Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 11(114)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10102458

Exploring the impact of social media on political discourse: a case study of the Makassar mayoral election

Jufri[1*], Aedah Binti Abd. Rahman[2] and H. Suarga[1]
Department of Software Engineering, Dipa Makassar University, Makassar, South Sulawesi, Indonesia[1]
Department of School of Science and Technology, Asia e University, Kuala Lumpur, Malaysia[2]

Received: 24-October-2023; Revised: 16-May-2024; Accepted: 19-May-2024

Abstract

Social media has become a significant force in today's modern society, influencing several areas, including politics, education, the economy, and the spread of information. This study examines how social media platforms influence political discourse, focusing on the Makassar mayoral race. This study looks into how Twitter, a well-known social media site, encourages more user participation in communications related to the Makassar mayoral election. In light of the Makassar mayoral election, this study employs several approaches, including data collection, acquisition, consolidation, and analysis, to address topics that are becoming increasingly popular on social media. Election dynamics are examined using the naïve Bayes approach. To increase the accuracy and efficiency of text mining operations, especially in result validation, text clustering, and classification, the k-means algorithm and support vector machines (SVM) were used. A hybrid method is employed to combine the benefits of k-means, SVM, and naïve Bayes. This method seeks to thoroughly grasp how social media affects conversations about the mayoral race, offering insightful information to political scientists and practitioners. The research on the Makassar mayoral race explores the influence of social media on political communication, highlighting Twitter's influence and a hybrid algorithm for sentiment analysis. It indicates the importance of social media strategy in political campaigns, providing insights for decision-makers, parties, and the public and recommending future research in this dynamic field.

Keywords

Social media, Twitter, Political, Algorithm, Election.

1.Introduction

Social media provides various communication channels, including political communication. Communication via social media is more attractive to people than mainstream media [1]. People can more easily approach politicians and public officials to convey their thoughts, criticism, and aspirations regarding political issues and agendas via social media [2]. People use social media for political communication, and they derive benefits from it. This research critically examines social media's role in political communication. Therefore, this research offers findings regarding the use of social media in political communication [3, 4]. The field of education emphasizes the importance of engaging in deep political conversations to cultivate democracy and its principles, further exploring the media's influence on political behaviour and beliefs [5].

Social media also facilitates mobilization efforts by political candidates or parties, such as campaign stickers and attire. However, it is not an adequate platform for campaign purposes due to limited reading and listening skills in Indonesia [6]. Political groups use social media to collect opinions and assign blame to unfavourable individuals while also promoting their stance [6]. Social media can serve as a platform for political participation during election season, enabling candidates to engage with the public and convey their vision and mission without physical presence [7]. This particular campaign is more cost-effective and time-efficient compared to standard campaigns [8]. Researchers are employing direct observation, demographic surveys, and social life components to analyze political communication data for mayoral candidates. [9].

Understanding popular moods is critical for candidates, political parties, and other stakeholders in the ever-changing political landscape of elections to assess public opinion and adjust their communication

[*] Author for correspondence

tactics [10]. The emergence of digital platforms has significantly transformed political communication, resulting in an enormous volume of textual data produced by social media posts, news stories, and online conversations. This abundance of information presents a previously unheard-of chance to assess public opinion on political figures and events. However, because of the sheer amount and intricacy of this data, advanced analytical methods that are precise and fast in processing and interpreting sentiment are required [11].

This study presents a new approach to text mining that integrates three potent analytical techniques [12]: k-means clustering, naïve Bayes, and support vector machines (SVM). Each method brings something different to the study. For example, SVM is good at stable classification, naïve Bayes is good at probabilistic modelling, and k-means is good at finding hidden patterns in the data [12]. Our method combines these approaches to offer a thorough sentiment analysis of political communication during the Makassar mayor election, a significant event that has drawn interest from both academics and practitioners.

Due to its importance in the political environment and the abundance of rich textual material it has produced, the Makassar mayor election is a perfect case study for our research. This research not only offers a technical explanation of the hybrid technique but also demonstrates its practical application in analyzing sentiment in political communication. This analysis aims to clarify public perceptions of the candidates, the main topics influencing election discourse, and the general dynamics of political communication during the election season. This research combines SVM, naïve Bayes, and k-means clustering into a hybrid approach for sentiment analysis in electoral processes. This approach improves theme analysis and sentiment categorization, offering a more complex understanding of popular sentiment in political communication. The study has implications across disciplines, including political science, computational linguistics, and data science, as it enhances the study of electoral communication, improves text mining methods, and enables the practical application of hybrid analytical approaches in problem-solving.

