Sentiment Analysis of Hate Speech Against Presidential Candidates of the Republic of Indonesia in the 2024 Election Using BERT

Fahriza Rizky Amalia
Informatics Engineering, Vacational School
Universitas Logistik dan Bisnis Internasional
West-Java
Bandung, Indonesia
fahrizarizkyamalia@gmail.com

Nisa Hanum Harani
Informatics Engineering, Vacational School
Universitas Logistik dan Bisnis Internasional
Jawa Barat
Bandung, Indonesia
nisa@ulbi.ac.id

Cahyo Prianto
Informatics Engineering, Vacational School
Universitas Logistik dan Bisnis Internasional
Jawa Barat
Bandung, Indonesia
cahyo@ulbi.ac.id

Abstract— The issue of hate speech on social media has become a matter of growing concern, particularly in the context of political discourse, as evidenced by the 2024 elections in Indonesia. Online platforms such as YouTube represent a significant medium for political discourse, frequently accompanied by negative or hateful commentary directed towards presidential candidates. The objective of this study is to analyze the sentiment of YouTube comments related to Indonesian presidential candidates in the 2024 elections using the BERT algorithm. The data was obtained through scraping using the YouTube API and subsequently categorized into three distinct categories of hate speech: The categories of hate speech are as follows: OFP (offensive personal), OFG (offensive group), and OFO (offensive others). In this study, the CRISP-DM method was employed, which encompasses the following stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The findings indicate that the BERT algorithm is capable of accurately classifying comments. The algorithm can be utilized to develop predictive applications that assist in the identification and management of hate speech on social media platforms.

Keywords— *Hate Speech, Election, Social Media, BERT*

I. INTRODUCTION

The internet provides individuals with a convenient platform for exchanging opinions with one another, enhancing the efficacy of communication. Contemporary social media platforms epitomize the essence of free speech, offering users the opportunity to express their opinions without the constraints of surveillance or censorship [1]. As reported by We Are Social, by 2023, over 4.8 billion people worldwide will be using social media. This significant increase has led to a notable shift in the way individuals access information, with political discussions becoming a primary source of information for many. In Indonesia, for instance, over 170 million people use social media, making it one of the main platforms for political discussions and debates [2]. The rapid dissemination of information on social media facilitates the propagation of hate speech, which can reach a vast audience with minimal effort. During the 2024 election campaign in Indonesia, there was a notable surge in the prevalence of hate speech on social media, particularly in the context of political discourse. A joint study conducted by Monas University and the Alliance of Independent Journalists (AJI) Indonesia revealed that during this period, there were 182,118 social media posts containing hate speech[3]

Political figures are frequently the subject of discourse, and those who adhere to a particular politician's views can also readily remain apprised of that politician's activities through social media. The tenet of freedom of speech represents a form of democracy, wherein every citizen is accorded the right to defend their opinions. However, individuals who are ideologically polarized may exploit this right to disseminate hate speech when they seek to criticize other individuals who espouse disparate political perspectives [4].

The extant literature demonstrates that hate speech can intensify political polarization, heighten social tensions, and precipitate acts of violence in the tangible world. [5]. In Indonesia, the potential for hate speech to incite excessive negative actions in the real world is a significant concern. These actions may include incitement and calls for violence against civilians, as well as other criminal acts. In 2016, conservative Islamic groups organized a series of large-scale demonstrations in Jakarta in response to the circulation of derogatory and religiously oriented social media posts targeting Christian gubernatorial candidate Basuki "Ahok" Tjahaja Purnama. The demonstrators demanded that Ahok be imprisoned for insulting Islam. Consequently, Ahok was sentenced to two years in prison [6]. This case illustrates the necessity of protecting freedom of speech, including the right to express political views and engage in discussions and arguments. Hate speech on social media not only affects individuals but can also target political parties, potentially triggering wider social conflicts.

A multitude of regulations have been established with the objective of curbing the dissemination of hate speech on social media platforms. One such strategy entails the implementation of a policy that restricts the visibility of posts or comments containing hateful language on the user's timeline [7]. This approach is designed to diminish the prevalence of hate speech on these digital spaces. It is not uncommon for politicians to utilize the social media platform YouTube as a means of engaging with the public. This platform can be utilized as a tool to gain insight into the public perception of presidential candidates. The primary challenge in evaluating the sentiment of hate speech is the variability of language utilized by users. Consequently, hate speech is obscured by the use of ambiguous or contextualized language, which can vary depending on cultural and social context [8]. Hate speech lacks discernible characteristics, rendering it difficult to identify specific text elements that contain hateful words. Hate speech can be defined as a broad, emotional, and discriminatory concept that can negatively impact the well-being of individuals [9].

In this study, data were obtained from YouTube by scraping YouTube data related to the 2024 presidential candidates Anies Baswedan, Prabowo Subianto, and Ganjar Prabowo, along with the supporting parties of each candidate. The data was collected on two different occasions: June 22, 2024, and July 5, 2024. The number of data points collected on each occasion was 11,051 and 18,851, respectively. The data was then subjected to preprocessing before being classified into sentiment categories. These categories were divided into three groups: OFP (Offensive Person) comments targeting individuals such as Anies, Prabowo, and Ganjar; OFG (Offensive Group) comments targeting a group of people or including parties supporting presidential candidates; and OFO (Offensive Others) comments targeting other entities outside OFP and OFG. To supplement the data, a Wikipedia trend analysis was conducted on each presidential candidate account in the 2024 election on the Wikipedia.com website. This entailed a comparison of the number of readers of articles in Indonesian and English. Furthermore, social media activities on Facebook, Twitter, and YouTube for each presidential candidate were also analyzed using Web Analytic Fanpage Karma.

This research employs a deep learning-based approach utilizing the BERT (Bidirectional Encoder Representations from Transformers) algorithm to analyze the sentiment of hate speech. BERT is a pre-trained language model developed by Google that has demonstrated optimal performance in a range of natural language processing (NLP) tasks. As demonstrated by Devlin et al. (2019), BERT exhibits high accuracy in a range of NLP tasks, including sentiment analysis and hate speech detection [10]. In a separate study, Liu et al. (2021) similarly demonstrated the superiority of BERT in understanding complex sentence contexts, making it particularly suitable for text analysis on social media [11]. The principal advantage of BERT is its capacity to comprehend the context of words within a sentence. The model processes a word in a sentence based on

its relationship to the sentence as a whole. BERT processes the full context by examining patterns that appear before or after the word [10].

