Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

Analysis of Public Opinion Sentiment towards the 2024 Presidential Election Based on Clustering Method with K-Means Algorithm

Mhd Anshor Harahap[1*], Muhammad Ikhsan[2]
[1,2] Computer Science Study Program, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[1]anshorharahap90@gmail.com,[2]mhdikhsan@uinsu.ac.id

open access

***Corresponding Author**

Article History:
Submitted: 15-10-2024
Accepted: 28-10-2024
Published: 04-11-2024

Keywords:
Sentiment Analysis; Clustering Method; K-Means Algorithm

ABSTRACT

The presence of social media, such as Twitter, Facebook and Instagram, provides a space for people to express their opinions freely and openly. Various sentiments, ranging from support to criticism of the candidates, work programs, and other political issues, have emerged along with the increasing public enthusiasm. Therefore, it is important to understand how public opinion is evolving and what is the main focus of public attention in the 2024 presidential election. The purpose of this research is to analyze the sentiment and views of the public about the presidential election using the Clustering approach and the K-Means method and to classify public opinion for various interests as well as optimizing social media information for the public interest. Based on the research conducted, the K-Means algorithm was successfully applied for sentiment analysis of public opinion on the 2024 presidential election, using tweet data taken through crawling Twitter as many as 220 tweets. From the dataset, 5 tweets were used for manual implementation of the K-Means algorithm calculation, through a series of pre-processing processes, including TF-IDF weighting. After the manual K-Means calculation, from 29 words generated from TF-IDF, the following clustering results were obtained: Cluster 0 (positive) contains 5 words, Cluster 1 (neutral) contains 18 words, and Cluster 2 (negative) contains 6 words. These results show that the K-Means algorithm can effectively cluster sentiment in public opinion data related to the 2024 presidential election based on patterns found in the words in the tweets.

INTRODUCTION

The 2024 presidential election is one of the most anticipated political events in Indonesia. As the largest democratic contestation in the country, this presidential election not only attracts public attention, but also triggers intensive discussions on various platforms, both conventional and social media. The presence of social media, such as Twitter, Facebook, and Instagram, provided space for people to express their opinions freely and openly. Various sentiments, ranging from support to criticism of candidates, work programs, and other political issues, emerged along with the increasing public enthusiasm. Therefore, it is important to understand how public opinion is evolving and what is the main focus of public attention in the 2024 presidential election.

Reporting from Katadata Media Network, the General Election Commission (KPU) determined the Permanent Voters List (DPT) for the General Election as many as 204,807,222 voters. The determination of the DPT was carried out through the Open Plenary Meeting of the Recapitulation of the National Level Permanent Voters List (DPT) for the 2024 Election, at the KPU building, Sunday (2/7/2023). Launching from Republika, based on the results of the DPT recapitulation, the majority of voters in the 2024 Election are dominated by generation Z and millennial groups. "A total of 66,822,389. or 33.60% of voters from the millennial generation," said KPU RI Commissioner Betty Epsilon Idroos at the DPT Recapitulation Open Plenary Meeting at the KPU office, Jakarta, Sunday (2/7/2023). The millennial generation is a term for people born from 1980 to 1994. Meanwhile, voters from generation Z are 46,800,161 voters or 22.85% of the total DPT for the 2024 elections.

Sentiment analysis of public opinion related to the 2024 presidential election is significant in identifying patterns, trends, and general feelings spread across various digital platforms. This information can provide insights for political observers, prospective leaders, and campaign teams in understanding the needs and concerns of the community more deeply. In addition, the results of this research can be used as input in designing more effective political communication strategies, as well as as consideration for election organizers to create a more conducive and transparent political atmosphere. By using sentiment analysis methods based on machine learning, it is expected to produce more accurate and objective opinion mapping, thus providing a comprehensive picture of the public's views on the 2024 Presidential Election.

To achieve a more structured opinion mapping, this research will apply a Clustering method based on the K-Means algorithm. The K-Means algorithm can group public opinion into several clusters based on similar characteristics and expressed sentiments, making it easier to identify the main issues that are developing. Through this

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

method, public opinion can be grouped into clusters such as support, criticism, and neutral towards the 2024 presidential election, as well as specific topics of public concern.

Research like this has been researched by several experts, such as that studied by (Faesal et al., 2020) entitled "Sentiment Analysis on Twitter User Tweet Data Against Online Store Sales Products Using the K-Means Method". In this study using the K-Means Method. The K-Means method is used for sentiment analysis on twitter user tweet data on online store sales products.

There is also something that has been researched by (Melati & Reza, 2024) entitled "Sentiment Analysis of Twitter Data Using the K-Means Clustering Method in the Case Study of Moving the Capital City of the Archipelago (IKN)". In this study using the K-Means Clustering method. The K-Means Clustering method is used for sentiment analysis of twitter data in the case study of moving the Archipelago Capital (IKN).

Previous research conducted focused on applying the K-Means algorithm to analyze consumer sentiment towards online store products, with the aim of grouping consumer opinions based on satisfaction with products and services. In contrast to these studies, this research focuses on analyzing the sentiment of public opinion towards the 2024 Presidential Election, which is a more complex and dynamic political phenomenon. The novelty proposed in this research is to apply the K-Means algorithm to perform clustering on public opinions collected from various social media platforms to identify emerging political issues and trends in support for presidential candidates. By adopting a different political context and topic, this research seeks to make a new contribution in the application of the K-Means algorithm in the realm of political sentiment analysis, which is different from previous studies that focus more on consumer opinions in a business context. It is hoped that this research can be an answer to the classification of job suitability searches.

LITERATURE REVIEW

Text Mining

Text mining is mining done by computers to get something new, something that is not known before or find back scattered information, which comes from information extracted automatically from different text data sources (Khan et al., 2020). Text mining is a technique used to handle classification, clustering, information extraction and information retrival problems. The stages of text mining in general are text preprocessing and feature selection (Joergensen E Munthe et al., 2022).

Sentiment Analysis

Sentiment analysis is one part of text mining studies which can be called computational studies to classify a person's opinions, emotions and attitudes towards entities (Hudaya, Fakhrurroja and Alamsyah, 2019). Sentiment analysis is also known as opinion mining or emotion artificial intelligence which is useful for processing natural language, text analysis and linguistic computing to identify, extract, calculate and study information in a structured manner (Pratama, Andrean and Nugroho, 2019). Sentiment analysis can be called the same as opinion mining because this method focuses on opinions that are positive, neutral or negative (Samsir et al., 2021). Sentiment analysis is widely used by researchers as a branch of research in the field of computer science. This sentiment analysis method extracts an opinion data by understanding and processing textual data automatically so that the sentiment contained in the opinion can be seen. Sentiment analysis can also be distinguished based on the data source. There are several levels that are most often used in sentiment analysis research, including at the document level and at the sentence level. Sentiment analysis is divided into 2 major groups, including Coarse-gained and Fine-gained (Pertiwi, 2019).

