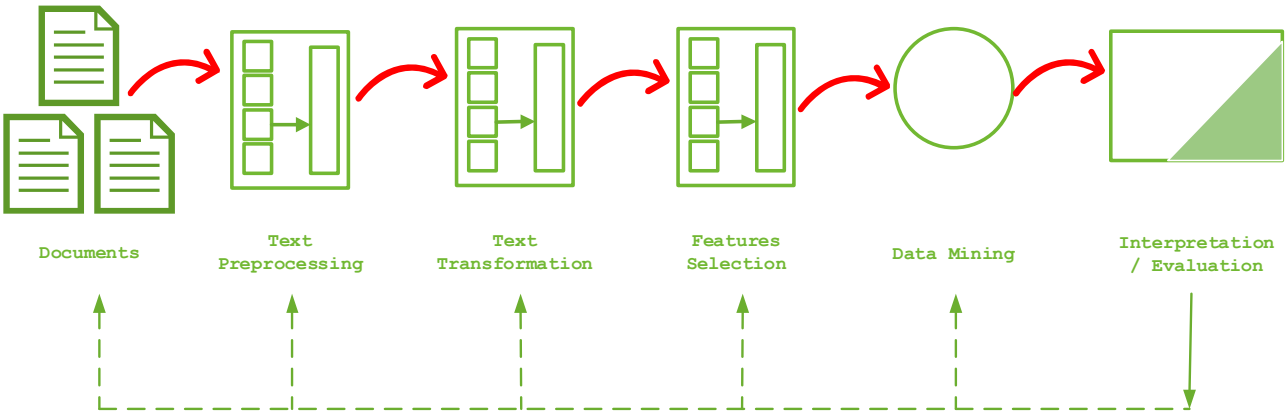

Fig. 1 Process Stage for Text Mining

It is clear that text mining is a subset of data mining to extract usable information for certain purposes from a collection of documents by looking for patterns in the text. Text categorization and text grouping can be done via text mining. Process Stage for Text Mining, which is implemented throughout the research:

A. Data Acquisition

Data Acquisition (DAQ) can be called data acquisition, a signal sampling process that measures the physical conditions that exist in the real world and converts the samples that have been generated to digital numeric values that a computer can process.

B. Data Exploration

At this stage, data exploration is carried out using statistical and mathematical functions and visualized in graphical form. This will make it easier to understand the data and database patterns.

C. Data Preprocessing

The Data Preprocessing stage is essential and cannot be missed. In data preprocessing, we present strategies for eliminating implicit noise from text data sets [40]. Data Preprocessing performed on the data includes the following sequence of stages:

1) Case Folding: Case folding is a process where the letters in each tweet are taken to all lowercase.

2) Tokenizing: Tokenizing is a process for cutting the input string performed on each compilation word.

3) Filtering: Filtering Function to delete if there is only one letter because it has no meaning. The frequent occurrences in sentences make the results of data extraction a lot better. One letter in question is, for example, y, g, k, and so on. Although the author comments that y is yes, g is no, and k is. So, the data extraction process is a word that is not easy to declare because it does not have a clear meaning.

4) Cleansing: Cleansing is the process of removing every character in a tweet, except for the alphabet. The goal is to reduce characters that have no meaning or are unwanted. The characters include numbers, @, # links from websites, and emojis.

5) Stemming: Stemming is changing a word into a basic word.

6) Clean Number: Deleting numbers always in front and behind the word. Even though in writing sentences, you always put a number at the beginning or end of each sentence to indicate that the sentence is repeated, in good Indonesian, it is wrong. Likewise, in a study, if you find a word that uses additional digits, it needs to be deleted. For example, rain2 means rain.

D. Classification

When the implementation is complete, tests from the implementation stage are continued. This stage uses test data to produce a high level of precision for classifying images on similar fruit shapes.

A supervised document classification method used the Naive Bayes algorithm. The Naive Bayes algorithm is a simple probabilistic classifier that creates a set of probabilities using a dataset's frequency distribution and value arrangement. The Naive Bayes technique performs a straightforward probabilistic classification process after computing the probability and configuration of values in a dataset [41]. The Nave Bayes method is also known as an algorithm with relatively simple calculations and a fairly quick learning process, and it is particularly well-suited for classifying data formed by several given categories. It is also known to have a higher level of accuracy than similar classification methods. A robust and anti-noise machine learning technique is the Naive Bayes (NB) algorithm [42].

A form of supervised learning built on the Bayes theorem is the Naive Bayes algorithm. According to the relational expression below, it presupposes that all features are independent of one another [43]. In relation to the independent feature vectors x_1 to x_n and the classification variable y:

$$P(y|x_1,\dots,x_n) = \frac{P(y)P(x_1,\dots,x_n|y)}{P(x_1,\dots,x_n)} \quad (1)$$

The following equation can be used to determine, using a priori probability values, the anticipated value of y when employing Naive Bayes [43].

$$\hat{y} = arg\max_y P(y)\prod_{i=1}^{n} P(x_i|y) \quad (2)$$

The prototype's classifier of choice, Naive Bayes, was chosen due to its superior speed when compared to more complex techniques. It organizes fresh data according to the likelihood that it belongs to a class. A Naive Bayes classifier is frequently the best option when the data set is tiny and has a lot of parameters. Naive Bayes classifiers are a family of probabilistic classifiers that employ Bayes' theorem under strict independence presumptions for the features [34].