This study investigates the political communication strategies employed by the Makassar community in relation to the forthcoming elections, with a specific focus on social media platforms. The aim is to comprehend the influence of the mayoral election. The study collects empirical data on Twitter's political communication dynamics during the election, examining small nuances and enhancing stakeholders' understanding of the political environment [13]. This study introduces a research framework, as shown in *Figure 1*, which centers on the application of social media in the realm of politics.

The research framework explains its aim to investigate specific political issues on Twitter, including the election process, policy debates, and attitudes towards specific issues [14]. The emphasis is on advancements in methods for monitoring political communication on social media during the voting process. [15]. While attitude analysis determines public opinion on a subject, text preprocessing organizes and standardizes text data. Topic modeling determines the points discussed. Network analysis focuses on understanding information spread across a network, while trend analysis tracks changes in sentiment and topic popularity over time [16]. Visualizations and summaries are essential to concisely summarize findings so they can reach stakeholders such as politicians and policymakers. Ethical and privacy considerations are especially important when handling sensitive political information and social media data [17]. Evaluation and iteration are important procedures that require continuous evaluation and updating to respond to assessments and changing social media dynamics. Summarization and visualization approaches are useful for condensing and presenting new research knowledge to stakeholders [18].

The research framework aims to investigate specific political issues on Twitter, including the election process, policy debates, and attitudes towards particular issues. It emphasizes advancements in methods for monitoring political communication on social media during the voting process [14]. Attitude analysis determines public opinion on a subject, while text preprocessing organizes and standardizes text data. Topic modeling identifies the main points of discussion [15, 16]. Network analysis focuses on understanding the spread of information across a network, and trend analysis tracks changes in sentiment and topic popularity over time. Visualizations and summaries are essential for concisely presenting findings to stakeholders such as politicians and policymakers. Ethical and privacy considerations are crucial when handling sensitive political information and social media data [17]. Continuous evaluation and iteration are necessary to adapt to assessments and the evolving dynamics of social media. Summarization and visualization techniques are valuable for condensing and presenting new research insights to stakeholders [18].

With a focus on the Makassar mayoral election, the paper offers a thorough analysis of the use of social media for political communication. Reviewing earlier research on digital communication channels highlights how social media has revolutionized political debate [19]. Examining social media's influence on political behavior and communication tactics critically is the driving force behind the study [20]. One of the goals is to improve stakeholders' comprehension of political dynamics and use text-mining techniques to analyze communication patterns among political leaders [21]. This paper makes three contributions: it introduces a new research paradigm for using social media analytics in political contexts, it uses mixed methods to analyze data, and it suggests ways to improve text mining techniques for political communication [22].

Literature review is explored in Section 2. Methods are discussed in Section 3. Section 4 covers the results and experimentation. Discussion of the results is presented in Section 5. The conclusions are provided in Section 6.