Coarse-gained sentiment analysis is an analysis that is only carried out on documents that broadly considers the entire contents of the document as a positive sentiment II-2 and negative sentiment. Fine-gained sentiment analysis is an analysis carried out only at the sentence level. The main focus is to determine sentiment in sentences only (Ardiani, Sujaini and Tursina, 2020). The application of sentiment analysis is usually used in various things such as consumer information, marketing, politics and social. In government or health and others, this sentiment analysis can be used to find out people's opinions on an issue that is happening so that the government can make the right solution based on the data that has been collected (Savitri et al., 2021).

X/Twitter

X/Twitter is an online social networking and microblogging service that allows users to send and read text-based messages of up to 140 characters but on November 07, 2017 increased to 280 characters known as tweets. Twitter is a micro-blogging site that is very popular in Indonesia. This can be seen from the number of Twitter users reaching 19.5 million users out of a total of 330 million users in the world. Twitter was founded in March 2006 by Jack Dorsey, and the social networking site was launched in July. Since its launch, Twitter has become one of the ten most visited sites on the Internet, and has been dubbed the "short message of the Internet."

According to the We Are Social report, there are around 27.5 million Twitter users aka X in Indonesia as of October 2023. That number puts Indonesia in fourth place globally. The United States is perched at the top of the world with 108.55 million Twitter/X users, followed by Japan and India with 74.1 million and 30.3 million users respectively.Under Indonesia, there is the UK which has 24.3 million Twitter/X users, after which there is Brazil with

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

24.15 million users, Turkey 22.75 million, Mexico 19.6 million, Saudi Arabia 17.9 million, and Thailand 16.2 million users. The high popularity of Twitter means that this service can be utilized as a means of necessity from various aspects, such as a means of expressing opinions, conveying aspirations, emergency communication media, political campaigns, and learning tools.

Presidential Election

As an event organized by the Indonesian nation every 5 years, the presidential election is always a highly anticipated people's party because from this event we as the people can hear the visions and missions of the sons and daughters of the nation. However, in a country that adheres to presidential democracy, the position of the President is very important, in addition to being the head of State as well as the head of government. The failure of the President can result in the democratic system itself failing to be implemented in practice. Because of the importance of the office of the president, the way in which he is chosen is important. This is because it will definitely affect the level of political effectiveness of the elected President (Putri Yolanda & Halim, 2020).

The filling of the positions of President and Vice President elected by the People's Consultative Assembly with a majority vote has been felt to be less democratic. Thus, there is a desire for direct election of the President and Vice President by the people (Article 6A paragraph (1) of the Third Amendment to the 1945 Constitution). In addition to encouraging people's participation in exercising their political rights, the system of direct election of the President is also seen as a democratic mechanism because it better represents the will of the people.

Clustering

Clustering is an important data grouping method to understand. This is part of data mining or data mining, which is the extraction of interesting patterns from large amounts of data. Clustering is also often interpreted as the process of grouping data into several clusters so that the data in a cluster has maximum similarity. According to (Noviyanto, 2020) Clustering refers to grouping documents, observations or cases in classes with similar objects. A cluster is a collection of documents that are similar to each other and different from documents in other clusters. Clustering is different from Clasification, in Clustering there is no target variable to be grouped. Clustering algorithms try to divide the data set into clusters whose members are relatively similar, where the similarity of documents in the same cluster is high, and the similarity of documents in other clusters is small.

K-Means

Data Clustering is one of the unsupervised data mining methods. There are two types of data clustering that are often used in the data grouping process, namely hierarchical data clustering and non-hierarchical data clustering. K-Means is one of the non-hierarchical data clustering methods that seeks to partition existing data into the form of one or more clusters/groups (Makarychev & Shan, 2022).

This method partitions data into clusters/groups so that data that has the same characteristics is grouped into the same cluster and data that has different characteristics is grouped into other groups. The purpose of this data clustering is to minimize the objective function set in the clustering process, which generally seeks to minimize variation within a cluster and maximize variation between clusters. The benefits of Clustering are as Object Identification (Recognition) for example in the fields of Image Processing, Computer Vision or robot vision. In addition, it is a Decision Support System and Data Mining such as market segmentation, regional mapping, marketing management etc. (Makarychev & Shan, 2022).

Data clustering using the K-Means method is generally carried out with the following basic algorithm (Faesal et al., 2020):

1. Determine the number of clusters
2. Randomly allocate data into clusters
3. Calculate the centroid/average of the data in each cluster
4. Allocate each data to the nearest centroid/average
5. Return to Step 3, if there is still data that moves clusters or if the change in centroid value is above the specified threshold value or if the change in the value of the objective function used is above the specified threshold value.

K-means Characteristics:

1. K-means is very fast in the clustering process.
2. K-means is very sensitive to random initial centroid generation.
3. It is possible for a cluster to have no members.
4. K-means clustering results are unique (always changing, sometimes good, sometimes bad).

Distance Space To Calculate the Distance Between Data and Centroid

Several distance spaces have been implemented in calculating distance (distance between data and centroid) including L1 (Manhattan/City Block) distance space, L2 (Euclidean) distance space, and Lp (Minkowski) distance space. The distance between two points x1 and x2 in Manhattan/City Block distance space is calculated using the following formula (Nascimento, 2022):

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

$$D_{L_1}(x_2, x_1) = \|x_2 - x_1\|_1 = \sum_{j=1}^{p} |x_{2j} - x_{1j}|$$

As for the L2 (Euclidean) distance space, the distance between two points is calculated using the between two points is calculated using the following formula:

$$D_{L_2}(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^{p} (x_{2j} - x_{1j})^2}$$

Where:

D_{L2} = squared Eucliden distance between object x2 and x1.
P = number of cluster variables.
x_{2j} = value or data from the 2nd object in the jth variable.
x_{1j} = value or data of the 1st object in the jth variable (Everitt, 1993).

Word2Vec

The word embeddings method named word2vec was developed by Mikolov in 2013. Word2vec is a set of several interrelated models used to generate word embeddings. Word embeddings is the name of a set of modeling languages and feature learning techniques in Natural Language Processing (NLP) where each word of the vocabulary has a vector that represents the meaning of the word and the words are mapped into a vector of real numbers (Kurniawan & Dr. Warih Maharani, S.T., 2022).

Word2Vec represents words in vectors that can carry the semantic meaning of words. This word embedding model is an unsupervised learning application that uses a neural network consisting of a hidden layer and a fully connected layer. The dimension of the weight matrix for each layer is the number of words in the corpus (collection of text) multiplied by the number of hidden cells/neurons in the hidden layer. The weight matrix in the hidden layer of the trained model is used to transform words into vectors. This weight matrix is like a lookup table, where each row represents each word and the column represents a vector of words (Nurdin, 2020).

This word2vec method consists of two main word embeddings algorithms, namely: continuous bag of word (CBOW) and skip-gram. The CBOW algorithm is used to see the specific length of a word in the input document. While the skip-gram algorithm is used to predict the context of a word by looking at the proximity of a word to other words that are positioned before or after the word. Architecturally, word2vector is actually just an artificial neural network that does not have many hidden layers, both in terms of nodes in each layer and the number of layers (Prabowo et al., 2019).