E. Evaluation

The confusion matrix represents the data produced by machine learning algorithms as predictions and actual circumstances (actual). In order to determine the values of accuracy, precision, and recall, the performance of a classification system is measured using a confusion matrix.

TABLE I
CONFUSION MATRIX

Actual	Prediction	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Table I displays a confusion matrix for binary classification that illustrates the experimental validity distribution between the data generated by the prediction system using the established model and the real data. When choosing measuring measures, four different confusion matrices are utilized as references:

- True Positive (TP). The True Positive (TP) indicator shows the percentage of data that matches the system's prediction of a positive value and the actual value.
- False Positive (FP). The percentage of data the system mistakenly predicts would be positive when the actual values are negative, which is known as a false positive (FP).
- False Negative (FN). False Negative (FN) measures the proportion of data that the system mistakenly predicts to be negative when it is positive.
- True Negative (TN). The percentage of anticipated negative data with negative values is called True Negative (TN).

The classification performance is then determined using the accuracy, precision, recall, and F1-Score measures, using the value received from each component.

- The percentage of accurate system predictions to all available prediction outcomes is known as accuracy. Equation (3) illustrates the method for calculating the accuracy value.

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

- Precision is the next statistic, which assesses the ratio of the total positive predictive result (TP and FP) to the positive anticipated value as measured by the actual value (TP). Precision is calculated using Equation (4).

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

- Recall is the next statistic, and it tries to quantify the proportion of expected positive data to all positive data. The recall value calculation formula is shown in equation (5).

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

- The F1-Score statistic, which measures how well Precision and Recall compare on average, is the last one. The F1-Score is calculated using equation (6).

$$F1-Score = \frac{2 \times precission \times recall}{precission+recall} \quad (6)$$

F. Visualization using Word Cloud

Word Cloud is a data visualization technique used to represent text data where the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using the word cloud. Word Cloud is widely used to analyze data from social networking websites. Ariadi and Fithriasari [44] demonstrated 82.2%, 83.9%, 82.2%, and 82.4% for accuracy, precision, recall, and F-Measure. These findings will be used as a guide and factored into the application of the Naive Bayesian classifications approach in this research proposal.

Use datasets or data obtained from typical news website sources whenever possible, according to a study to classify hoax news [44], [45], even though the categorization and precision are good and high, the results are still regarded as being questionable because the data sources used to be not a hoax or to have been confirmed hoax. Because it is derived from the mafindo website, which can be viewed through the https://www.mafindo.or.id/ page, the authors can use data confirmed and shown to be a hoax. This application's benefit is that it uses verified and accurate hoax information from the mafindo website, which can be viewed at https://www.mafindo.or.id/. The generated dataset is trustworthy, legitimate, and dependable because it uses information from the website.

III. RESULT AND DISCUSSION

A. Data Acquisition

The data collection method used crawling techniques, which manually collected information on the web.turnbackhoax.id. Finally, the following data in Table II is obtained.

B. Data Preprocessing

It is essential to complete the Data Preprocessing stage. We outline methods for removing implicit noise from text data sets during data preprocessing. The stages of data preprocessing that are conducted on the data are as follows:

1) Removing Mentions and Retweets: Case folding is a process where the letters in each tweet. Data that had been crawled were filtered to produce clean data. The results can be seen in the picture. In the picture, a cleanup of words that had a mention of '@' and a retweet of 'RT' was carried out. The result is represented in Table III.

TABLE II
FEATURES

	Content	Type	Tag	Number of Tags	Source	Validator	Category
0	Benda Logam Menempel pada Lengan setelah divaksin	FALSE	covid-19	1.0	1	Riza Dwi	Misleading Content
1	Foto Demodex yang Hidup di Bulu Mata	FALSE	NaN	NaN	1	Renanda Dwina Putri	False Context
2	Foto Anak Berlumuran Darah Korban Serangan di ...	FALSE	NaN	NaN	1	Renanda Dwina Putri	False Context
3	Foto Dampak Abu Vulkanik di Tonga	FALSE	NaN	NaN	1	Fakta Fathia IS	False Context
4	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	FALSE	NaN	NaN	1	Renanda Dwina Putri	Manipulated Content
...	...	...	...	...	...	...	...
245	Video Jokowi Mengatakan Berminat Menjadi Presi...	FALSE	NaN	NaN	1	luthfiyah OJ	Manipulated Content
246	Pemerintahan Joe Biden Memberikan 30 Juta Doll...	FALSE	NaN	NaN	1	evarizma Zahra	Misleading Content
247	Anies Resmi Ditahan KPK	FALSE	NaN	NaN	1	Ani Nur MR	Manipulated Content
248	Video Tanggul Kali Pemali Brebes Jebol pada 13...	FALSE	NaN	NaN	1	Rahmah a n	False Context
249	Akun Whatsapp Wakil Wali Kota Denpasar Kadek A...	FALSE	NaN	NaN	1	NaN	Impostor Content