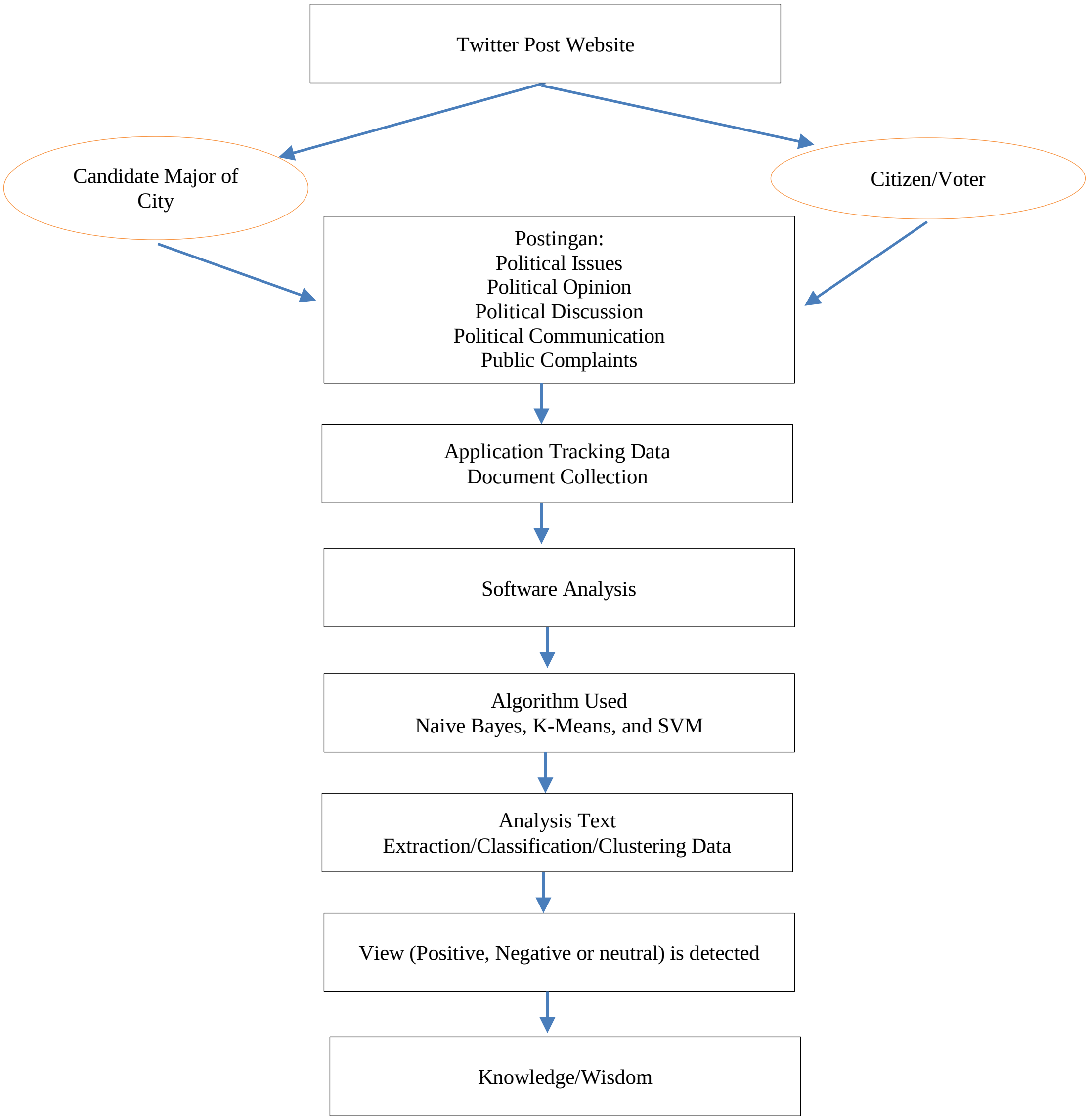


Figure 1 Research framework

2.Literature review

The rapid advancement of internet-based information technology has significantly impacted communication, creating digital channels like social media platforms, news sites, and blogs. These platforms enable the distribution and consumption of information across various fields, including entertainment, education, politics, and economics [18]. The importance of social media in human experience extends to knowledge collection and interpersonal connections. The shift in communication methods is characterized by a shift from traditional face-to-face encounters to online platforms, with social media being a widely used mode of internet-based communication [23]. Traditional media, such as electronic and print media, have been transformed into internet-based platforms that serve various areas, including education, culture, social affairs, economics, law, and politics. For example, election campaigns use social media channels to distribute regional candidates' vision, mission, and operational plans, increasing public knowledge and engagement [24].

Political communication is a critical field involving various communication strategies used by political actors to engage, educate, and influence others [25]. As technology advances, it becomes increasingly important to adapt to the competitive dynamics of a particular era. Political communication is one of the ten distinct forms of communication, including interpersonal, intergroup, rumor dissemination, public, and media communication [26]. Public opinion, collective perspectives, and the emotions of the broader populace all significantly impact political communication, with notable instances where prevailing sentiment can overthrow long-standing governing systems [27].

Social media networks are digital platforms that promote human engagement and interconnectedness [28]. They enhance interpersonal relationships and facilitate the dissemination of information and data [29]. These platforms integrate personal communication for individualized sharing and public media for a wider audience [30]. They enable the collection, dissemination, and interaction of information and data, fostering collaboration and leisure activities [31]. Social media platforms, such as Twitter, Instagram, blogs, Facebook, WhatsApp, and Line, offer user-generated content and facilitate the exchange of information and data [32].

Political communication involves the application of communication principles in various political campaigns, aiming to influence public perception and promote different ideas. Using a strategic approach, it encompasses creating, implementing, and managing campaigns by candidates, parties, governments, lobbyists, and interest groups. [33]. The importance of political communication lies in the visual representation of political individuals, and the use of modern technology in campaigns allows for targeted messages and the creation of political imagery. Political branding is a direct outcome of this process [34].

Data mining is a systematic procedure that employs artificial intelligence, statistical techniques, and machine learning algorithms to derive valuable insights from extensive databases. [35]. This includes building models, using filters, and analyzing extensive information to identify patterns. This holistic methodology combines various scientific fields, enabling the creation of valuable information and increasing the value of data. Artificial intelligence, machine learning, statistics, databases, and information retrieval have a strong history in data mining. The widespread adoption of broader data mining techniques has been driven by the availability of vast amounts of data and its diverse and extensive nature [36].