METHOD

The place and time of this research was conducted at the State Islamic University of North Sumatra. The relevance and urgency of collecting public opinion data with a leatherative method for better understanding and handling. This research utilizes the Waterfall method as an approach. The decision to use this type of method is based on the belief that Waterfall is a research method that produces a specific product and tests the extent to which the product is effective. In this research, the research methodology is a guideline in conducting research so that what is achieved does not deviate from the predetermined objectives. The following is an outline of the steps in this research:

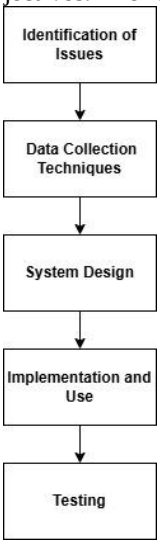

Figure 1. Diagram of Research Method

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

The data in this study were obtained through observation and documentation, which were then used as the basis for application development. In this research effort to obtain comprehensive and precise data, the author approaches the method with web scrapping or web crawling on twitter social media to obtain user tweets. Design the K-Means algorithm flowchart for grouping public opinion on the 2024 presidential election, the flowchart has been designed to see the K-Means Algorithm process in grouping public opinion on the 2024 presidential election, the following is below the K-Means Algorithm florchar image:

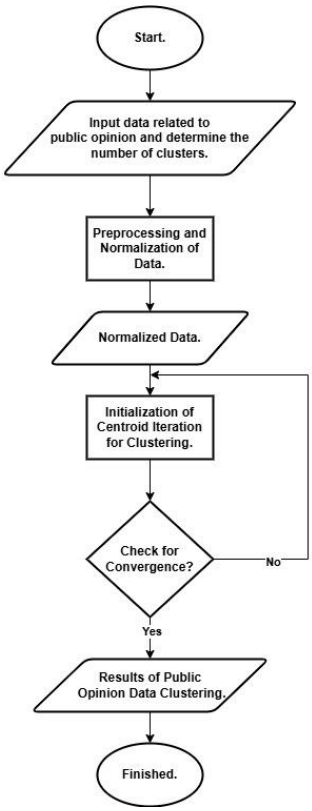

Figure 2. Flowchart of K-Means Algorithm

RESULT

DATA ANALYSIS

This research aims to analyze the sentiment of public opinion towards the 2024 presidential election through data taken from Twitter. In the initial stage, user sentiment data was collected using a Python library, which allows researchers to extract data from Twitter automatically. This data collection was done without time limitation, thus covering all tweets related to the 2024 presidential election since the platform began to be used. This approach allowed the research to cover a wide range of perspectives and opinions, covering a long period of time and a variety of socio-political contexts that influence people's sentiments.

For data analysis, the K-Means algorithm was used due to its effective ability to group text data into clusters. This algorithm is able to identify sentiment patterns based on people's tweets, be it positive, negative, or neutral sentiments, related to the 2024 presidential election. In addition, K-Means offers flexibility in grouping similar sentiments into one cluster, so researchers can more easily analyze how public opinion develops over time. The results of this clustering are expected to provide deeper insights into public perceptions and opinions, as well as assist policy makers or related parties in understanding trends and public attitudes towards the 2024 Presidential Election.

The data analysis steps in this study begin with the collection of tweet data using a Python library to access data from Twitter, followed by pre-processing to clean the data from irrelevant elements, such as punctuation, numbers, URLs, and stop words. After that, text conversion into numerical representation using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) is performed so that the algorithm can process it. Next, the K-Means algorithm is applied to group tweets into clusters based on sentiment similarity, such as positive, negative, or neutral. After the clusters are formed, the results of this clustering are analyzed to identify trends and patterns of public sentiment towards the 2024 presidential election. The final results are then visualized using graphs or diagrams to facilitate interpretation and conclusion making.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

Data Representation

In the research of sentiment analysis of public opinion on the 2024 presidential election using the K-Means algorithm, using 220 crawled tweets from Twitter, data representation plays a crucial role in preparing the information needed for accurate analysis. This stage involves the process of converting tweet text data into a numerical or vector form that can be processed by the K-Means algorithm. Techniques such as TF-IDF are often used to ensure each tweet can be effectively represented based on the words used. Proper data representation is essential to obtain accurate clustering results and help the K-Means algorithm better categorize sentiment, so that trends and patterns of public opinion related to the 2024 Presidential Election can be clearly identified.

Table 1. Dataset Crawling

No.	Tweet
1	@Hasbil_Lbs @basuki_btp @aniesbaswedan Mas Wowo and Mas Gibran bring different experiences and expectations to the presidential race.
2	I don't know why this 2024 presidential election fell to gemoy
3	@tvOneNews pity pak prabowo diapusi tok ama survey institute 3x presidential election
...	...
220	It seems like this year is the last year we Indonesians will have a democratic presidential election.

In table 1 Dataset Crawling, data representation is performed starting with tokenization, where each tweet is broken down into small units such as words or phrases. After that, text cleaning is performed to remove irrelevant elements, such as links and icon emots that do not support sentiment analysis. Stop words are also removed to improve the precision of the analysis, by removing common words that have no informative value. Next, the data was converted into feature vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) method, which helps assess how important a word is in the context of sentiment related to the 2024 presidential election. These TF-IDF weights provide a more accurate representation of the frequency of word occurrence, allowing the K-Means algorithm to build an effective model for clustering tweet sentiment. This data representation process not only strengthens the understanding of public opinion, but also shows how this technique can be applied in sentiment analysis on dynamic platforms such as Twitter.

Data Pre-processing

The crawled data needs to go through a pre-processing stage before it can be analyzed. Pre-processing aims to transform datasets that are initially unstructured and contain a lot of noise into clean and ready-to-process data. This stage includes several steps, such as case folding to convert all text into lowercase letters, cleansing to remove irrelevant elements such as punctuation marks or symbols, tokenizing to break the text into words, slang conversion to replace informal words with standard words, filtering to remove unimportant words, and Stemming which returns words to their basic form. All of these steps are designed to ensure the data is in optimal condition before it is analyzed further.

Text Cleaning

Here are the steps of cleaning text data in brief:

1. Convert to lowercase: All text is converted to lowercase for consistency.
2. Removing URLs: Removes the link from the text.
3. Removing mentions and hashtags: Delete @username and #tag.
4. Removes numbers: Clears the text of unnecessary numbers.
5. Remove punctuation: Removes punctuation marks such as periods, commas, etc.
6. Remove extra spaces: Removes redundant spaces between words.