250 rows × 7 columns

TABLE III
THE PROCESS OF REMOVING MENTIONS AND RETWEETS

	Content	Type	Tag	Number of Tags	Source	Validator	Category	clean_content
0	Benda Logam Menempel pada Lengan setelah divaksin	FALSE	covid-19	1.0	1	Riza Dwi	Misleading Content	Benda Logam Menempel pada Lengan setelah divaksin
1	Foto Demodex yang Hidup di Bulu Mata	FALSE	NaN	NaN	1	Renanda Dwina Putri	False Context	Foto Demodex yang Hidup di Bulu Mata
2	Foto Anak Berlumuran Darah Korban Serangan di ...	FALSE	NaN	NaN	1	Renanda Dwina Putri	False Context	Foto Anak Berlumuran Darah Korban Serangan di ...
3	Foto Dampak Abu Vulkanik di Tonga	FALSE	NaN	NaN	1	Fakta Fathia IS	False Context	Foto Dampak Abu Vulkanik di Tonga
4	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	FALSE	NaN	NaN	1	Renanda Dwina Putri	Manipulated Content	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”

TABLE IV
THE PROCESS OF REMOVING SYMBOLS

	Content	Type	Tag	Number of Tags	Source	Validator	Category	clean_content	remove_http
0	Benda Logam Menempel pada Lengan setelah divaksin	False	covid-19	1.0	1	Riza Dwi	Misleading Content	Benda Logam Menempel pada Lengan setelah divaksin	Benda Logam Menempel pada Lengan setelah divaksin
1	Foto Demodex yang Hidup di Bulu Mata	False	NaN	NaN	1	Renanda Dwina Putri	False Context	Foto Demodex yang Hidup di Bulu Mata	Foto Demodex yang Hidup di Bulu Mata
2	Foto Anak Berlumuran Darah Korban Serangan di ...	False	NaN	NaN	1	Renanda Dwina Putri	False Context	Foto Anak Berlumuran Darah Korban Serangan di ...	Foto Anak Berlumuran Darah Korban Serangan di ...
3	Foto Dampak Abu Vulkanik di Tonga	False	NaN	NaN	1	Fakta Fathia IS	False Context	Foto Dampak Abu Vulkanik di Tonga	Foto Dampak Abu Vulkanik di Tonga
4	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	False	NaN	NaN	1	Renanda Dwina Putri	Manipulated Content	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”

TABLE V
CLEANSING RESULTS

	Content	Type	Tag	Number of Tags	Source	Validator	Category	clean_content	remove_http	remove_hastag
0	Benda Logam Menempel pada Lengan setelah divaksin	False	covid-19	1.0	1	Riza Dwi	Misleading Content	Benda Logam Menempel pada Lengan setelah divaksin	Benda Logam Menempel pada Lengan setelah divaksin	Benda Logam Menempel pada Lengan setelah divaksin
1	Foto Demodex yang Hidup di Bulu Mata	False	NaN	NaN	1	Renanda Dwina Putri	False Context	Foto Demodex yang Hidup di Bulu Mata	Foto Demodex yang Hidup di Bulu Mata	Foto Demodex yang Hidup di Bulu Mata
2	Foto Anak Berlumuran Darah Korban Serangan di ...	False	NaN	NaN	1	Renanda Dwina Putri	False Context	Foto Anak Berlumuran Darah Korban Serangan di ...	Foto Anak Berlumuran Darah Korban Serangan di ...	Foto Anak Berlumuran Darah Korban Serangan di ...
3	Foto Dampak Abu Vulkanik di Tonga	False	NaN	NaN	1	Fakta Fathia IS	False Context	Foto Dampak Abu Vulkanik di Tonga	Foto Dampak Abu Vulkanik di Tonga	Foto Dampak Abu Vulkanik di Tonga
4	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	False	NaN	NaN	1	Renanda Dwina Putri	Manipulated Content	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”

TABLE VI
DUPLICATE REMOVAL RESULTS

	Content	Type	Tag	Number of Tags	Source	Validator	Category	clean_content	remove_http	remove_hastag
0	Benda Logam Menempel pada Lengan setelah divaksin	False	covid-19	1.0	1	Riza Dwi	Misleading Content	Benda Logam Menempel pada Lengan setelah divaksin	Benda Logam Menempel pada Lengan setelah divaksin	Benda Logam Menempel pada Lengan setelah divaksin
1	Foto Demodex yang Hidup di Bulu Mata	False	NaN	NaN	1	Renanda Dwina Putri	False Context	Foto Demodex yang Hidup di Bulu Mata	Foto Demodex yang Hidup di Bulu Mata	Foto Demodex yang Hidup di Bulu Mata
2	Foto Anak Berlumuran Darah Korban Serangan di ...	False	NaN	NaN	1	Renanda Dwina Putri	False Context	Foto Anak Berlumuran Darah Korban Serangan di ...	Foto Anak Berlumuran Darah Korban Serangan di ...	Foto Anak Berlumuran Darah Korban Serangan di ...
3	Foto Dampak Abu Vulkanik di Tonga	False	NaN	NaN	1	Fakta Fathia IS	False Context	Foto Dampak Abu Vulkanik di Tonga	Foto Dampak Abu Vulkanik di Tonga	Foto Dampak Abu Vulkanik di Tonga
4	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	False	NaN	NaN	1	Renanda Dwina Putri	Manipulated Content	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”