The objective of this research is to utilize sentiment analysis to anticipate the emergence of texts containing public hate speech directed at the 2024 presidential candidates, namely "Anies Baswedan," "Prabowo Subianto," and "Ganjar Prabowo," along with the supporting parties of each candidate. It is anticipated that this research will enhance the comprehension of the dynamics of hate speech that occurs on social media.

II. RESEARCH METHODS

This research employs the CRISP-DM research method, a data mining standardization framework developed by three pioneering figures in the data mining market. The CRISP-DM method was developed by Daimler Chrysler (formerly Daimler-Benz), SPSS (ISL), and NCR at various workshops from 1997 to 1999 [12]. CRISP-DM represents a more comprehensive and thoroughly documented alternative to existing data mining methodologies. The CRISP-DM methodology offers a standardized process for data mining that can be integrated into the problem-solving strategy of a business or research unit. Each phase is structured and clearly defined, facilitating ease of use [13]. Figure 1 depicts the six stages of this research methodology, which include business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each stage comprises distinct steps [14].

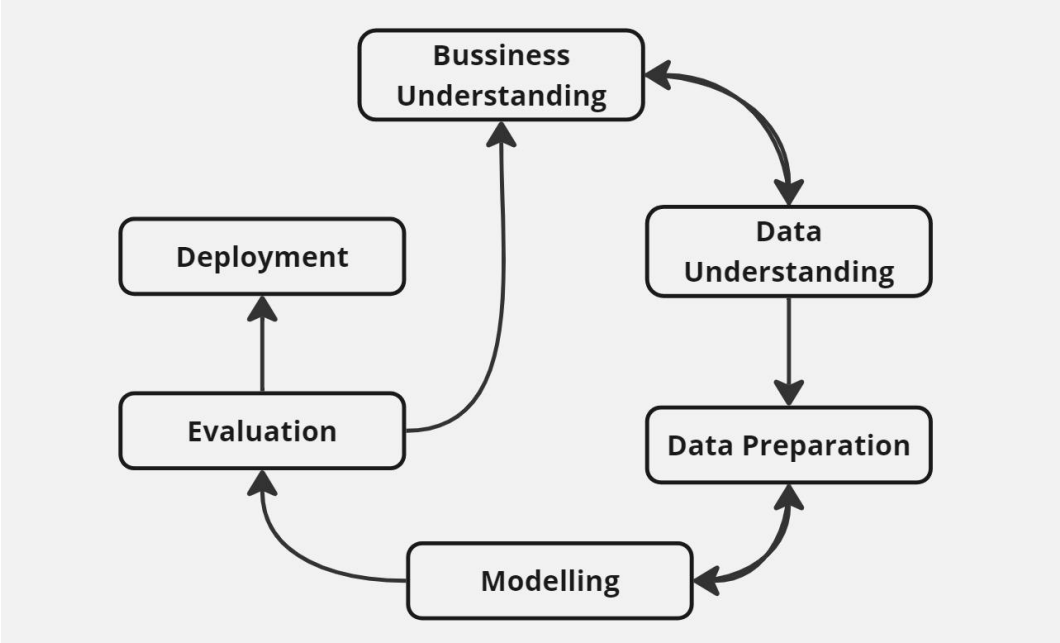


Figure 1 Research Methods

A. Business Understading

The initial stage of the data mining process is the Business Understanding stage. This stage is of paramount importance, as it establishes the foundation for the entire project. This stage is concerned with gaining an understanding of the underlying business goals and needs of the project [15]. The primary objective is to guarantee that all analytical endeavors align with the business requirements and that the ultimate outcome will confer substantial value. This stage entails the identification of the business problem that requires resolution. To this end, in-depth discussions are held with key stakeholders, with the objective of gaining insight into the business context, the challenges faced, and the

opportunities present. Once the business problem has been identified, the project objectives must be clearly and specifically defined. The project objectives must indicate the manner in which the data analysis will be employed to resolve the business problem. Another crucial step at this stage is the definition of project success criteria. These criteria are utilized to assess whether the project has successfully attained its objectives and delivered the anticipated value. In this research, the objectives are twofold: firstly, to analyze the sentiment of hate speech directed at the presidential candidates of the Republic of Indonesia in the 2024 election, and secondly, to define how the results of the analysis can be implemented to inform the public or policy makers.

B. Data Understanding

The objective of data understanding is to identify and comprehend the data that will be utilized in the data mining project. This stage encompasses the process of gathering data pertinent to the analysis, conducting exploratory analysis to identify potential patterns and issues, and examining the structure and quality of the data [16]. In this study, the variables employed encompass a range of key aspects pertaining to the sentiment analysis of YouTube comments pertaining to presidential candidates and their respective supporting parties in the 2024 election. The variables in question are as follows:

1) The Independent Variable
    - Social Media (YouTube): The data set was derived from comments on YouTube pertaining to the 2024 Election presidential debate video and the news of Anies, Prabowo, and Ganjar's candidacy. The data were collected on two different occasions: June 22, 2024, and July 5, 2024.
    - WikipediaTrend: The data were collected from the Wikipedia.com website, which is used to ascertain the number of articles viewed at a given time. The data was collected from January 1, 2024, to April 17, 2024.
    - Web Analytic Fanpage Karma: he data was collected from the official Facebook social media accounts of each presidential candidate. The data were collected between April 8, 2024, and July 7, 2024.
2) The Dependent Variable
    - Sentiment in Comments Sentiments from comments are classified into three categories based on the target of hate speech: The categories of offensive entities are as follows:
        - ✓ OFP (Offensive Person): Targets a specific individual, such as a presidential candidate.
        - ✓ OFG (Offensive Group): Targets groups or parties that support the candidate.
        - ✓ OFO (Offensive Other): argets other entities that do not fall under OFP or OFG.
3) The Control Variable
    - BERT Algorithm Model: The algorithm used to classify and analyze the sentiment of the collected comments.

C. Data Preparation

Data preparation is the process of transforming raw data into a format that is suitable for analysis. This stage is of great consequence, as the quality of the data utilized will impact the ultimate outcomes of the analysis. This stage encompasses a multitude of procedures, including data cleansing, data transformation (normalization, tokenization, and stopwords), filtering data based on keywords, providing positive, neutral, and negative sentiments, employing the NRC Lexicon, retrieving negative sentiments, and dividing the dataset into three categories (OFP, OFG, and OFO).