Table 2. Text Cleaning

No.	Tweet	Cleaning Tweet
1	@Hasbil_Lbs @basuki_btp @aniesbaswedan Mas Wowo and Mas Gibran bring different experiences and expectations to the presidential race.	mas wowo and mas gibran bring different experiences and expectations to the presidential race
2	I don't know why this 2024 presidential election fell to gemoy	I don't know why this presidential election fell to Gemoy.
3	@tvOneNews pity pak prabowo diapusi tok ama survey institute 3x presidential election	pity pak prabowo diapusi tok ama survey institute x presidential election
...	...	...
220	It looks like this year will be the last year that we in Indonesia will have a democratic presidential election.	It looks like this year is the last year we will have a democratic presidential election in Indonesia.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

Folding & Slang Words

The 2nd step is Folding & Slang Words, which aims to unify different forms of words that have the same meaning (folding) and improve the use of informal or slang words into more standardized words. Here are the steps taken for your data:

1 Folding: Equalizing variations of a word form into one standardized form, such as "I" into "me," "really" into "very."
2 Slang Replacement: Replacing slang or informal words with standardized word forms. For example, "bgt" becomes "really," "ga" becomes "no," and "wkwk" is removed because it does not provide meaning.
3 Foreign Language Word Substitutions (if any): Replace commonly used foreign words in Indonesian with more formal words, such as "glad" with "happy."

Table 3. Folding & Slang Words

No.	Cleaning Tweet	Folding & Slang Words
1	mas wowo and mas gibran bring different experiences and expectations to the presidential race	mas wowo and mas gibran bring different experiences and expectations to the presidential race
2	I don't know why this presidential election fell to Gemoy.	don't know why this presidential election fell to gemoy.
3	pity pak prabowo diapusi tok ama survey institute x presidential election	pity pak prabowo diapusi tok ama survey institute x presidential election
...	...	...
220	It looks like this year is the last year we will have a democratic presidential election in Indonesia.	It looks like this year is the last year we will have a democratic presidential election in Indonesia.

Tokenization

The third step is Tokenization, which means breaking the text into smaller units such as words or tokens. Each word in the sentence will be separated so that it can be further analyzed, for example to calculate frequency or used in machine learning models.

1. Breaking Sentences into Words (Tokens): Each tweet is separated by spaces or punctuation marks that separate the words.
2. Removing unnecessary punctuation marks: Punctuation marks such as periods, commas, exclamation marks are removed to leave only meaningful words.
3. Rough Tokenization (Word Tokenization): No special groupings or phrases, each word is considered a separate unit.

Table 4. Tokenization of

No.	Folding & Slang Words	Tokenization
1	mas wowo and mas gibran bring different experiences and expectations to the presidential race	['mas', 'wowo', 'dan', 'mas', 'gibran', 'bring', 'experience', 'dan', 'hope', 'yang', 'different', 'in', 'bursa', 'election', 'president']
2	don't know why this presidential election fell to gemoy.	['not', 'know', 'why', 'election', 'president', 'this', 'fall', 'heart', 'to', 'gemoy']
3	pity pak prabowo diapusi tok ama survey institute x presidential election	['pity', 'pak', 'prabowo', 'diapusi', 'tok', 'ama', 'institution', 'survey', 'x', 'election', 'president']
...	...	...
220	It looks like this year is the last year we will have a democratic presidential election in Indonesia.	['seems', 'year', 'this', 'is', 'year', 'last', 'we', 'country', 'indonesia', 'with', 'election', 'president', 'in', 'democracy']

Stopword Removal

The fourth step is Stopwords Removal. Stopwords are common words that usually have no important meaning in text analysis, such as "which", "and", "at", "this", and so on. Removing stopwords helps to reduce irrelevant words so that the focus of the analysis is more on meaningful words.

1. Stopwords List: Prepare a list of common stopwords in Indonesian and English.
2. Removing Stopwords: Eliminate all the words in the stopwords list from the previous tokenization result.
3. Remaining Words: After the stopwords are removed, only important words that are more relevant for analysis will remain.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

Table 5. Removal of Stopwords

No.	Tokenization	Stopword Removal
1	['mas', 'wowo', 'dan', 'mas', 'gibran', 'bring', 'experience', 'dan', 'hope', 'yang', 'different', 'in', 'bursa', 'election', 'president']	['mas', 'wowo', 'mas', 'gibran', 'bring', 'experience', 'hope', 'different', 'exchange', 'election', 'president']
2	['not', 'know', 'why', 'election', 'president', 'this', 'fall', 'heart', 'to', 'gemoy']	['know', 'election', 'president', 'fall', 'heart', 'gemoy']
3	['pity', 'pak', 'prabowo', 'diapusi', 'tok', 'ama', 'institution', 'survey', 'x', 'election', 'president']	['pity', 'prabowo', 'diapusi', 'tok', 'ama', 'institute', 'survey', 'election', 'president']
...	...	...
220	['seems', 'year', 'this', 'is', 'year', 'last', 'we', 'country', 'indonesia', 'with', 'election', 'president', 'in', 'democracy']	['seems', 'year', 'last', 'country', 'indonesia', 'election', 'president', 'democracy']

Stemming

The next step is Stemming. Stemming is the process of converting words to their base or root form. For example, the word "election" is changed to "select", "issue" is changed to "exit", and so on.

1. Identify bound words: Separate words that are bound forms (such as affixes) to their root words.
2. Use a Stemming algorithm: In Indonesian, Porter's algorithm or Nazief & Adriani algorithm is often used for Stemming.
3. Check the result: Make sure the transformed word is really the corresponding base form.

Table 6. Stemming

No.	Stopword Removal	Stemming
1	['mas', 'wowo', 'mas', 'gibran', 'bring', 'experience', 'hope', 'different', 'exchange', 'election', 'president']	['mas', 'wowo', 'mas', 'gibran', 'bawa', 'pengalaman', 'harap', 'beda', 'bursa', 'pilih', 'presiden']
2	['know', 'election', 'president', 'fall', 'heart', 'gemoy']	['know', 'choose', 'president', 'fall', 'heart', 'gemoy']
3	['pity', 'prabowo', 'diapusi', 'tok', 'ama', 'institute', 'survey', 'election', 'president']	['pity', 'prabowo', 'apus', 'tok', 'ama', 'institute', 'survey', 'choose', 'president']
...	...	...
220	['seems', 'year', 'years', 'last', 'country', 'indonesia', 'election', 'president', 'democracy']	["year", "last", "indonesia", "election", "president", "democracy"]

TF-IDF Weighting

After going through the stage of labeling and cleaning sentiment data, the next step is weighting using the TF-IDF (Term Frequency-Inverse Document Frequency) method. At this stage, each word or term in the document will be calculated its frequency (TF), which is how often the word appears in the document. Then, the value will be multiplied by IDF (Inverse Document Frequency), which measures how common or rare the word appears throughout the document. The combination of TF and IDF gives higher weight to words that are specific and relevant to the document, and lowers the weight of words that appear frequently in various documents but are less meaningful. For example, from the 220 tweets analyzed, there are sample TF and DF (Document Frequency) values taken from 5 tweets to show how this weighting is done. This TF-IDF approach aims to identify the most significant key words in documents, making it easier to analyze or predict sentiment effectively.