TABLE VII
TOKENIZATION AND STEMMING RESULTS

	Content	Type	Tag	Number of Tags	Source	Validator	Category	clean_content	remove_http	remove_hastag
0	[benda, logam, tempel, lengan, vaksin]	False	covid-19	1.0	1	Riza Dwi	Misleading Content	Benda Logam Menempel pada Lengan setelah divaksin	Benda Logam Menempel pada Lengan setelah divaksin	Benda Logam Menempel pada Lengan setelah divaksin
1	[foto, demodex, hidup, bulu, mata]	False	NaN	NaN	1	Renanda Dwina Putri	False Context	Foto Demodex yang Hidup di Bulu Mata	Foto Demodex yang Hidup di Bulu Mata	Foto Demodex yang Hidup di Bulu Mata
2	[foto, anak, lumur, darah, korban, serang, yaman]	False	NaN	NaN	1	Renanda Dwina Putri	False Context	Foto Anak Berlumuran Darah Korban Serangan di ...	Foto Anak Berlumuran Darah Korban Serangan di ...	Foto Anak Berlumuran Darah Korban Serangan di ...
3	[foto, dampak, abu, vulkanik, tonga]	False	NaN	NaN	1	Fakta Fathia IS	False Context	Foto Dampak Abu Vulkanik di Tonga	Foto Dampak Abu Vulkanik di Tonga	Foto Dampak Abu Vulkanik di Tonga
4	[foto, mcdonalds, , hey, crypto, bro, , s, we,...	False	NaN	NaN	1	Renanda Dwina Putri	Manipulated Content	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”

2) Removing Symbols: Furthermore, cleaning of words with symbols on each character was performed, and the result is represented in Table IV.

3) Cleansing Process: Furthermore, cleansing was done to remove every character on the content, except the alphabet, to remove unwanted characters such as numbers, '#', and links from the website; the result is represented in Table V.

4) Duplicate Removal: Content from social media had a lot of repetitive text, so the data needs to be done to remove duplicates. The result is represented in Table VI.

5) Stop word Removal: Furthermore, the stop word removal process was performed where the hyphen used was in the Indonesian dictionary, and the second eliminates the non-standard hyphen contained in the dataset.

```
from nltk.corpus import stopwords
stopwords_indonesia = stopwords.words ('indonesian')

from Sastrawi.StopWordRemover. StopWordRemoverFactory import StopWordRemoverFactory, StopWordRemover, ArrayDictionary

stop_factory = StopWordRemoverFactory().get_stop_words()
more_stopwords = ['yg', 'utk', 'cuman', 'deh', 'Btw', 'tapi', 'gua', 'gue', 'lo', 'lu',
                  'kalo', 'trs', 'jd', 'nih', 'ntar', 'nya', 'lg', 'gk', 'ecusli', 'dpt',
                  'dr', 'kpn', 'kok', 'kyk', 'donk', 'yah', 'u', 'ya', 'ga', 'km', 'eh',
                  'sih', 'eh', 'bang', 'br', 'kyk', 'rp', 'jt', 'kan', 'gpp', 'sm', 'usah',
                  'mas', 'sob', 'thx', 'ato', 'jg', 'gw', 'wkwk', 'mak', 'haha', 'iy', 'k',
                  'tp', 'haha', 'dg', 'dri', 'duh', 'ye', 'wkwkwk', 'syg', 'btw', 'nerjemahin',
                  'gaes', 'guys', 'moga', 'kmrn', 'nemu', 'yukkk', 'wkwkw', "klas", 'iw', 'ew',
                  'lho', 'sbnry', 'org', 'gtu', 'bwt', 'klrga', 'clau', 'lbh', 'cpet',
                  'ku', 'wke', 'mba', 'mas', 'sdh', 'kmrn', 'oi', 'spt', 'dlm', 'bs', 'krn',
                  'jgn', 'sapa', 'spt', 'sh', 'wakakaka', 'sihhh', 'hehe', 'ih', 'dgn', 'la',
                  'kl', 'ttg', 'mana', 'kmna', 'kmn', 'tdk', 'tuh', 'dah', 'kek', 'ko', 'pls',
                  'bbrp', 'pd', 'mah', 'dhhh', 'kpd', 'tuh', 'kzl', 'byar', 'si', 'sii', 'sy',
                  'hahahaha', 'weh', 'dlu', 'tuhh'
]
data = stop_factory + more_stopwords

dictionary = ArrayDictionary(data)
str = StopWordRemover(dictionary)

print(data)
```

Fig. 2 Stop words Removal Process

6) Tokenizing and Stemming: Furthermore, a tokenizing process was undertaken to cut each input string that will be done in each sentence, then stemming was carried out to turn a word into a base word. The result is represented in Figure 3 and Table VII.

```
def clean_tweets(Content):

  #tokenize tweets
  tokenizer = TweetTokenizer (preserve_case=False, strip_handles=True, reduce_len=True)
  Content_tokens = tokenizer.tokenize(Content)

  Content_clean = []
  for word in Content_tokens:
    if (word not in data and # remove stopwords
        word not in emoticons and # remove emoticons
        word not in string.punctuation): # remove punctuation
        #tweets_clean.append(word)
        stem_word = stemmer.stem(word) # stemming word
        Content_clean.append(stem_word)

  return Content_clean
df['Content'] = df['remove_hastag'].apply(lambda x: clean_tweets(x))
```

Fig. 3 Tokenizing and Stemming Process

7) Document Filtering: Afterwards, a feature selection process was carried out to select records and features relevant to text mining work. Of the several features available, only one selection feature was chosen, namely content, because, in this content, some words and sentences indicate whether the news contained false news or true news. There was one class feature in this classification that was worth the wrong and clarification. News is false value if it contains content that does not match the relevant facts, news is worth clarifying if it is appropriate and relevant to the facts which represent in Table VIII. The results of the process in the selection of features in the dataset have two

features, namely type, and content, which can be seen in Table IX.