Text mining is a valuable tool for data analysis, but it has inherent challenges. Extracting knowledge from textual databases, focusing on identifying and interpreting new knowledge [37]. This process is crucial for data extraction and search, but finding text data with intricate patterns is challenging. Researchers have extensively studied text mining techniques, contributing to the field's refinement and its associated patterns [38]. It can be categorized into text grouping, text clustering, association rule extraction, and trend analysis [39]. Supervised learning trains data using labels to generate new data. It uses accurate and inaccurate data to produce reliable results. Functions are generated using test data labels, and output data is monitored for adjustments. The training results can be used for accurate classifications or predictions [40]. SVM is a binary classification algorithm used for classification and regression tasks in various domains. It uses a hyperplane to separate data into distinct classes, with support vectors at specific distances. SVM is beneficial in information retrieval for classification tasks, especially in large-dimensional text data [41].

Twitter, a popular social media platform, is experiencing rapid growth due to its large user base and the daily generation of over 500 million tweets. With over 313 million active users and a significant number of tweets, Twitter applications have the potential to generate revenue. However, the large number of tweets presents challenges in efficiently distributing and organizing relevant information for users, groups, and government agencies [42]. Researchers are exploring sentiment analysis or opinion mining in natural language use, particularly its irregular form. Researchers from various fields, including prediction, classification, clustering, communication, events, commerce, movies, and finance, frequently use Twitter [43]. Twitter has become a crucial social media platform for political communication, transforming how politicians connect with the public, conduct campaigns, and encourage popular participation in political discourse by utilizing various components within the platform [44].

The term frequency-inverse document frequency (TF-IDF) algorithm is a computational tool used to determine the importance of frequently occurring words in a document. It computes TF-IDF values for every word in each document. The TF-IDF method assigns weight values to words based on their associations. The word's frequency within the document positively influences the strength of the association, while the total frequency of papers containing that word throughout the document collection inversely influences it [45].

3.Methods

Using a mixed-methods approach, this study analyzed social media information about the Makassar mayor election using both qualitative and quantitative text-mining approaches. The following is the structure of the methodology:

1. Data collection and content pertaining to the Makassar mayoral election was the primary source of data, with additional information obtained from Twitter and other social media sites. The collection covered 2019 to 2020, going beyond election day to record conversations and responses after the poll.
2. Keyword-based crawling and keywords pertinent to the Makassar mayoral election were employed in the data collection. This required automated crawling methods that extracted relevant posts, comments, and hashtags using Python scripts and programs like RapidMiner. Particular search phrases were used to filter and find information that directly mentioned or discussed the election.
3. Use of social media application programming interface (API) another aspect of the collecting process was the utilization of social media API or application programming interfaces. These allow for the methodical acquisition of social media data based on predetermined criteria, such as hashtags, date periods, and keywords.
4. Preprocessing steps: the gathered data was cleaned (to remove spam, duplication, and irrelevant information), normalized (to standardize text for consistency), and segmented (to divide the text into manageable bits for analysis) before analysis.
5. Sentiment and relevance filtering: as part of the preprocessing, the posts, comments, and hashtags were categorized according to their relevance to the election and their sentiment (positive, negative, or neutral). In order to guarantee that the ensuing studies were concentrated on pertinent and contextually significant data, this classification was necessary.
6. The use of an organized methodology for a targeted and effective examination of the social media conversation around the mayoral contest guaranteed that the information was pertinent and of adequate calibre to permit in-depth sentiment analysis and trend detection.

This research offers a novel text mining approach for social media political conversation analysis that integrates three potent analytical methods (k-means clustering, naïve Bayes and SVM). Every technique offers distinct benefits to the analysis.

1. K-means clustering is well known for its capacity to find latent patterns in data, allowing text to be sorted into distinct groups according to similarities. Finding recurrent themes or concepts in political discourse may benefit from this.
2. Text has been categorized using naïve Bayes algorithms based on the likelihood that it belongs in a particular category (like positive, negative, or neutral emotion). When dealing with big data sets, this method is quite useful since it can swiftly scan text to determine emotion or themes.
3. SVM, a common option for differentiating between various phrases or thematic groups in text, offers strong categorization capabilities. SVM is effective in situations when class distinction is challenging and can handle large amounts of data.

This research aims to do a comprehensive sentiment analysis of political communication during the Makassar Mayoral Election by combining the advantages of various techniques. This hybrid approach intends to improve the analysis of attitudes and subjects in political communication by providing a deeper knowledge of public views, major themes influencing electoral discourse, and the overall dynamics of political communication during the election season.

3.1Research conceptual

Conceptual research in text mining involves systematically analyzing unstructured data to extract valuable information using natural language processing (NLP) techniques and algorithms [46]. The goal is to convert unprocessed data into an organized format, enabling researchers to identify trends and make data-driven decisions. The approach aims to understand the interconnections among keywords, allowing a deeper understanding of textual content. Challenges in conceptual text mining, such as ambiguity and contextual reliance, are significant, but advancements in deep learning methodologies have improved our ability to tackle these issues. *Figure 2* shows the conceptual framework for this study.