Table 7. Sample Data

No.	Sentiment
1	['mas', 'wowo', 'mas', 'gibran', 'bawa', 'pengalaman', 'harap', 'beda', 'bursa', 'pilih', 'presiden']
2	['know', 'choose', 'president', 'fall', 'heart', 'gemoy']
3	['pity', 'prabowo', 'apus', 'tok', 'ama', 'institute', 'survey', 'choose', 'president']
4	['pilpres', 'choose', 'president', 'pildun', 'choose', 'world']
5	['choose', 'president', 'year', 'why', 'bring', 'deg', 'deg', 'an']

The first step is TF calculation, here is a formula to determine the TF value of each word.

$$\text{TF} = \frac{\text{Jumlah kemunculan term dalam kalimat}}{\text{Jumlah total kata dalam kalimat}}$$

For the first sentence: ['mas', 'wowo', 'mas', 'gibran', 'bawa', 'pengalaman', 'harap', 'beda', 'bursa', 'pilih', 'presiden']

a. Number of words in the sentence = 11

b. Number of occurrences of the word "mas" = 2

$$\text{TF}_{\text{mas}} = \frac{2}{11} = 0.1818$$

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

Here are the overall results in the table:

Table 8. TF Value

Term	D1 (TF)	D2 (TF)	D3 (TF)	D4 (TF)	D5 (TF)
mas	0.1818	0	0	0	0
wowo	0.0909	0	0	0	0
gibran	0.0909	0	0	0	0
...	...	...	...	...	...
funny	0	0.125	0	0	0

The first step of IDF calculation, Here is a formula to determine the IDF value of each word. $\text{IDF} = \log\left(\frac{\text{D}+1}{\text{df}+1}\right) + 1$. The word "mas" appears in 5 out of 5 sentences: $\text{IDF} = \log\left(\frac{5+1}{1+1}\right) + 1 = \log(3) + 1 = 2.0986$. Here are the overall results in the table:

Table 9. IDF value

Term	df	IDF
Mas	1	2.0986
Wowo	1	2.0986
Gibran	1	2.0986
...	...	...
Funny	1	2.0986

The next step will calculate the TF-IDF for the word "mas" in each document: $\text{TF} - \text{IDF} = \text{TF} \cdot \text{IDF}$ maka $\text{TF} - \text{IDF}_{\text{mas}} = 0.1818 \text{ x } 2.0986 = 0.3813$
Here are the overall results in the table:

Table 10. TF-IDF value

Term	D1	D2	D3	D4	D5
mas	0.3813	0	0	0	0
wowo	0.189	0	0	0	0
gibran	0.189	0	0	0	0
...	...	...	...	...	...
funny	0	0.125	0	0	0

The next step is normalization using Min-Max Scaling, Min-Max Scaling Formula: $\text{Normalisasi} = \frac{\text{X}-\text{min}}{\text{max}-\text{min}}$ then $\text{Normalisasi} = \frac{0.3813-0}{0.3813-0} = 1$
Here are the overall results in the table:

Table 11. Data Normalization

Term	D1	D2	D3	D4	D5
mas	1	0	0	0	0
wowo	0.4955	0	0	0	0
gibran	0.4955	0	0	0	0
heart	0	0.4286	0	0	0
funny	0	0.4286	0	0	0

K-Means Implementation

The implementation of K-Means Clustering in the sentiment analysis of public opinion on the 2024 Presidential Election includes several stages. First, collect public opinion data, for example from social media or surveys, then preprocess the data such as text cleaning, tokenization, stopwords removal, and Stemming. After that, convert the text data into numerical representation using methods such as TF-IDF. Determine the appropriate number of clusters (k), then apply the K-Means algorithm to cluster the data based on sentiment patterns (positive, negative, or neutral). Evaluate the clustering results with metrics such as inertia, to ensure that the clusters formed can optimally represent people's sentiments.

Table 12. Data Set

Term	D1	D2	D3	D4	D5
Mas	1	0	0	0	0
Wowo	0.4955	0	0	0	0
Gibran	0.4955	0	0	0	0
...	...	...	...	...	...
Funny	0	0.4286	0	0	0

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

After determining the dataset, it is necessary to determine the number of clusters to be formed. The clusters that will be formed include: Cluster 0 (C0) = Positive, Cluster 1 (C1) = Neutral, Cluster 2 (C2) = Negative. Determining the initial centroid center C randomly from the above dataset, 3 clusters and centroid centers were selected:

Table 13. Initial Centroid

Initial Centroid					
Cluster (C0)	0.2381	0.4286	0.7143	0.6667	0.5556
Cluster (C1)	0	0	0	0	0.5556
Cluster (C2)	0	0	1	0	0

Allocate all data/objects into the nearest cluster. Here are the results of data allocation to cluster distances. The results of the distance to the cluster are obtained from the calculation with the formula:

$$d_{x,y} = \sqrt{\sum_{i=1}^{n} (Xi - Yi)^2}$$

$$d(1,0) = \sqrt{(1-0.2381)^2 + (0-0.4286)^2 + (0-0.7143)^2 + (0-0.6667)^2} \sqrt{+(0-0.5556)^2} = 1.423936203$$

$$d(1,1) = \sqrt{(1-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0.5556)^2} = 1.143980489$$

.....

$$d(5,2) = \sqrt{(0.4955-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 + (0-0)^2} = 1.116028785$$

After doing the calculation, the overall result of the distance to the cluster is obtained as follows:

Table 14. Calculation of Distance to cluster

Term	D1	D2	D3	D4	D5	Distance to Cluster			Results Cluster
						C0	**C1**	**C2**	
mas	1	0	0	0	0	1.423936	1.14398	1.414214	Cluster 1 (Neutral)
wowo	0.4955	0	0	0	0	1.230186	0.744454	1.116029	Cluster 1 (Neutral)
gibran	0.4955	0	0	0	0	1.230186	0.744454	1.116029	Cluster 1 (Neutral)
...	...	...	...	...	...	...	...	...	...
funny	0	0.4286	0	0	0	1.148954	0.701705	1.087979	Cluster 1 (Neutral)

Redetermine the new centroid center point based on the average. The average value in question is the starting point of the centroid that can be obtained then the new centroid is obtained from the formula = result value / many results.