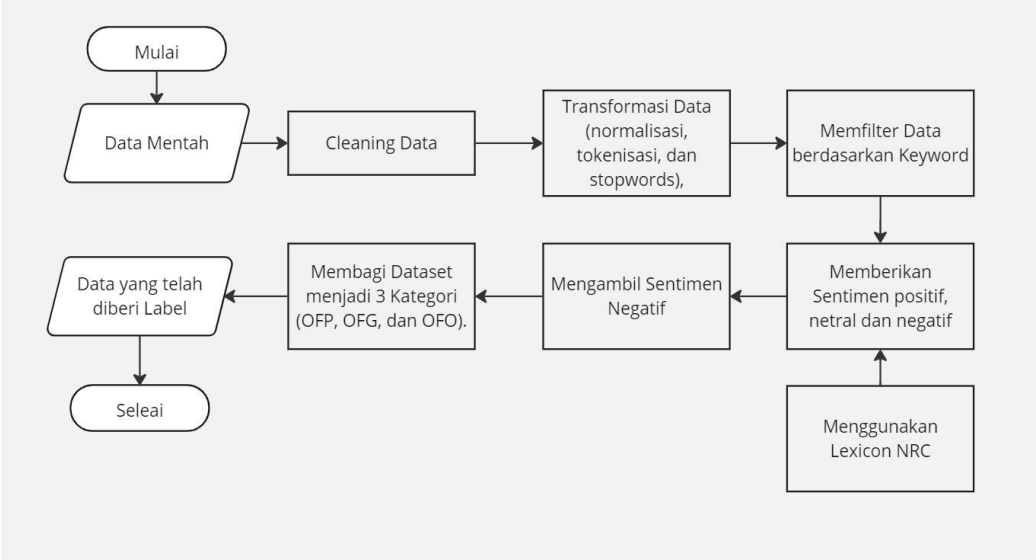


Figure 2 Preprocessing Data

The initial step is to undertake data cleansing, with the objective of ensuring that the data employed is free from extraneous elements and focused on pertinent information. This entails the removal of irrelevant words, such as URLs, hashtags, usernames, as well as special characters and emojis, from the dataset. Following the data cleaning process, each letter character in the dataset will be converted into lowercase letters in order to reduce unnecessary differences and facilitate a more consistent and accurate analysis. This will be done without changing the meaning of the original text.

Following the completion of data cleaning, the subsequent data transformation is divided into three distinct phases. Normalization is the process of modifying a collection of words that do not adhere to the standards set forth by the Big Indonesian Dictionary (KBBI) to align with standard Indonesian language conventions, including the elimination of abbreviations and the replacement of slang terms with their standard Indonesian equivalents. Subsequently, the data undergoes tokenization, whereby a single sentence is divided into its constituent words. In addition, numbers and symbols are also removed. The stopword process is employed to eliminate words that are not pertinent or significant, such as conjunctions.

The subsequent phase of the process entails the application of a keyword-based filter, with the objective of identifying content pertaining to the presidential candidates in the 2024 election. Subsequently, the data will be assigned positive, neutral, and negative sentiments through the use of the IndoBERT model, which has been trained using the Hugging Face library. IndoBERT was developed for the Indonesian language and is a pre-trained model using transformers that follows the BERTbase configuration. In the feature engineering process, IndoBERT employs a vectortex representation of tokens through the ids input and the last layer, or the closest layer, the dense layer, and then takes the cls token to be embedded. Following this, the training process is carried out to obtain positive, neutral, and negative sentiments [17].

The subsequent step is to undertake a comparison of the words contained within the data set with those listed in the NRC Lexicon. In this research, the data that will be subjected to further analysis is that which pertains to negative sentiment. Subsequently, the data is classified into three categories: OFP, OFG, and OFO. OFP (offensive person) texts target individuals, such as Anies, Prabowo, and Ganjar. OFG (offensive group) texts target a group of people or include parties that support presidential candidates. OFO (offensive other) texts target other entities outside OFP and OFG [7].

D. Modeling

The modeling stage involves the construction and evaluation of data analysis models to predict or classify data in accordance with the project objectives. This stage encompasses a number of crucial steps, including the selection of suitable modelling techniques, the construction and evaluation of models, and the assessment of model performance. This stage is of paramount importance to the overall success of the project. In this study, the BERT algorithm is employed for sentiment analysis of comment data, model training with prepared data, and model optimization to ensure accurate results. Previous studies have indicated that BERT is the optimal choice for detecting hate speech in Arabic, which suggests that pre-trained models can accurately identify hate speech in other languages as well [18].

The model will be constructed using BERT, with training conducted over three epochs and a training data and validation data division ratio of 60:40. Training is conducted for three epochs, as this is typically sufficient for BERT models [19].

The BERT model employs a training and validation data division with a ratio of 60:40, which is optimal for a medium-sized dataset. Allotting 60% of the data for training allows the model to learn effectively without compromising the amount of data necessary for evaluation. By allocating 40% of the data for validation, it is possible to obtain a more reliable estimation of the model's performance. Furthermore, the provision of additional data for validation facilitates more effective detection and prevention of overfitting. This is particularly crucial when intricate models such as BERT can be overfitted with constrained training data [20].

K-fold cross-validation is employed to guarantee that the model is evaluated on diverse subsets of data, rather than on a single data set. This process allows for the determination of the accuracy of the model by solving a model performance evaluation problem that divides the data into folds and ensures that each fold is used as a test set at multiple cross-validation points. The objective is to circumvent the phenomenon of overfitting, which arises in machine learning when the model exhibits an excessive degree of learning from the minutiae and noise present in the training data, thereby negatively impacting its performance on novel data. In other words, the model becomes overly specialized to the training data, thereby losing the capacity to generalize to new data. Furthermore, K-fold cross-validation is employed to guarantee that the model is evaluated on disparate subsets of data, rather than on a single data set. This process allows for the determination of the accuracy of the model through the resolution of a model performance evaluation problem, whereby the data is divided into folds and each fold is utilized as a test set at multiple cross-validation points. The objective is to circumvent the phenomenon of overfitting, which arises in machine learning when the model exhibits an excessive degree of adaptation to the training data, thereby impairing its capacity to generalize effectively to novel inputs.

The data set is divided into K folds, with each fold serving as a test set. In this study, a five-fold cross-validation approach was employed, with the data set divided into five parts. In the initial iteration, the first section was utilized for model testing, while the last section was designated for model training. In the subsequent iteration, the second section served as the testing set, and the last section was used as the training set. This process was repeated until every fold of the five folds was used as a testing set [21].