TABLE VIII
FEATURE SELECTION RESULTS

	Content	Category
0	benda logam tempel lengan vaksin	Misleading Content
1	foto demodex hidup bulu mata	False Context
2	foto anak lumur darah korban serang yaman	False Context
3	foto dampak abu vulkanik tonga	False Context
4	Foto McDonalds “Hey Crypto Bro’s WE ARE HIRING”	Manipulated Content

TABLE IX
FEATURE SELECTION RESULTS FROM A DATASET

Content	Type	Tag	Number of Tags	Source	Validator	Category
Benda Logam Menempel pada Lengan setelah divaksin	FALSE	covid-19	1	1	Riza Dwi	Misleading Content
Foto Demodex yang Hidup di Bulu Mata	FALSE			1	Renanda Dwina Putri	False Context
Foto Anak Berlumuran Darah Korban Serangan di Yaman	FALSE			1	Renanda Dwina Putri	False Context
Foto Dampak Abu Vulkanik di Tonga	FALSE			1	Fakta Fathia IS	False Context
Foto McDonalds â€œHey Crypto Broâ€™s WE ARE HIRINGâ€	FALSE			1	Renanda Dwina Putri	Manipulated Content
Video â€œMELALUI SERTIFIKAT HALAL, MUI KUASAI RATUSAN TRILIUNAN RUPIAHâ€	FALSE			1	Adi Syafitrah	Misleading Content
Meteor Jatuh di Wilayah Indonesia Pada Tanggal 7 Mei 2022	FALSE			1	Novita Kusuma Wardhani	False Content
Akun WhatsApp Wakil Bupati Sukoharjo Tawarkan Bantuan Sejumlah Dana	FALSE			1	Novita Kusuma Wardhani	Impostor Content
Foto Ratu Mongolia Terakhir Sebelum Dieksekusi Tahun 1938	FALSE			1	Renanda Dwina Putri	False Context
Video Satelit India Bertabrakan dengan Stasiun Luar Angkasa Internasional	FALSE			1	Renanda Dwina Putri	False Context
...	...			...	...	...
Kotoran Putih Pada Bayi Baru Lahir Dapat Dihilangkan Apabila Rajin Mengonsumsi Air Kelapa Selama Hamil	FALSE			1	Gabriela Nauli Sinaga	Misleading Content
Ade Armando Meninggal Dunia	FALSE			1	Gabriela Nauli Sinaga	Misleading Content
Video Dukungan Jokowi 3 Periode	FALSE			1	Gabriela Nauli Sinaga	Misleading Content

C. Data Classification

This approach used supervised learning, where the data already had a verified label. The labeling process has also been done by verificators who understand the fact-checking of news so that there is already a label when the data is entered for processing. Using a tokenizer for the arguments in the Get_Data.ipnyb file, each piece of data is processed with the following three parameters:

1) preserve_case = False: uppercase letters are converted to lowercase characters.

2) strip_handles = True: a handler that removes usernames that begin with "@" from the text will be applied.

3) reduce_len = True: to reduce words like "hello" to "hello" in length.

Based on Figure 4, from 994 content, the classification of misleading content reached 33.9%, incorrect content was 24.85%, imitation content reached 13.48%, false content as much as 11.07%, manipulated content was at 9.86%, parody content was 3.22%, satire content was 2.31%, and connection content reached 1.31 %.

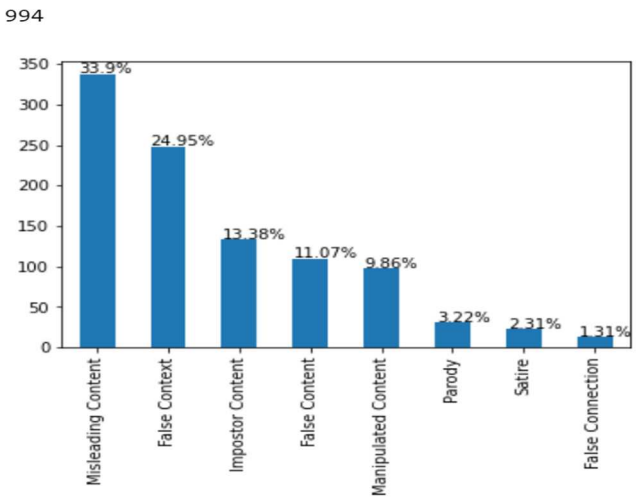

Fig. 4 Classification Analysis Results

D. Evaluation Results

Based on the classification using naïve Bayes, an accuracy of 77% was obtained with training data of 82%. The results can be seen in the following figure:

	precision	recall	f1-score	support
False Connection	0.00	0.00	0.00	13
False Content	0.98	0.56	0.72	110
False Context	0.79	0.90	0.84	248
Impostor Content	0.92	0.92	0.92	133
Manipulated Content	0.89	0.51	0.65	98
Misleading Content	0.71	0.96	0.82	337
Parody	1.00	0.03	0.06	32
Satire	0.00	0.00	0.00	23
accuracy			0.79	994
macro avg	0.66	0.49	0.50	994
weighted avg	0.79	0.79	0.75	994

Fig. 5 Evaluation Results

E. Visualization using Word Cloud

WordCloud below visualizes the words that appear most often on each label. Word cloud results on the overall content data showed that the most used words within the news content were the content, followed by the ones, photos, and context. The word cloud below in Figure 7 was obtained from Misleading Content Classification, False Context, and Copy Content.