The first step calculates the average in cluster 0 as follows:

Table 15. Cluster 0

Term	D1	D2	D3	D4	D5	Cluster Result
select	0.2381	0.4286	0.7143	0.6667	0.5556	Cluster 0 (Positive)
president	0.2381	0.4286	0.7143	0.6667	0.5556	Cluster 0 (Positive)
election	0	0	0	1	0	Cluster 0 (Positive)
pildun	0	0	0	1	0	Cluster 0 (Positive)
world	0	0	0	1	0.6667	Cluster 0 (Positive)

Cluster 0 (D1) = (0.2381 + 0.2381 + 0 + 0 + 0) / 5 = 0.09524
Cluster 0 (D2) = (0.4286 + 0.4286 + 0 + 0 + 0) / 5 = 0.17144
Cluster 0 (D3) = (0.7143 + 0.7143 + 0 + 0 + 0) / 5 = 0.28572
Cluster 0 (D4) = (0.6667 + 0.6667 + 1 + 1 + 1) / 5 = 0.86668
Cluster 0 (D5) = (0.5556 + 0.5556 + 0 + 0 + 0.6667) / 5 = 0.35558

Then calculate the average in cluster 1 as follows:

Table 16. Cluster 1

Term	D1	D2	D3	D4	D5	Cluster Result
mas	1	0	0	0	0	Cluster 1 (Neutral)
wowo	0.4955	0	0	0	0	Cluster 1 (Neutral)
gibran	0.4955	0	0	0	0	Cluster 1 (Neutral)
...	...	...	...	...	...	...
funny	0	0.4286	0	0	0	Cluster 1 (Neutral)

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

Cluster 1 (D1) = (1 + 0.4955 + 0.4955 + 0.4955 + 0.4955 + 0.4955 + 0.4955 + 0.4955 + 0 + 0 + 0 + 0 + 0 + 0 + 0) / 18 = 0.38897
Cluster 1 (D2) = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0.4286 + 0.4286 + 0.6469 + 0 + 0 + 0 + 0 + 0 + 0.4286 + 0.4286 + 0.4286) / 18 = 0.18574
Cluster 1 (D3) = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0) / 18 = 0
Cluster 1 (D4) = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0) / 18 = 0
Cluster 1 (D5) = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0.5556 + 0.5556 + 0.5556 + 0.5556 + 0.5556 + 0.5556 + 0 + 0 + 0) / 18 = 0.0926

Then calculate the average in cluster 2 as follows:

Table 17. Cluster 2

Term	D1	D2	D3	D4	D5	Cluster Result
Pity	0	0	1	0	0	Cluster 2 (Negative)
Sir	0	0	1	0	0	Cluster 2 (Negative)
Prabowo	0	0	1	0	0	Cluster 2 (Negative)
Trick	0	0	1	0	0	Cluster 2 (Negative)
institution	0	0	1	0	0	Cluster 2 (Negative)
survey	0	0	1	0	0	Cluster 2 (Negative)

Cluster 2 (D1) = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0) / 6 = 0
Cluster 2 (D2) = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0) / 6 = 0
Cluster 2 (D3) = (1 + 1 + 1 + 1 + 1 + 1) / 6 = 1
Cluster 2 (D4) = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0) / 6 = 0
Cluster 2 (D5) = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0) / 6 = 0

So that the new centroid values are obtained, among others:

Table 18. New Centroid

	New Centroid				
Cluster (C0)	0.09524	0.17144	0.28572	0.86668	0.35558
Cluster (C1)	0.38897	0.18574	0	0	0.0926
Cluster (C2)	0	0	1	0	0

Calculate the distance to the cluster as done in the initial process:

$d(1,0) = \sqrt{(1-0.09524)^2 + (0-0.17144)^2 + (0-0.28572)^2 + (0-0.86668)^2} = \sqrt{+(0-0.35558)^2} = 1.344317525$

$d(1,1) = \sqrt{(1-0.38897)^2 + (0-0.18574)^2 + (0-0)^2 + (0-0)^2} = \sqrt{+(0-0.0926)^2} = 0.645315247$

$d(1,2) = \sqrt{(1-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 + (0-0)^2} = 1.414213562$

$d(5,2) = \sqrt{(0.4955-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 + (0-0)^2} = 1.116028785$

After doing the calculation, the overall result of the distance to the cluster is obtained as follows:

Table 19. Calculation of Distance to cluster

Term	D1	D2	D3	D4	D5	Distance to Cluster			Results Cluster
						C0	**C1**	**C2**	
mas	1	0	0	0	0	1.344318	0.645315	1.414214	Cluster 1 (Neutral)
wowo	0.4955	0	0	0	0	1.071824	0.233287	1.116029	Cluster 1 (Neutral)
gibran	0.4955	0	0	0	0	1.071824	0.233287	1.116029	Cluster 1 (Neutral)
...	...	...	...	...	...	...	...	...	...
funny	0	0.4286	0	0	0	1.017059	0.467818	1.087979	Cluster 1 (Neutral)

In the K-Means calculation, if the centroid has not matched the starting point, the calculation process will continue and if the centroid point is the same as the starting point, the calculation process stops. The results of the first and second stages have no centroid changes, so the calculation is complete with the results obtained. Then the results are in accordance with the cluster grouping. Here are the results of the clustering:

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

Table 20. Final Cluster Results

Term	D1	D2	D3	D4	D5	Cluster Result
mas	1	0	0	0	0	Cluster 1 (Neutral)
wowo	0.4955	0	0	0	0	Cluster 1 (Neutral)
gibran	0.4955	0	0	0	0	Cluster 1 (Neutral)
heart	0	0.4286	0	0	0	Cluster 1 (Neutral)
funny	0	0.4286	0	0	0	Cluster 1 (Neutral)

Model Implementation

After designing and building the system, the next stage is deployment to evaluate the extent to which the system meets the original objectives as well as test its performance and reliability under real conditions. In this study, the K-Means method was implemented using RapidMiner software, which facilitates data analysis and clustering based on identified patterns. The implementation process starts with importing a clean dataset, then adding the K-Means operator, determining the number of clusters (K) based on analysis or trials, and setting parameters such as distance and number of iterations. Once all the settings are complete, the clustering process is run, resulting in the grouping of data based on similar characteristics in the dataset. RapidMiner helps to simplify the application and improve the accuracy of the results through visualization and automatic application of the K-Means algorithm.

RapidMiner Operator View

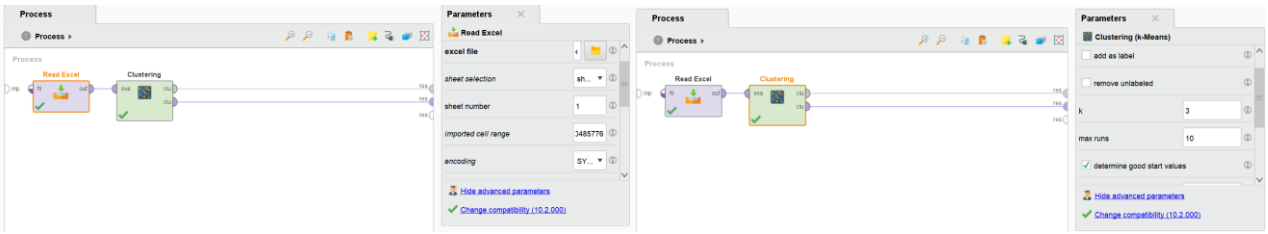

Figure 3. RapidMiner Operator View

In Figure 3 the RapidMiner Operator Display shown in the figure serves to create a center point or centroid of the data in the dataset imported from the xlsx format file. This operator is used to group the data into several clusters by identifying patterns based on the proximity between data. For further configuration, the clusters are set to three, namely Cluster 0, Cluster 1, and Cluster 2, each of which will be filled based on the characteristics of the data. In addition, the max runs parameter is set to 10, which means the algorithm will be repeated up to 10 times to search for optimal results, minimizing the variability of the clustering results. This operator automatically calculates and places the centroid at the center of the data in each cluster, allowing for a more in-depth analysis of the data distribution and patterns formed. This process helps in a more structured segmentation of the data, providing clearer insights for the interpretation of the results.