E. Evaluation

The evaluation stage is designed to ascertain whether the model in question meets the requisite business objectives and success criteria. This stage is of particular importance, as it serves to confirm whether the model is ready to be deployed in an operational environment or requires further refinement. The present research will conduct an evaluation to ensure that the BERT model is capable of categorizing comments based on sentiment and hate speech categories (OFP, OFG, and OFO). In addition, validation

metrics such as accuracy, precision, recall, and F1 metrics can be utilized to evaluate the modeling results. [22].

In order to ascertain the degree of accuracy of the positive model and the proportion of data classified as positive, precision is employed. This can be expressed through the following equation:

$$\text{Precision} = \frac{tp}{tp+fp} \quad (1)$$

The effectiveness of a model in identifying all positive instances within a dataset, including instances that are incorrectly classified as negative, is quantified by the recall measure. This can be expressed through the following equation:

$$\text{Recall} = \frac{tp}{tp+fn} \quad (2)$$

The F-1 score is employed to achieve an equilibrium between precision and recall. The F-1 value can be expressed as follows:

$$\text{F1} = 2\text{x}\frac{PxR}{P+R} \quad (3)$$

Accuracy is an evaluation metric used to measure the proportion of correct predictions out of all predictions made by the model, including both positive and negative correct predictions. The accuracy of a model can be expressed using the following equation:

$$\text{Accurancy} = \frac{(tp+tn)}{(tp+fp+tn+fn)} \quad (4)$$

Description: tp = true positive
tn = true negative
fp = false positive
fn = false negative

F. Devloyment

The final stage of the process is the deployment of the model to an operational environment, where it is made available for business use. This stage encompasses the creation of a deployment plan, the integration of the model into existing systems, the monitoring and maintenance of the model, and the provision of training and documentation for end users. In this study, the Bidirectional Encoder Representations from Transformers (BERT) model is employed to develop a system capable of analyzing and predicting comment data, with the objective of categorizing comments based on the sentiment associated with specific hate speech categories (OFP, OFG, and OFO).

The system will be constructed using the Python library Gradio, which has been developed with the specific purpose of creating interactive demonstrations of machine learning models. The Gradio platform facilitates the conversion of intricate machine learning models into straightforward and user-friendly web applications, enabling anyone to interact with the pre-trained model. Gradio offers the advantage of being applicable to a wide range of machine learning models, including those in the domains of natural language processing (NLP), computer vision, and more complex models. Additionally, it provides a multitude of user interface (UI) components, such as text input, file upload, plotting, and more, which are ready to use. Furthermore, Gradio enables the rapid transformation of a model into a web application that can be accessed through a browser and supports deployment to diverse platforms, including web, mobile, and cloud.

III. RESULTS AND DISCUSSION

A. Business Understading

The primary issue addressed in this study is the role of social media, particularly YouTube, in the dissemination of hate speech related to politics in Indonesia during the 2024 elections. Political figures are frequently the subject of discussion on social media, where followers can readily monitor and engage in discourse regarding the activities and developments of their preferred politicians. While freedom of speech is a fundamental tenet of a democratic system, individuals with strongly held political views or those engaged in political polarization frequently utilize these platforms to disseminate hate speech directed at those with differing political perspectives.

This dynamic of disseminating hate speech not only affects individuals, but can also exacerbate political polarization, increase social tensions, and precipitate violent acts in the physical world. The 2016 demonstrations in response to social media posts related to the Ahok case provide a clear illustration of how hate speech can incite excessive negative actions [6].

The objective of this study is to examine the role of social media, particularly YouTube comments, in the dissemination of hate speech directed at presidential candidates and their respective supporters during the 2024 election. The BERT algorithm will be used to identify patterns of hate speech, categorizing comments based on hate speech categories (OFP, OFG, and OFO), and to evaluate the performance of the algorithm. It is hoped that this research will provide insights into how the dynamics of hate speech spreading on social media affect the reputation of candidates, political parties, as well as overall political stability in Indonesia, and help develop strategies to mitigate its negative impact.

B. Data Understanding

The data presented here has been extracted from the YouTube social media platform via the use of the YouTube API. Table 1 represents the outcome of the data collection process, which has been organised according to the date of data collection. The data has been divided into two distinct datasets, namely Dataset 1 and Dataset 2.

Table 1 Total Youtube Comment Data

	Total Data	Data Retrieval
Dataset 1	11.0851	22 Juni 2024
Dataset 2	18.851	05 Juli 2024

1) WikipediaTrend

In this study, the Wikipedia Trend library, which can be accessed through the Python programming language, is employed to ascertain the number of articles on the Wikipedia.com website that are viewed within a specified time frame. Furthermore, this library is employed as a point of comparison prior to the mining of data from YouTube regarding online behavior related to Indonesian presidential candidates [23].

To ascertain the extent of comparison between readers of each candidate, data was gathered from January 1, 2024 to April 17, 2024, encompassing articles in Indonesian and English. Table 2 illustrates the quantity of data collected for each candidate.

Table 2 WikipediaTrend Results for each Candidate

Candidate Name	Language	Number of Viewers
Anies Baswedan	Id	863.422
	En	172.464
Prabowo Subianto	Id	1.575.040
	En	503.193
Ganjar Pranowo	Id	485.138
	Eg	103.681

The subsequent graph illustrates a comparative analysis of the candidates with the most comprehensive profile information.