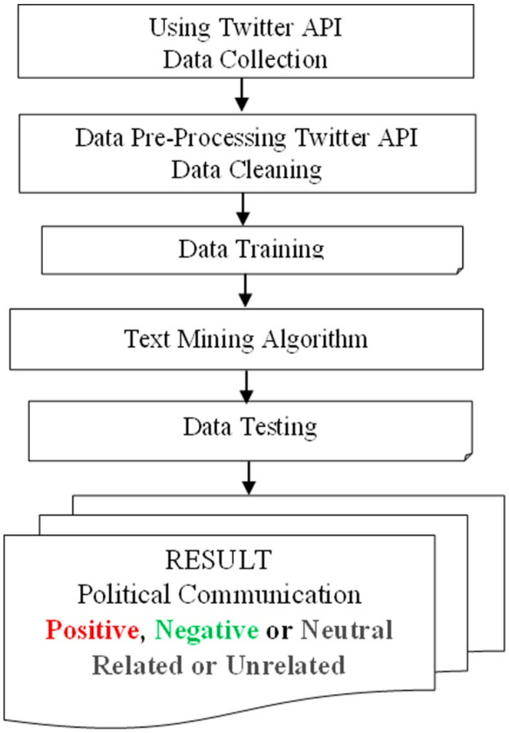


Figure 2 Conceptual work flow of the study

The data collection process for Twitter entails selecting social media channels and dividing the data into training and test datasets. Text mining algorithms, such as filters and wrappers, are used to convert unstructured text into structured numerical representations. Sentiment analysis classifies politically affiliated or unrelated messages in social media posts based on their relevance to politics and prevalent sentiment. This results in collecting "related" messages, sorted into positive, negative, or neutral groups.

3.2New improved methodology for social media text mining in political communication

The objective of this project is to enhance the process of analyzing political communication on social media by integrating data collecting, pre-processing, sentiment analysis, topic modeling, and visualization. The methodology that includes gathering data, preparing it for analysis, conducting sentiment analysis, performing topic modelling, and creating visual representations is necessary, following a series of steps, as seen in *Figure 3:*

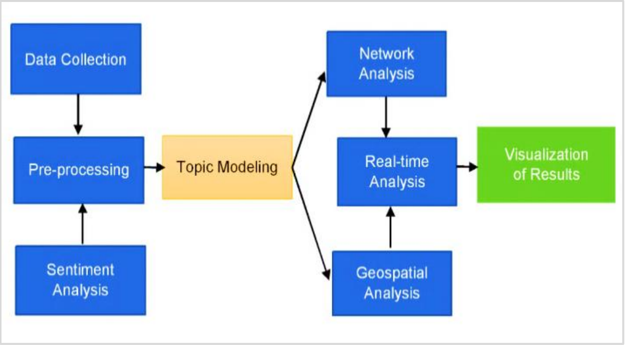


Figure 3 Improved approach to support social media text mining

This text-mining approach aims to analyze political communication data on social media platforms and obtain desired insights. It involves data collection and processing using an API, pre-processing techniques to eliminate noise, and text standardization through tokenization, stemming, and lemmatization [46]. Sentiment analysis is used to determine the prevailing sentiment in social media posts on specific subjects or issues. Topic modeling helps identify significant subjects and themes in political communication, making it easier to categorize postings and improve comprehension of commonly discussed issues [47]. Real-time analytics is recommended for rapid analysis and response to emerging patterns. Geospatial analysis determines spatial patterns and the dispersion of political communication, potentially uncovering regional differences in mood and political inclinations. Visualizations, such as dashboards, charts, and graphs, help policymakers, political analysts, and the general public better understand the insights obtained by text-mining algorithms.

3.3Text mining using hybrid algorithm naïve Bayes, SVM, and k-means technique

A novel hybrid method, which integrates the naïve Bayes and k-means techniques, is suggested for text mining applications. This approach aims to improve the efficiency and accuracy of text classification and grouping tasks. This approach leverages the advantages of both algorithms, with a comprehensive operational framework outlined [48]. The integration of naïve Bayes and k-means is a technique that categorizes texts into groups based on text data. The naïve Bayes algorithm is used for initialization, categorizing texts into groups. The k-means algorithm is then used for clustering, using Naïve Bayes' class labels as initial centroids. This approach improves clustering results. The k-means algorithm splits texts into K clusters, using representative documents as training data. Naïve Bayes classifiers are constructed for every cluster, and when a new document requires categorization, it is allocated to the closest cluster. The hybrid algorithm's performance is assessed using metrics like accuracy, precision, recall, or F1-score. Iterative model refinement and parameter changes are crucial for achieving enhanced results [49].