Display of K-Means Clustering Results

Cluster Model

```
Cluster 0: 34 items
Cluster 1: 1222 items
Cluster 2: 4 items
Total number of items: 1260
```

Row No.	id	cluster	Term	D1	D2	D3	D4	D5	D6	D7
1	1	cluster_1	10	0	0	0	0	0	0	0
2	2	cluster_0	2024	0	0	0	0	0	0	0
3	3	cluster_1	abah	0	0	0	0	0	0	0
4	4	cluster_1	abang	0	0	0	0	0	0	0
5	5	cluster_1	abu	0	0	0	0	0	0	0
6	6	cluster_1	acara	0	0	0	0	0	0	0
7	7	cluster_1	aceh	0	0	0	0	0	0	0
8	8	cluster_1	ada	0	0	0	0	0	0	0
9	9	cluster_1	adalah	0	0	0	0	0	0	0
10	10	cluster_1	adanya	0	0	0	0	0	0	0
11	11	cluster_1	adil	0	0	0	0	0	0	0
12	12	cluster_1	adili	0	0	0	0	0	0	0

Figure 4. Display of K-Means Clustering Results

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

In Figure 4 the Result View of the clustering model in RapidMiner shows that the sentiment analysis data has been successfully grouped into three clusters. Cluster 0 contains 34 words, Cluster 1 has 1,222 words, and Cluster 2 consists of 4 words, for a total of 1,260 words. This clustering is based on attributes such as Term (word), as well as frequency or weight values from D1 to D220, which reflect the distribution of words in different documents or sentiment categories. These clusters identify unique word usage patterns in the dataset, helping to categorize words based on their occurrence and contribution to positive, negative or neutral sentiment. With these results, we can understand more about the distribution of sentiment generated by certain words and how they affect the overall sentiment analysis. These groupings provide a clearer picture of the data structure and dominant patterns in the sentiment analysis performed.

Centroid Table View in RapidMiner

Attribute	cluster_0	cluster_1	cluster_2
D1	0	0.002	0.049
D2	0.012	0.001	0.085
D3	0.007	0.002	0.065
D4	0	0.001	0.151
D5	0.006	0.001	0.057
D6	0	0.003	0.041
D7	0	0.002	0.052
D8	0.019	0.002	0.070
D9	0.020	0.002	0.124
D10	0	0.003	0.047
D11	0	0.002	0.119
D12	0.007	0.004	0.021
D13	0	0.002	0.052

Figure 5. Centroid Table View in RapidMiner

In Figure 5 the Centroid Table View in RapidMiner displays the centroid values for each cluster generated by a clustering algorithm, such as K-Means. This table contains the average value of the attribute or feature that represents the center of each cluster, indicating the general characteristics of the data grouped in that cluster. Through this view, users can easily identify the differences between clusters based on the key attributes analyzed. Its purpose is to provide clearer insights into the relative position and pattern of data within each cluster, making it easier to interpret the clustering results, such as identifying dominant categories or distinct patterns of behavior between clusters. This view is also important in aiding more targeted decision-making, as users can understand the central characteristics of each data group.

Folder View in RapidMiner

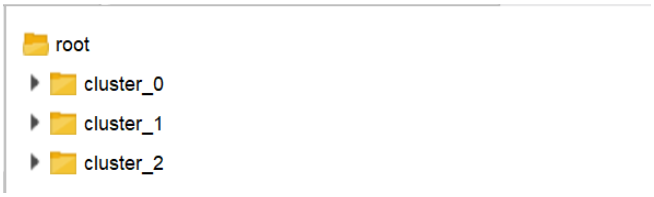

Figure 6. Folder View in RapidMiner

In Figure 6 the Folder View in RapidMiner or similar software serves to display the structure of files and folders used in the project, similar to a file explorer in an operating system. This view makes it easy for users to access, manage, and organize datasets, models, processes, and output results stored in the project directory. Its purpose is to provide easier navigation in finding and opening related files, thus speeding up the user's workflow in running and editing analytic processes. With Folder View, users can efficiently manage various data sources, experimental results, and other components, ensuring that all important files are organized and easily accessible when needed.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

RapidMiner Visualizations View

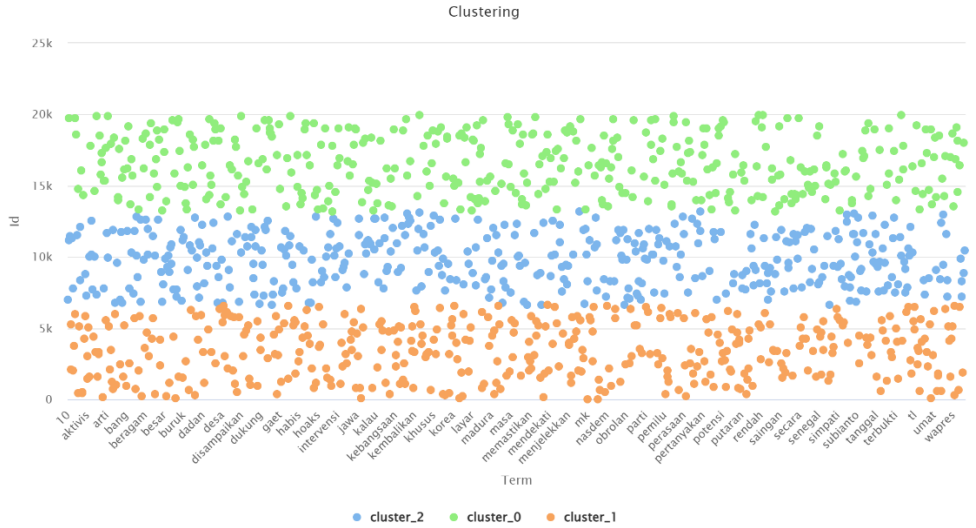

Figure 7. RapidMiner Visualizations View

In Figure 7 the Visualization View in cluster analysis in RapidMiner serves to present the results of data grouping in an easy-to-understand graphical form, such as a scatter plot or dendrogram diagram. These visualizations allow users to see the distribution and relationships between clusters in an intuitive way, making it easier to identify patterns, outliers, and unique characteristics of each group. Its purpose is to provide a clear visual picture of the data structure, which helps in decision-making and a better understanding of the dynamics in the dataset. With visualization, users can quickly evaluate cluster analysis results, communicate findings to others, and conduct further exploration of relevant data.