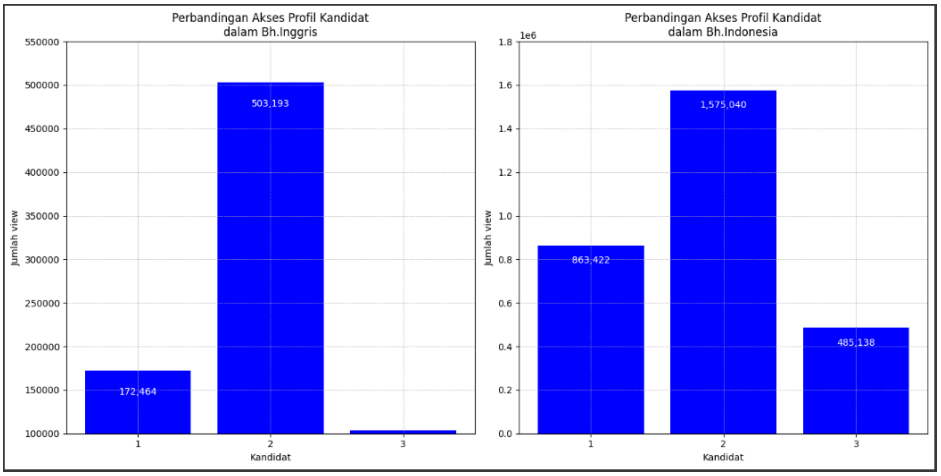


Figure 3 Comparison of WikipediaTrend for each Candidate

Figure 3 illustrates that the profile of candidate 02, Prabowo Subianto, was accessed significantly more frequently on the Indonesian language page, with a total of 1,575,040 visitors, compared to 863,422 for candidate 01, Anies Baswedan, and 485,138 for candidate 03, Ganjar Pranowo. In comparison, the English page on candidate 02 Prabowo Subianto's profile was also more accessed, with 172,464 visitors, compared to candidate 01 Anies Baswedan and candidate 03 Ganjar Pranowo, with 172,464 and 103,681, respectively. Based on the number of visitors, it can be concluded that the three presidential candidates are the subject of considerable domestic discourse, as evidenced by the high number of visitors to the Indonesian pages.

2) Presidential Candidate Account Analysisc

Following an examination of the popularity trends of the candidates in question, based on the data provided by WikipediaTrends, we will proceed to present an overview of the official accounts on Facebook, the social media platform. This analysis employs the Fanpage Karma web analytics application, which is capable of examining the social media account activity

of each candidate account. The results of the Fanpage Karma web analysis, accessed from April 8, 2024 to July 7, 2014, are presented below:

Table 3 Facebook Account Web Analytics Results for each Candidate

Categories	Candidate 01 Anies Baswedan	Candidate 02 Prabowo Subianto	Candidate 03 Ganjar Pranowo
Fans	2.1 M	10.9 M	2.4 M
Engagement	0.31%	0.044%	0.17%
Follower Growth Weekly	-0.31%	-0.011%	0.24%
Post Interaction	0.13%	0.074%	0.21%
Post per Day	2.3	0.6	0.8

The results of the Facebook Web Analytics for each candidate account in Table 3 indicate that candidate 02 Prabowo Subianto has a higher score than candidate 01 Anies Baswedan and candidate 03 Ganjar Praowo in almost every category. These results demonstrate that candidate 02 Prabowo Subianto had a greater number of interactions, conversations, and followers than the other candidates. Candidate 02 effectively illustrated that the utilization of Facebook as a social media platform can facilitate engagement with the public.

C. Data Preparation

The retrieved data is in its original, unprocessed form, containing elements that are not required when conducting further analysis. Therefore, it is necessary to prepare the data for further processing. The initial stage of data processing is data cleaning, which involves the removal of irrelevant words and symbols. This includes URLs, hashtags, usernames, as well as special characters and emojis. Once the data has been cleaned, it is converted into lowercase letters or case folding as a whole. The next data transformation is normalization, which is the process of normalizing the words in the dataset. Tokenization is then carried out, which separates sentences into individual words. Stopwords are removed from the dataset, as they are not important or have no special meaning. This is done by utilizing the NLTK corpus and creating a stopword dictionary manually. The results of data preparation on the YouTube comment dataset are presented below:

	textDisplay	cleaned_comment	normalized_text	tokenized_text	final_text
0	Ganjarukem dan anisa soryyy...yyeee..	ganjarukem dan anisa soryyyyyeee	ganjar dan anies maaf	[ganjar, dan, anies, maaf]	ganjar anies
1	Anis cerdas sekali berbicara	anis cerdas sekali berbicara	anies cerdas sekali berbicara	[anies, cerdas, sekali, berbicara]	anies cerdas berbicara
2	Nanya tentang pertahanan ke menteri pertahanan...	nanya tentang pertahanan ke menteri pertahanan...	tanya tentang pertahanan ke menteri pertahanan...	[tanya, tentang, pertahanan, ke, menteri, pert...	tanya pertahanan menteri pertahanan tentara in...
3	Inilah yang membuag banyak orang memilih Prabo...	inilah yang membuag banyak orang memilih prabo...	inilah yang membuat banyak orang memilih prabo...	[inilah, yang, membuat, banyak, orang, memilih...	memilih prabowo terbukti mengambang berkali de...
4	Siapapun presidennya\nKita tetap jadi kuli dun...	siapapun presidennya\nkita tetap jadi kuli dunia	siapapun presidennya kita tetap jadi kuli dunia	[siapapun, presidennya, kita, tetap, jadi, kul...	presidennya kuli dunia

Figure 3 Results of Data Preparation Dataset1

	textDisplay	cleaned_comment	normalized_text	tokenized_text	final_text
0	Kasihan banget..	kasihan banget	kasihan banget	[kasihan, banget]	
1	Tukang puisi	tukang puisi	suka puisi	[suka, puisi]	puisi
2	Pa Anies dijadikan pencuci nama baik dari olok...	pa anies dijadikan pencuci nama baik dari olok...	pak anies dijadikan pencuci nama baik dari olo...	[pak, anies, dijadikan, pencuci, nama, baik, d...	anies pencuci baik olok olok metrotipu metrotv...
3	Kita sdh terlalu kenyang dengan kata2, bravo p...	kita sdh terlalu kenyang dengan kata2 bravo pa...	kita sudah terlalu kenyang dengan kata-kata ba...	[kita, sudah, terlalu, kenyang, dengan, kataka...	kenyang katakata bagus anies
4	Emang ente udah punya partai pendukung Nies..m...	emang ente udah punya partai pendukung niesmod...	emang kamu sudah punya partai pendukung anies ...	[emang, kamu, sudah, punya, partai, pendukung,...	partai pendukung anies modal nasdem pks demokr...