The SVM linear algorithm was originally employed in mathematical formulations for the purpose of binary classification. The training data and its corresponding labels were provided (xn,yn), n = 1……N, xn € RD, tn € {-1, +1}, SVM limited optimization of the learning process [50]. Formula as follows Equation 1.

$$\min_{w,\varepsilon n} \tfrac{1}{2} w^T w + C \textstyle\sum_{n=1}^{N} \varepsilon n \qquad (1)$$

$$s.t.\, w^T x_n t_n \geq 1 - \ \varepsilon n \,\forall n \ \varepsilon n \ \leq 0 \,\forall n$$

- w represents the weight vector of the hyperplane used for classification.
- $w^T w$ is a measure of the margin width; SVM aims to maximize this margin while classifying data points.
- ½ $w^T w$ is used for mathematical convenience, as its derivative is more straightforward to compute.
- ϵn are slack variables introduced to allow misclassification of difficult or noisy data points.
- C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification error. A larger value of C puts more emphasis on minimizing classification error.
- $\sum{=}1\ \sum n{=}1N\epsilon n$ is the sum of the slack variables for all data points, representing the total penalty for misclassifications.

The naive Bayes classification algorithm is a classic Bayesian classification technique known for its straightforward algorithmic structure and efficient computational capabilities. Its advantage lies in its ability to estimate necessary parameters using limited training data, making the overall covariance matrix unnecessary [51]. The naive Bayes classifier is effective in various practical scenarios, such as document classification and word prediction. However, its reliance on feature independence can lead to inaccurate classification outcomes. The naive Bayes classifier is particularly suitable for large datasets with numerous features, where intricate models may overfit or require excessive computational time [52]. The classification of opinion documents can be divided into two stages: training on pre-categorized documents and classification on pre-existing documents. The algorithm's simplicity and accuracy make it an ideal choice for data analysis [53].

The algorithm classifies documents by identifying the category with the highest probability among all examined, represented mathematically using Equation 2.

$$V_{MAP} = \arg \begin{matrix} max \\ vjev \end{matrix} \Bigl(\frac{P\left(X_1, X_2, X_3, ..X_n \middle| V_j\right) P\left(V_j\right)}{P(X_1, X_2, X_3, ...X_n} \Bigr) \qquad (2)$$

For P(x1,x2,x3,.... xn) the value is constant for all categories (Vj) this equation can be expressed mathematically using Equation 3:

$$V_{MAP} = \frac{argmax}{VjeV} \ (P\left(x_1, x_2, x_3, \ldots.x_n \middle| V_j\right) P\left(V_j\right)) \ (3)$$

- V_{MAP} stands for the category that maximizes the posterior probability.
- MAP (Maximum A Posteriori) estimation involves choosing the hypothesis that is most probable given the observed data.
- vj refers to a possible category in the set of all categories V.
- $X1$,$X2$,$X3$,…,Xn represent the features of the document, which could be things like word frequencies, presence of certain keywords, etc.
- $P(X1,X2,X3,…,Xn∣Vj)$ is the likelihood, the probability of observing the document's features given that it belongs to category vj.
- $P(Vj)$ is the prior probability of category vj. This reflects how common or rare this category is in the general context or in the training data.

- $P(X1,X2,X3,\ldots,Xn)$ is the evidence, the overall probability of observing the document's features. This term is the same for all categories, so it's often ignored in the actual computation of the argmax since it doesn't affect which category is most probable.

Clustering refers to the systematic procedure of classifying data into distinct groups by considering their shared characteristics and differences [54]. This process is a component of unsupervised learning, wherein data is segregated into distinct clusters. The k-means algorithm is a widely used clustering technique in which a given dataset is divided into k distinct groups. The best assignment of data points to clusters is determined by minimizing the distance between each data point and its allocated cluster [55].

- To ascertain the suitable value for k, representing the desired number of clusters to be formed, a thorough investigation must be conducted.
- Generate k randomly selected values to serve as the initial cluster center, generally referred to as the centroid.
- The utilization of the Euclidean distance formula allows for the computation of distances between individual data points and the centroid. The technique described above is repeatedly performed until the shortest distance between each data point and centroid is determined. The subsequent formula is well recognized as the Euclidean distance formula, Equation 4.

$$d(x_i,\mu_j) = \sqrt{\Sigma(x_{i-}\mu_j)^{\ \ 2}} \qquad (4)$$

x_i = Criteria data to-i, i = 1, 2,…, n
μ_j = Criteria data centroid cluster to-j, j=1, 2, 3, …, m
d = Shortest distance between criteria data