Testing

Blackbox testing basically focuses on evaluating whether the program meets the predefined functional requirements. In this case, testing is based on system specifications, such as functions, menu options, and compatibility of the models used in the research. This approach is done by running the program that has been developed, then observing whether the results match the expected requirements. As a clearer illustration, there is a system performance table below that provides a concrete picture of the extent to which the program meets the set standards and objectives.

Table 21. Blackbox Testing

No.	Scenario Work	Expected results	Testing Results	Conclusion
1	Input Data and processed using RapMiner.	Users can input data and be processed by RapMiner.	The user has successfully input data and processed by RapMiner according to what was input.	Validation
2	Data Pre-processing.	RapidMiner can process the input data.	RapidMiner successfully processes the input data.	Validation
3	Application of the K-Means method to data processing in RapidMiner.	Can apply the K-Means method to data processing in RapidMiner.	Successfully applied the K-Means method to data processing in RapidMiner.	Validation
4.	Data can be clustered using K-Means.	The system can process data to generate clusters using K-Means.	The system successfully processes the data to generate clusters using K-Means.	Validation
5.	RapidMiner visualizes the cluster results with the K-Means method.	RapidMiner can visualize cluster results with the K-Means method.	RapidMiner can visualize cluster results with the K-Means method.	Validation

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.4821

CONCLUSION

Based on the entire series of research that has been carried out, the author can formulate a number of conclusions that Based on the research conducted, the K-Means algorithm was successfully applied for sentiment analysis of public opinion on the 2024 Presidential Election, using tweet data taken through Twitter crawling as many as 220 tweets. From the dataset, 5 tweets are used for manual implementation of the K-Means algorithm calculation, through a series of pre-processing processes, including TF-IDF weighting. After the manual K-Means calculation, from 29 words generated from TF-IDF, the following clustering results were obtained: Cluster 0 (positive) contains 5 words, Cluster 1 (neutral) contains 18 words, and Cluster 2 (negative) contains 6 words. These results show that the K-Means algorithm can effectively cluster sentiment in public opinion data related to the 2024 presidential election based on patterns found in the words in the tweets. Based on the model implementation, the K-Means algorithm was successfully applied for sentiment analysis of public opinion on the 2024 Presidential Election using tweet data taken through Twitter crawling as many as 220 tweets. The process involves a series of pre-processing stages, including TF-IDF weighting, which is then implemented on RapidMiner for clustering. As a result, the data has been successfully grouped into three clusters: Cluster 0 contains 34 words, Cluster 1 includes 1,222 words, and Cluster 2 consists of 4 words, totaling 1,260 words. This clustering is based on attributes such as Term (word) and frequency weights from D1 to D220, which represent the distribution of words in different sentiment categories. These results show that K-Means is effective in clustering words based on sentiment patterns in public opinion data related to the 2024 presidential election.

REFERENCES

Abbas, M., Rioboo, R., Ben-Yelles, C. B., & Snook, C. F. (2021). Formal modeling and verification of UML Activity Diagrams (UAD) with FoCaLiZe. *Journal of Systems Architecture*, *114*. https://doi.org/10.1016/j.sysarc.2020.101911

Al-Fedaghi, S. (2021). Validation: Conceptual versus Activity Diagram Approaches. *International Journal of Advanced Computer Science and Applications*, *12*(6), 287–297. https://doi.org/10.14569/IJACSA.2021.0120632

Faesal, A., Muslim, A., Ruger, A. H., & Kusrini, K. (2020). Sentimen Analisis Terhadap Komentar Konsumen Terhadap Produk Penjualan Toko Online Menggunakan Metode K-Means. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *19*(2), 207–213. https://doi.org/10.30812/matrik.v19i2.640

Joergensen E Munthe, C., Astuti Hasibuan, N., & Hutabarat, H. (2022). Penerapan Algoritma Text Mining Dan TF-RF Dalam Menentukan Promo Produk Pada Marketplace. *Resolusi : Rekayasa Teknik Informatika Dan Informasi*, *2*(3), 110–115. https://doi.org/10.30865/resolusi.v2i3.309

Khan, M., Khan, S. S., & Alharbi, Y. (2020). Text Mining Challenges and Applications, A Comprehensive Review. *International Journal of Computer Science and Network Security*, *20*(12), 138–148. http://paper.ijcsns.org/07_book/202012/20201215.pdf

Kurniawan, F. W., & Dr. Warih Maharani, S.T., M. T. (2022). *Analisis Sentimen Twitter Bahasa Indonesia Menggunakan*. *6*(2), 7821–7829.

Makarychev, K., & Shan, L. (2022). Explainable k-means: don't be greedy, plant bigger trees! *Proceedings of the Annual ACM Symposium on Theory of Computing*, 1629–1642. https://doi.org/10.1145/3519935.3520056

Melati, R., & Reza, M. (2024). Analisis Sentimen Data Twitter Menggunakan Metode *K-Means Clustering* Pada Studi Kasus Pemindahan Ibu Kota Nusantara (Ikn). *Technology Acceptance Model) Jurnal TAM*, *15*(1), 66–73.

Mohammadi Sarab, M., Shahrokhi, M., & Tabatabaei, O. (2020). A Structured Approach for Display of the Most Practical Theories in ELT. *Journal of Critical Studies in Language and Literature*, *1*(3), 36–55. https://doi.org/10.46809/jcsll.v1i3.27

M. Taufik Aufa, Jasmir, & Rohaini, E. (2024). Perancangan Sistem Informasi Pelayanan Pengaduan Masyarakat Kelurahan Bagan Pete Kota Jambi Berbasis Website. Jurnal Informatika Dan Rekayasa Komputer(JAKAKOM), 4(1), 937–945. https://doi.org/10.33998/jakakom.2024.4.1.1673

Nascimento, M. (2022). In Search of Star Clusters: An Introduction to the K-Means Algorithm. *Journal of Humanistic Mathematics*, *12*(1), 243–255. https://doi.org/10.5642/jhummath.202201.19

Putri Yolanda, H., & Halim, U. (2020). Partisipasi Politik Online Generasi Z Pada Pemilihan Presiden Indonesia 2019. *CoverAge: Journal of Strategic Communication*, *10*(2), 30–39. https://doi.org/10.35814/coverage.v10i2.1381

Riyani, A., Zidny Naf'an #2, M., & Burhanuddin, A. (2019). Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen. *Jlk*, *2*(1), 23–27.

Rofiqi, M. A., Fauzan, A. C., Agustin, A. P., & Saputra, A. A. (2019). Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, *1*(2), 58–64. https://doi.org/10.28926/ilkomnika.v1i2.18

Rosaly, R., & Prasetyo, A. (2019). Pengertian Flowchart Beserta Fungsi dan Simbol-simbol Flowchart yang Paling Umum Digunakan. *Https://Www.Nesabamedia.Com*, *2*, 2. https://www.nesabamedia.com/pengertian-flowchart/https://www.nesabamedia.com/pengertian-flowchart/