Figure 4 Results of Data Preparation Dataset 2

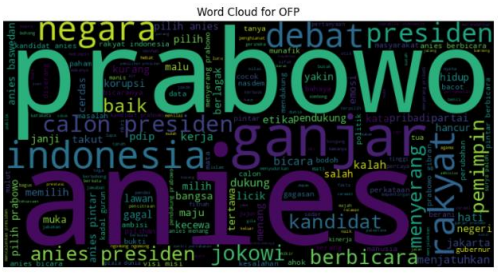
Once the data has been cleaned, the subsequent step is to compare the words in the dataset with the lexicon that will be utilized. The data employed in this analysis is that contained in dataset 1. The results of the analysis conducted using the NRC lexicon for each candidate are as follows:

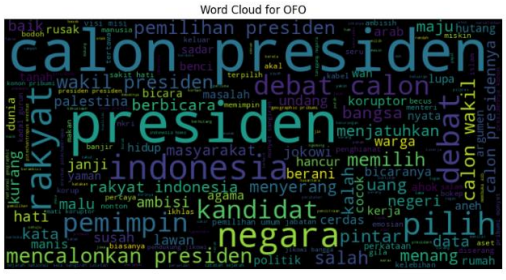
Table 4 NRC Lexicon Sentiment Analysis Results

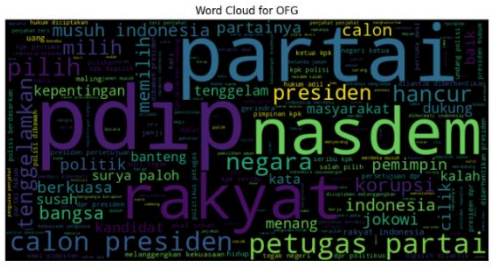
No	Sentiment	Anies Baswedan Score	Prabowo Subianto Score	Ganjar Pranowo Score
1	Anger/Marah	920	1108	425
2	Anticipation/antisipasi	1053	1258	545
3	Disgust/Menjijikan	536	523	217
4	Fear/Taku	1137	1367	527
5	Joy/Gembira	839	922	425
6	Negative	1813	2029	798
7	Positive	3589	4154	1821
8	Sadness/Sedih	520	596	230
9	Surprise/Kejutan	690	818	361
10	Trust/Percaya	2041	2313	1009

An examination of the polarity between the positive and negative sentiments associated with each candidate, as illustrated in Table 4, reveals that candidate 02, Prabowo Subianto, has received a more balanced range of positive and negative sentiments. This finding suggests that candidate 02 is more favourably regarded by netizens on YouTube social media.

In this research, the data that will be processed to the next stage is negative sentiment data. Subsequently, datasets 1 and 2 are merged into a single CSV file and categorized into three categories: OFP, OFG, and OFO. The following section presents the word cloud visualization for each category:







:)

Figure 5 Wordcloud Visualization

D. Modeling

The data modeling stage is conducted using the BERT (Bidirectional Encoder Representations from Transformers) algorithm. Prior to commencing the modeling stage, the data preparation process is undertaken to construct a model. This involves the partitioning of the data into three distinct categories: training data, validation data, and test data.

```
# Menggunakan kolom final_text sebagai fitur dan label sebagai target
X = dataset['final_text']
y = dataset['label']

# Membagi dataset menjadi data pelatihan dan data sementara
X_train, X_temp, y_train, y_temp = train_test_split(X, y, test_size=0.4, random_state=42, stratify=y)

# Membagi data sementara menjadi data validasi dan data pengujian
X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42, stratify=y_temp)
```

Figure 6 Split Data

Figure 6 illustrates a script for data sharing. Variable x is utilized to store feature data from the dataset, specifically within the final_text column. Variable y is employed to store target data, whereby this label denotes the class or category of each data point.

The function train_test_split(X, y, test_size=0.4, random_state=42, stratify=y) is used to divide a dataset into two parts: training data (train) and temporary data (temp). By determining the proportion of data that will be designated as temporary data (40%), the remaining 60% will become training data. Subsequently, a random seed value must be identified to guarantee that the division result remains consistent each time the script is executed. Additionally, it is essential to ensure that the proportion of classes represented in the training data and temporary data aligns with the proportion of classes present in the original dataset. This approach is beneficial in preventing bias in data sharing, particularly when the dataset is not balanced.

The script train_test_split(X_temp, y_temp, test_size=0.5, random_state=42, stratify=y_temp) is employed for the purpose of dividing the temporary data set into two distinct subsets: namely, the validation data set and the test data set. The value of test_size=0.5 serves to determine the proportion of the temporary data set that will be utilized as the test data set, which is set at 50%. The remaining 50% of the data set will be designated as the validation data set.

The training data is employed to educate the model, which will discern patterns from the training data to make predictions. While the validation data is utilized to assess the performance of the model during the training process, this data can prevent overfitting. The testing data is used to evaluate the final performance of the model on data that the model has never seen before. The sum of the split data results is shown in Figure 7.

```
Ukuran data pelatihan: 2111
Ukuran data validasi: 704
Ukuran data pengujian: 704
```

Figure 7 Dataset Split Result

In order to prevent overfitting and to obtain more accurate performance estimates, cross-validation techniques such as K-fold cross-validation are employed during the modelling process, as illustrated in Figure 8.

```
# Mengonversi label menjadi tipe integer setelah memastikan distribusi yang benar
label_mapping = {'OFP': 0, 'OFO': 1, 'OFG': 2}
```

Figure 8 Converting to Integer

As illustrated in Figure 8, the script transforms each category into an integer format. Following the conversion of the data to integer, cross-validation techniques such as K-fold cross-validation are employed during modeling to prevent overfitting and to enhance the precision of performance estimates. This is demonstrated in Figure 9.

```
# Menggunakan StratifiedKFold
kf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

train_scores = []
test_scores = []

for fold, (train_index, val_index) in enumerate(kf.split(X, y)):
    print(f'Fold {fold + 1}')
    print('-' * 10)

    X_train, X_val = X.iloc[train_index], X.iloc[val_index]
    y_train, y_val = y.iloc[train_index], y.iloc[val_index]

# Mendefinisikan epochs
    epochs = 3

    # Scheduler untuk pembelajaran bertahap
    total_steps = len(train_loader) * epochs
    scheduler = get_linear_schedule_with_warmup(optimizer, num_warmup_steps=0, num_training_steps=total_steps)

    # Melatih model
    for epoch in range(epochs):
        print(f'Epoch {epoch + 1}/{epochs}')
        print('-' * 10)

        train_acc, train_loss = train_epoch(model, train_loader, optimizer, device, scheduler)
        print(f'Train loss {train_loss} accuracy {train_acc}')

        val_acc, val_loss = eval_model(model, val_loader, device)
        print(f'Validation loss {val_loss} accuracy {val_acc}')
```

Figure 9 Training the Model Using K-fold Cross Validation

Figure 9 illustrates the code for creating a StratifiedKFold object with five folds (data partitions). The shuffle parameter is set to True, which randomizes the data while it is divided into folds. The random_state parameter is set to 42 to ensure reproducibility of the randomization results. Subsequently, two empty lists, train_scorse and test_scores, are created to store metric values (such as accuracy) at each fold for training and validation data.