- The data should be classified according to its closeness to the centroid. The cluster that has the greatest level of similarity to the centroid will be selected, as indicated by the data point with the closest value.
- Adjust the centroid value. The computation of the updated centroid value entails finding the mean of the cluster, which may be accomplished by utilizing the subsequent formula, Equation 5:

$$C_j = \frac{1}{n_k}\sum_{j=1}^{N_k} X_j \qquad (5)$$

C_j = New centroid
N_k = Centroid criteria data in clusters are combined into clusters
J = Total amount of data to -j

- Proceed with the iterative procedure starting with step 3 and continuing until step 5, repeating the process until the cluster remains unchanged and no adjustments are made to any centroid. It is particularly important to ensure that there are no alterations between the most recent centroid and the centroid before it. The last iteration will be employed as a parameter for establishing the data cluster.
- The term-frequency value is determined using numerous formulas, Equation 6, namely:

$$Tft,d = 1 + {}^{10}\mathrm{Log}\ tf \qquad (6)$$

Each variable is defined and described as follows:
tf : where each variable is explained as follows
Tft,d : term frequency or number of words t in document d or local weighting

- Calculate the inverse document-frequency value using the given formula as follows Equation 7:

$$Idf_t = 10\log n/dft \qquad (7)$$

Idft : Inverse document-frequency, also known as global weighting
n : lots of documents
dft : number of documents that contain the word t.

- The weight value (Wt,d) can be calculated by multiplying Equation 1 and Equation 2, resulting in the expression presented in Equation 8.

$$Wt,d = tft,dx\ idft \qquad (8)$$

Where:
Tft,d : Term frequency refers to the count of words in a document, also known as local weighting
Idft : inverse document frequency or global weighting
Wt,d : word final weight value

3.4How the naïve Bayes, SVM, and k-means algorithms are integrated, and the steps taken to optimize this hybrid approach

The process is carried out by integrating naïve Bayes, SVM, and k-means algorithms into a hybrid text mining approach, which is specifically intended to increase the effectiveness and accuracy of text classification and clustering tasks [56]. The unique advantages of each algorithm are combined in this hybrid algorithm to improve text data processing and analysis, especially when it comes to political communication on social media [57]. The following procedures and actions are taken to optimize:

1. Prepare data: To begin, data must be collected and pre-processed using various methods, including

tokenization, stemming, and lemmatization, to standardize text and remove noise.

2. Using naïve Bayes for initial classification: The naïve Bayes technique was initially used to classify processed text data into initial groups. Due to its effectiveness and probabilistic basis, naïve Bayes is a good choice for managing uncertainty in text data, where the text is first categorized according to its content.
3. K-means clustering: The texts are grouped into k clusters using the k-means technique after the first classification. The class labels generated by naïve Bayes are used as the initial centroid to build clusters. By using predictive insights from naïve Bayes to guide k-means spatial clustering, this integration aims to improve clustering results.
4. SVM-based refinement: At this point, the SVM algorithm is introduced to improve the classification of texts in clusters. SVM are well known for their ability to represent complex class boundaries and perform well in high-dimensional environments. The model can more accurately identify various subjects or attitudes in text input by using SVM.
5. Iterative optimization: Metrics including accuracy, precision, recall, and F1-score are used to continuously evaluate the performance of the hybrid algorithm. These assessments continue to improve the model. This may require changing algorithm parameters, redefining the number of k-means clusters, or updating the SVM decision boundaries.
6. Comprehensive assessment. After optimization, the hybrid model undergoes one final test to see how well it can reliably categorize and group text data. To verify the robustness of the results, one must assess model performance against a predefined set of benchmarks or apply cross-validation techniques.

This hybrid approach aims to distinguish SVM, naïve Bayes, and k-means' strengths in classification, NLP, and large data sets. It integrates and optimizes these algorithms to analyze complex political communication in social media, enabling more extensive analysis of public opinion and political polarization [58].

3.5Validate a new method that enhances text mining techniques for analyzing political communication

The validation method for a novel strategy targeted at improving text mining technologies in the field of political communication involves a series of processes specifically developed to assess its accuracy. The validation process encompasses the execution of numerous procedures [59]. The confusion matrix is a widely-used method for assessing the effectiveness of supervised machine learning algorithms. The confusion matrix is a valuable tool in the field of machine learning because it allows for the precise measurement of uncertain confusion when evaluating performance [60]. By integrating semantic concerns, there is the possibility of further improving this analysis. Therefore, the researcher concludes that by expanding the measurement space, there is a possibility of achieving better performance [61]. Within the framework of performance measurement using the confusion matrix, there are four specific terms that represent the results of the categorization process. These words refer to the true negative (TN) value, which represents the amount of correctly detected negative data, and the false positive (FP) value, which refers to negative data that is mistakenly identified as positive data. A true positive (TP) refers to the accurate identification of positive data, whereas a false negative (FN) is an error in classifying positive data as negative. [62]. The confusion matrix formula for determining accuracy is presented below using Equation 9.

$$\text{Accurate}= \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \qquad (9)$$

Where:
TP =The number of positive data grouped correctly..
TN =The number of negative data grouped correctly.
FN =The number of negative data but classified as wrong.
FP=The number of positive data but classified as wrong.