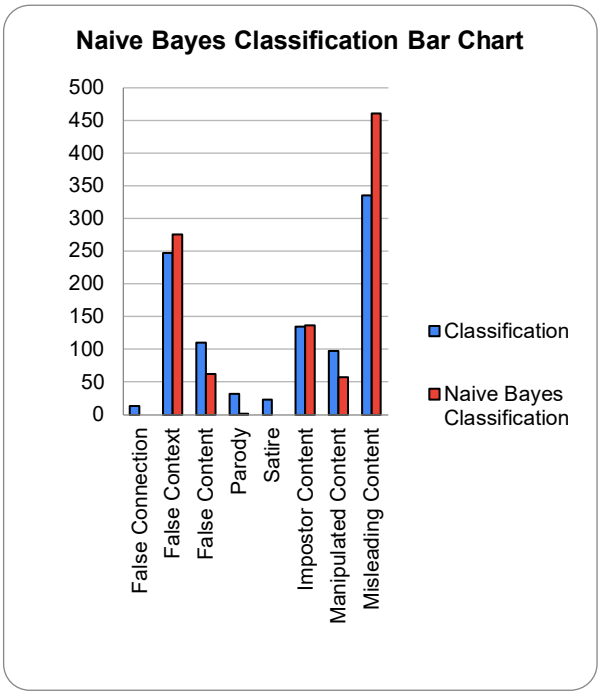

Fig. 6 Naive Bayes Classification Bar Chart

Fig. 7 Word cloud entire content

Fig. 8 Word cloud misleading content

After classification using the naive Bayes classification, the output produced by the naive Bayes model was obtained. We can make a visualization comparison, which is shown in Figure 6. Word cloud results on misleading content data showed that the most used words in news content include video, followed by *tinggal, dunia*, and covid, which are visualized in Figure 8.

IV. Conclusion

This study can classify data using a naive Bayes data mining approach model so that the classification can be done automatically by a system. The study produced an accuracy of 77% obtained with training data of 82%. From 994 contents, the classification of misleading content was 33.9%, false content was 24.85%, imitation content reached 13.48%, fake content was at 11.07%, manipulated content was 9.86%, parody content was 3.22%, satire content was 2.31%, and connection content reached 1.31%. This work produced datasets with the naïve Bayes method of news data and news that has been valid. Then, the dataset will be used in making applications to produce prototypes of computer program applications for development opportunities in future research.

Acknowledgment

This work is supported by the Research Institute and Community Service (LPPM) Universitas Pembangunan Nasional "Veteran" Jakarta (UPNVJ), Faculty of Computer Science UPNVJ, Information Systems Study Program UPNVJ for providing funding support and assisting the implementation of this research. The authors acknowledge financial support from Universitas Pembangunan Nasional "Veteran" Jakarta (UPNVJ), Decree No. 719/UN61.0/HK.02/2022 of the Rector of UPN Veteran Jakarta about Recipients of Internal Grant Funding for Research UPN Veterans Jakarta for the Year 2022.

References

[1] A. Arista and B. S. Abbas, "Using the UTAUT2 model to explain teacher acceptance of work performance assessment system," International Journal of Evaluation and Research in Education (IJERE), vol. 11, no. 4, p. 2200, Dec. 2022, doi:10.11591/ijere.v11i4.22561.

[2] A. Arista, "Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19," Sinkron, vol. 7, no. 1, pp. 59–65, Jan. 2022, doi:10.33395/sinkron.v7i1.11243.

[3] T. Tjahjanto, A. Arista, and E. Ermatita, "Information System for State-owned inventories Management at the Faculty of Computer Science," Sinkron, vol. 7, no. 4, pp. 2182–2192, Oct. 2022, doi:10.33395/sinkron.v7i4.11678.

[4] T. Theresiawati, H. B. Seta, and A. Arista, "Implementing quality function deployment using service quality and Kano model to the quality of e-learning," International Journal of Evaluation and Research in Education (IJERE), vol. 12, no. 3, p. 1560, Sep. 2023, doi: 10.11591/ijere.v12i3.25511.

[5] U. Rusdiana, I. Ernawati, N. Falih, and A. Arista, "Comparison of Distance Metrics on Fuzzy C-Means Algorithm Through Customer Segmentation," in *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2021, pp. 307–311.

[6] W. Cholil, F. Panjaitan, F. Ferdiansyah, A. Arista, R. Astriratma, and T. Rahayu, "Comparison of Machine Learning Methods in Sentiment Analysis PeduliLindungi Applications," in *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, IEEE, 2022, pp. 276–280.