Subsequently, a loop is initiated for each fold, wherein the enumerate() function is employed to ascertain the fold index and the training and validation data indexes. The kf.split(X,y) function performs the division of the x (feature) and y (label) data into training and validation data, in accordance with the specified fold.

Subsequently, the x feature data should be divided into x_train training data and x_val validation data, based on the index obtained from kf.split(). The y label data should then be divided into y_train training data and y_val validation data, based on the same index. The number of epochs in model training should then be determined, and the total steps in training should be calculated with the help of total_steps and scheduler, using get_linear_schedule_with_warmup, in order to set the learning rate gradually.

The training loop trains the model for a specified number of epochs, creating a function to train the model at a single epoch. This function returns the loss and accuracy metrics. The train_epoch function is used to train the model on a specified number of epochs, and the eval_model function is used to evaluate the model on validation data, returning the loss and accuracy metrics.

E. Evalualition

At this juncture, the objective is to evaluate the model's performance based on the outcomes of the preceding process. The outcomes of training the model using cross-validation or K-fold cross-validation with k=5 are illustrated in Figure 10.

```
Accuracy on the test set: 0.9730113636363636
Precision: 0.9755480826689312
Recall: 0.9730113636363636
F1 Score: 0.9729208100240178
```

Figure 10 Results of Training the Model with K-fold Cross Validation

The overall result depicted in Figure 10 represents the culmination of a process involving the division of the data into multiple subsets, the training of a model on each subset, the evaluation of the model on a separate validation data set, and the storage of the resulting metric value for each subset. In order to prevent overfitting, the data is divided into several folds, with the model trained on each fold. This allows the model to be more generalized and provides a more accurate performance estimate. The average metric value of all folds is used to evaluate the model's performance, as this provides a more accurate picture of the model's performance. Based on the chart, the model evaluation results for accuracy, precision, recall, and F1-score are 97%.

Figure 11 depicts the metric table of the confusion matrix test results, which are utilized to assess the performance of the classification model. The confusion matrix table illustrates the number of correct and incorrect predictions made by the model for each class. In other words, this table compares the actual value (the actual class) with the predicted value (the class predicted by the model).

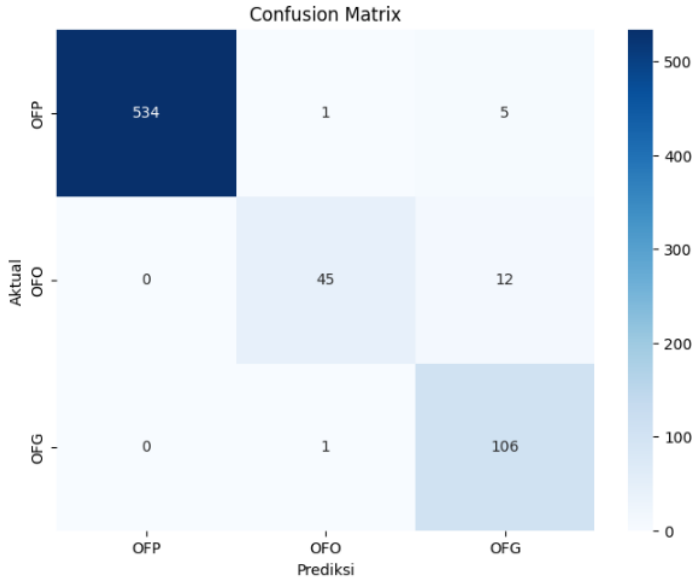


Figure 11 Confusion Matrix

The X-axis (prediction) depicts the classification predicted by the model (OFP, OFO, OFG), while the Y-axis (actual) illustrates the actual or true classification of the data (OFP, OFO, OFG). Each cell in the table represents the quantity of data that falls into a specific combination of actual and predicted class. The main diagonal depicts the number of data points correctly classified. A value of 534 signifies that 534 data points with the actual class designation OFP were correctly identified as such by the model. In contrast, the values outside the main diagonal represent the number of data points incorrectly classified. A value of 45 indicates that 45 data points with the actual class designation OFO were incorrectly predicted as OFP.

As evidenced by the confusion matrix, the model demonstrates a notable capacity for classifying the OFP class, though it exhibits less proficiency in categorizing the OFO and OFG classes. The confusion matrix outcomes will be employed to assess the model's

performance by calculating accuracy, precision, recall, and F1-score, as illustrated in Figure 12.

	precision	recall	f1-score	support
0	1.00	0.99	0.99	540
1	0.96	0.79	0.87	57
2	0.86	0.99	0.92	107
accuracy			0.97	704
macro avg	0.94	0.92	0.93	704
weighted avg	0.98	0.97	0.97	704

Figure 12 Metric Evaluation Results

Figure 12 illustrates that class 0 or OFP exhibits optimal performance, with precision, recall, and F1-score values approaching 1. This indicates that the model is highly effective in classifying data accurately as OFP. In class 1 or OFO, the model exhibits a relatively high precision but a low recall. This indicates that the model is proficient at identifying data belonging to the OFO class when predicting OFO class, yet it fails to correctly identify some actual OFO instances. In contrast, the model demonstrates a relatively high precision and an excellent recall in class 2 or OFG. This indicates that the model is highly effective in identifying data belonging to the OFG class and exhibits minimal misclassification of other data as OFG. Overall, the classification model exhibits commendable performance, particularly for the OFP and OFG classes. However, there is scope for enhancement in the OFO class, particularly with regard to recall.

F. Devloyment

Once the modeling process is complete, the next step is to apply the model in order to facilitate the prediction of comment data. This will allow the model to be utilized in categorizing comments based on the sentiment associated with the hate speech categories (OFP, OFG, and OFO). The implementation of this approach will result in the creation of a new sentence prediction website, which will be developed using the Python library Gradio. The web view of this application is illustrated in Figure 13.

Figure 13 Comment Prediction Web View

Figure 13 illustrates the new text box, which serves as a field for users to input the comment text they wish to test. Users may enter comments directly into the column. The output text box displays the model's prediction results. If the comment contains hate speech, the prediction label (OFP, OFG, or OFO) will be displayed.

IV. CONCLUSION

The findings of the analysis demonstrate that this research has effectively identified patterns of hate speech in YouTube comments pertaining to presidential candidates and their respective political parties. The detection of negative sentiments suggests the existence of a notable polarization among supporters and opponents of the candidates. This result is consistent with the research objective, which was to analyze sentiment and categorize comments based on hate speech categories (OFP, OFG, and OFO). The BERT algorithm demonstrated efficacy in classifying comments into the three hate speech categories. The OFP (Offensive Personal) and OFG (Offensive Group) categories demonstrate high precision, recall, and F1-score levels, indicating that the model is capable of accurately identifying hate speech within those two categories. However, there is room for improvement in the OFO (Offensive Others) category, especially in terms of recall, which indicates that the model occasionally fails to identify comments that should be classified as such. The implementation of the model in the form of a web-based application demonstrates significant potential for facilitating the automatic detection of hate speech. Users can readily test new comments through the provided interface and receive immediate prediction results, indicating whether the comments fall into the OFP, OFG, or OFO categories.

REFERENCES

[1] Y. Chen, H. Sack, and M. Alam, "Analyzing social media for measuring public attitudes toward controversies and their driving factors: a case study of migration," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, Dec. 2022, doi: 10.1007/s13278-022-00915-7.

[2] W. A. Social, "Digital 2023: Global Overview Report," 2023. [Online]. Available: https://wearesocial.com/

[3] N. Muhamad, "Twitter, Medsos dengan Ujaran Kebencian Terbanyak pada Kampanye Pemilu 2024," Katadata. [Online]. Available: https://databoks.katadata.co.id/datapublishembed/167538/twitter-medsos-dengan-ujaran-kebencian-terbanyak-pada-kampanye-pemilu-2024

[4] Y. Tang and N. Dalzell, "Classifying Hate Speech Using a Two-Layer Model," *Stat. Public Policy*, vol. 6, no. 1, pp. 80–86, Jan. 2019, doi: 10.1080/2330443X.2019.1660285.

[5] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*, vol. 6, pp. 13825–13835, Feb. 2018, doi: 10.1109/ACCESS.2018.2806394.

[6] M. H. Siregar, "Riset: Ujaran Kebencian Terhadap Capres Meningkat di Media Sosial Jelang Pemilu 2024," The Conversation. [Online]. Available: https://theconversation.com/riset-ujaran-kebencian-terhadap-capres-meningkat-di-media-sosial-jelang-pemilu-2024-222060

[7] J. M. Molero, J. Perez-Martin, A. Rodrigo, and A. Penas, "Offensive Language Detection in Spanish Social Media: Testing from Bag-of-Words to Transformers Models," *IEEE Access*, vol. 11, pp. 95639–95652, 2023, doi: 10.1109/ACCESS.2023.3310244.

[8] K. Florio, V. Basile, M. Polignano, P. Basile, and V. Patti, "Time of your hate: The challenge of time in hate speech detection on social media," *Appl. Sci.*, vol. 10, no. 12, Jun. 2020, doi: 10.3390/APP10124180.

[9] S. Gite *et al.*, "Textual Feature Extraction Using Ant Colony Optimization for Hate Speech Classification," *Big Data Cogn. Comput.*, vol. 7, no. 1, Mar. 2023, doi: 10.3390/bdcc7010045.

[10] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor

[11] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[12] I. Budiman, T. Prahasto, and Y. Christyono, "Data Clustering Menggunakan Metodologi CRISP-DM Untuk Pengenalan Pola Proporsi Pelaksanaan Tridharma," *J. Sist. Inf. Bisnis*, vol. 1, no. 3, pp. 15–16, 2014, doi: 10.21456/vol1iss3pp129-134.

[13] A. P. Fadillah, "Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ)," *J. Tek. Inform. dan Sist. Inf.*, vol. 1, no. 3, pp. 260–270, 2015, doi: 10.28932/jutisi.v1i3.406.

[14] A. Rianti, N. W. A. Majid, and A. Fauzi, "CRISP-DM: Metodologi Proyek Data Science," *Pros. Semin. Nas. Teknol. ...*, pp. 107–114, 2023, [Online]. Available: http://ojs.udb.ac.id/index.php/Senatib/article/view/3015

[15] F. N. Dhewayani, D. Amelia, D. N. Alifah, B. N. Sari, and M. Jajuli, "Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM," *J. Teknol. dan Inf.*, vol. 12, no. 1, pp. 64–77, 2022, doi: 10.34010/jati.v12i1.6674.

[16] D. Kurniawan and M. Yasir, "Optimization Sentimen Analysis using CRISP-DM and Naive Bayes Methods Implemented on Social Media," *Cybersp. J. Pendidik. Teknol. Inf.*, vol. 6, no. 2, p. 74, 2022, doi: 10.22373/cj.v6i2.12793.

[17] D. Alita and A. R. Isnain, "Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier," *J. Komputasi*, vol. 8, no. 2, pp. 50–58, 2020, doi: 10.23960/komputasi.v8i2.2615.

[18] Z. Boulouard, M. Ouaissa, M. Ouaissa, M. Krichen, M. Almutiq, and K. Gasmi, "Detecting Hateful and Offensive Speech in Arabic Social Media Using Transfer Learning," *Appl. Sci.*, vol. 12, no. 24, 2022, doi: 10.3390/app122412823.

[19] M. I. Amal, E. S. Rahmasita, E. Suryaputra, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Terhadap Isu Kebocoran Data Kartu Identitas Ponsel di Twitter," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 3, pp. 645–660, 2022, doi: 10.28932/jutisi.v8i3.5483.

[20] K. K. Dobbin and R. M. Simon, "Optimally splitting cases for training and testing high dimensional classifiers," *BMC Med. Genomics*, vol. 4, 2011, doi: 10.1186/1755-8794-4-31.

[21] A. Peryanto, A. Yudhana, and R. Umar, "Klasifikasi Citra Menggunakan Convolutional Neural Network dan K Fold Cross Validation," *J. Appl. Informatics Comput.*, vol. 4, no. 1, pp. 45–51, 2020, doi: 10.30871/jaic.v4i1.2017.

[22] N. Putu, V. D. Saraswati, N. Yudistira, and P. P. Adikara, "Analisis Sentimen terhadap Perundungan Siber pada Twitter menggunakan Algoritma Bidirectional Encoder Representations from Transformer (BERT)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 2, pp. 909–916, 2023, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/12345

[23] C. Prianto, N. H. Harani, and I. Firmansyah, "Analisis Sentimen Terhadap Kandidat Presiden Republik Indonesia Pada Pemilu 2019 di Media Sosial Twitter," *J. Media Inform. Budidarma*, vol. 3, no. 4, p. 405, 2019, doi: 10.30865/mib.v3i4.1549.