[7] A. Arista and K. N. M. Ngafidin, "An Information System Risk Management of a Higher Education Computing Environment," International Journal on Advanced Science, Engineering and Information Technology, vol. 12, no. 2, p. 557, Apr. 2022, doi:10.18517/ijaseit.12.2.13953.

[8] Y. Wang, M. McKee, A. Torbica, and D. Stuckler, "Systematic Literature Review on the Spread of Health-related Misinformation on Social Media," Social Science & Medicine, vol. 240, p. 112552, Nov. 2019, doi: 10.1016/j.socscimed.2019.112552.

[9] S. Kumar, A. Mallik, A. Khetarpal, and B. S. Panda, "Influence maximization in social networks using graph embedding and graph neural network," Information Sciences, vol. 607, pp. 1617–1636, Aug. 2022, doi: 10.1016/j.ins.2022.06.075.

[10] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," Information Processing & Management, vol. 57, no. 2, p. 102025, Mar. 2020, doi:10.1016/j.ipm.2019.03.004.

[11] Tim APJII, "Buletin APJII," *APJII 91 Edition, July 2021*, Jul. 2021. Accessed: May 07, 2023. [Online]. Available: https://apjii.or.id/assets/media/buletin_apjii_edisi_91_-_juli_2021_bulletin.pdf

[12] J. Canavilhas and T. de M. Jorge, "Fake News Explosion in Portugal and Brazil the Pandemic and Journalists' Testimonies on Disinformation," Journalism and Media, vol. 3, no. 1, pp. 52–65, Jan. 2022, doi: 10.3390/journalmedia3010005.

[13] M. D. Molina, S. S. Sundar, T. Le, and D. Lee, "'Fake News' Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content," American Behavioral Scientist, vol. 65, no. 2, pp. 180–212, Oct. 2019, doi: 10.1177/0002764219878224.

[14] I. Ali, M. N. B. Ayub, P. Shivakumara, and N. F. B. M. Noor, "Fake News Detection Techniques on Social Media: A Survey," Wireless Communications and Mobile Computing, vol. 2022, pp. 1–17, Aug. 2022, doi: 10.1155/2022/6072084.

[15] P. N. Vasist and S. Krishnan, "Demystifying fake news in the hospitality industry: A systematic literature review, framework, and an agenda for future research," International Journal of Hospitality Management, vol. 106, p. 103277, Sep. 2022, doi:10.1016/j.ijhm.2022.103277.

[16] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," Applied Soft Computing, vol. 100, p. 106983, Mar. 2021, doi: 10.1016/j.asoc.2020.106983.

[17] D. Ehrenfeld and M. Barton, "Online Public Spheres in the Era of Fake News: Implications for the Composition Classroom," Computers and Composition, vol. 54, p. 102525, Dec. 2019, doi:10.1016/j.compcom.2019.102525.

[18] L. Soetekouw and S. Angelopoulos, "Digital Resilience Through Training Protocols: Learning To Identify Fake News On Social Media," Information Systems Frontiers, Jan. 2022, doi:10.1007/s10796-021-10240-7.

[19] A. Herasimenka, J. Bright, A. Knuutila, and P. N. Howard, "Misinformation and professional news on largely unmoderated platforms: the case of telegram," Journal of Information Technology & Politics, vol. 20, no. 2, pp. 198–212, May 2022, doi:10.1080/19331681.2022.2076272.

[20] H. Mustofa and A. A. Mahfudh, "Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes," Walisongo Journal of Information Technology, vol. 1, no. 1, p. 1, Nov. 2019, doi:.21580/wjit.2019.1.1.3915.

[21] J. Lee, K. Kim, G. Park, and N. Cha, "The role of online news and social media in preventive action in times of infodemic from a social capital perspective: The case of the COVID-19 pandemic in South Korea," Telematics and Informatics, vol. 64, p. 101691, Nov. 2021, doi: 10.1016/j.tele.2021.101691.

[22] E. Park, J. Park, and M. Hu, "Tourism demand forecasting with online news data mining," Annals of Tourism Research, vol. 90, p. 103273, Sep. 2021, doi: 10.1016/j.annals.2021.103273.

[23] K. Park and H. Rim, "Social media hoaxes, political ideology, and the role of issue confidence," Telematics and Informatics, vol. 36, pp. 1–11, Mar. 2019, doi: 10.1016/j.tele.2018.11.001.

[24] F. Tchakounté, K. Amadou Calvin, A. A. A. Ari, and D. J. Fotsa Mbogne, "A smart contract logic to reduce hoax propagation across social media," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, pp. 3070–3078, Jun. 2022, doi:10.1016/j.jksuci.2020.09.001.

[25] C. Moreno-Castro, E. Vengut-Climent, L. Cano-Orón, and I. Mendoza-Poudereux, "Exploratory study of the hoaxes spread via WhatsApp in Spain to prevent and/or cure COVID-19," Gaceta Sanitaria, vol. 35, no. 6, pp. 534–541, Nov. 2021, doi:10.1016/j.gaceta.2020.07.008.

[26] C. P. J. Sinaga and J. Yonatia, "Kampanye Penangkalan Hoax Melalui Aplikasi Gawai," Serat Rupa Journal of Design, vol. 2, no. 2, p. 119, Jul. 2018, doi: 10.28932/srjd.v2i2.805.

[27] M. Syaiful, M. Akbar, and T. Bahfiarti, "Analyze Fact Checking of Haram Sinovac Vaccine Hoax on Twitter Social Media Status," *International Journal of Science and Applied Science: Conference Series*, vol. 5, no. 1, p. 2021, 2021, doi: 10.20961/ijsascs.v5i1.62059.

[28] D. Fardiah, F. Darmawan, and R. Rinawati, "Fact-checking Literacy of Covid-19 Infodemic on Social Media in Indonesia," Komunikator, vol. 14, no. 1, pp. 14–29, May 2022, doi: 10.18196/jkm.14459.

[29] D.-H. Vu, "Privacy-preserving Naive Bayes classification in semi-fully distributed data model," Computers & Security, vol. 115, p. 102630, Apr. 2022, doi: 10.1016/j.cose.2022.102630.

[30] A. Nurdina and A. B. I. Puspita, "Naive Bayes and KNN for Airline Passenger Satisfaction Classification: Comparative Analysis," Journal of Information System Exploration and Research, vol. 1, no. 2, Jul. 2023, doi: 10.52465/joiser.v1i2.167.

[31] A. B. Yilmaz, Y. S. Taspinar, and M. Koklu, "Classification of Malicious Android Applications Using Naive Bayes and Support Vector Machine Algorithms," *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering* , vol. 10, no. 2, pp. 269–274, 2022, doi: 10.1039/b000000x.

[32] "Comparison of LSTM, SVM, and naive bayes for classifying sexual harassment tweets," Journal of Soft Computing Exploration, vol. 3, no. 2, Sep. 2022, doi: 10.52465/joscex.v3i2.85.

[33] Mussalimun, E. H. Khasby, G. I. Dzikrillah, and Muljono, "Comparison of K- N earest Neighbor (K -NN) and Naïve Bayes Algorithm for Sentiment Analysis on Google Play Store Textual Reviews," 2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), Sep. 2021, doi: 10.1109/icitacee53184.2021.9617217.

[34] A. Ali, W. Samara, D. Alhaddad, A. Ware, and O. A. Saraereh, "Human Activity and Motion Pattern Recognition within Indoor Environment Using Convolutional Neural Networks Clustering and Naive Bayes Classification Algorithms," Sensors, vol. 22, no. 3, p. 1016, Jan. 2022, doi: 10.3390/s22031016.

[35] C. Sirichanya and K. Kraisak, "Semantic data mining in the information age: A systematic review," International Journal of Intelligent Systems, vol. 36, no. 8, pp. 3880–3916, May 2021, doi:10.1002/int.22443.

[36] S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, "A Survey on Concept-Level Sentiment Analysis Techniques of Textual Data," 2021 IEEE World AI IoT Congress (AIIoT), May 2021, doi:10.1109/aiiot52608.2021.9454169.

[37] Y. Eroglu, "Text Mining Approach for Trend Tracking in Scientific Research: A Case Study on Forest Fire," Fire, vol. 6, no. 1, p. 33, Jan. 2023, doi: 10.3390/fire6010033.

[38] F. Gurcan and N. E. Cagiltay, "Research trends on distance learning: a text mining-based literature review from 2008 to 2018," Interactive

Learning Environments, vol. 31, no. 2, pp. 1007–1028, Sep. 2020, doi: 10.1080/10494820.2020.1815795.

[39] J. Park, D. Yang, and H. Y. Kim, "Text mining-based four-step framework for smart speaker product improvement and sales planning," Journal of Retailing and Consumer Services, vol. 71, p. 103186, Mar. 2023, doi: 10.1016/j.jretconser.2022.103186.

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

[41] S. Jancy Sickory Daisy and A. Rijuvana Begum, "Smart material to build mail spam filtering technique using Naive Bayes and MRF methodologies," Materials Today: Proceedings, vol. 47, pp. 446–452, 2021, doi: 10.1016/j.matpr.2021.04.630.

[42] H. Zhang, N. Cheng, Y. Zhang, and Z. Li, "Label flipping attacks against Naive Bayes on spam filtering systems," Applied Intelligence, vol. 51, no. 7, pp. 4503–4514, Jan. 2021, doi:10.1007/s10489-020-02086-4.

[43] M. Suh and M. Jeong, "Development of Bus Routes Reorganization Support Software Using the Naïve Bayes Classification Method," Sustainability, vol. 14, no. 8, p. 4400, Apr. 2022, doi:10.3390/su14084400.

[44] D. Ariadi and K. Fithriasari, "Indonesian News Classification Using Naive Bayesian Classification Method and Support Vector Machine With Confix Stripping Stemmer," *Jurnal Sains dan Seni ITS*, vol. 4, no. 2, pp. 2337–3520, 2015, doi: 10.12962/j23373520.v4i2.10966.

[45] H. Muhabatin, C. Prabowo, I. Ali, C. L. Rohmat, and D. R. Amalia, "Klasifikasi Berita Hoax Menggunakan Algoritma Naïve Bayes Berbasis PSO," Informatics for Educators and Professional : Journal of Informatics, vol. 5, no. 2, p. 156, Jun. 2021, doi:10.51211/itbi.v5i2.1531.