Through the examination of the metrics derived from the confusion matrix, researchers can acquire valuable insights regarding the efficacy of the text mining method. This facilitates informed decision-making and aids in implementing necessary modifications to enhance the precision and reliability of the model.

3.6Required hardware and software

For detailed hardware and software requirements relating to text mining or data analysis as described, general recommendations may include:

a. Hardware:

- Computer with Intel i7 multi-core processor to handle large data sets efficiently.

- Minimum random-access memory (RAM) 8 GB, recommended 16 GB or more for handling large data sets.
- Sufficient storage space, preferably solid state drive (SSD), for fast data access and storage of large data sets.

b. Software:

- Operating system Microsoft Windows 10
- Python environment that includes libraries for data analysis (pandas), machine learning, and NLP.
- Twitter is a medium for political discourse analysis and data collection that affects political discourse, information sharing, and election-related attitude.
- Database software for storing and managing my structured query language (MySQL) data.
- Jupyter Notebook development tool for writing and testing code.
- RapidMiner functions as a tool for managing, processing and analyzing data.

4. Result

This research provides a major foundation for scientific involvement in studying Indonesian voting procedures. The author presents a comprehensive analysis of the results obtained from research carried out in this special field. The results are categorized into different categories, each of which is aligned with a specific research objective. Visual aids, such as tables, charts, and graphs, are used to enhance the presentation of research findings and facilitate effective data communication. In addition, this study provides a comprehensive analysis of the research findings, accompanied by a careful explanation of the subsequent conclusions.

4.1Data collection

Process crawling data, collecting Twitter tweet data since 2019 until 2020 using Twitter data crawling. This process involves using RapidMiner tools and Python for data crawling. The data is retrieved using keywords related to the Makassar mayoral election. The Python programming language is used to extract tweet data, using the "scrape" module to extract relevant information. This automated process ensures accurate and relevant data for our analysis, using the following script:

```
twitter_search = "Pilwali makassar since:2021-01-01
until:2021-01-10
filename = f"{twitter_search.replace(' ', '_').replace(':', '-
').replace('#', '')}_{datetime.date.today().strftime('%Y-%m-
%d')}.json"
USING_TOP_SEARCH = True
snscrape_params = '--jsonl --max-results'
twitter_search_params = ''
if USING_TOP_SEARCH:
   twitter_search_params += "--top"
snscrape_search_query = f"snscrape {snscrape_params}
{max_results} twitter-search {twitter_search_params}
'{twitter_search}' > {filename}"
print(snscrape_search_query)
os.system(snscrape_search_query)
import pandas as pd
import ast
import json
tweets_df = pd.read_json(filename, lines=True)
NAMA_FILE_CSV = 'pilwalii.csv'
new_columns = {
   'conversationId': 'Conv. ID',   'url': 'URL',
       'date':  'Date',      'rawContent':  'Tweet',      'id':
'ID',   'replyCount': 'Replies',   'retweetCount': 'Retweets',
             'likeCount':    'Likes',           'quoteCount':
'Quotes',    'bookmarkCount':  'Bookmarks',     'lang':
'Language',
    'links': 'Links',   'media': 'Media',   'retweetedTweet':
'Retweeted Tweet',   'username': 'Username'}
if len(tweets_df) == 0:
    print('Pencarian tidak ditemukan coba ganti keyword
lain, keywordmu: ', twitter_search)
   exit()
else:      tweets_df   =   tweets_df.loc[:,   ['url',   'date',
'rawContent', 'id',
                       'replyCount', 'retweetCount', 'likeCount',
'quoteCount',
                     'conversationId', 'lang', 'links',
                                   'media',  'retweetedTweet',
'bookmarkCount', 'username']]
  tweets_df = tweets_df.rename(columns=new_columns)
 tweets_df['Media'] = tweets_df['Media'].apply(lambda x:
x[0]['fullUrl']  if  isinstance(x,  list)  and  x  and
isinstance(x[0], dict) and 'fullUrl' in x[0] else None)
  tweets_df['Links'] = tweets_df['Links'].apply(lambda x:
x[0]['url'] if isinstance(x, list) and x and isinstance(x[0],
dict) and 'url' in x[0] else None)
  display(tweets_df)
  tweets_df.to_csv(NAMA_FILE_CSV, index=False)
```

The script code above displays the format of data that can be retrieved, stored in the form of a JavaScript object notation (JSON) file, and then the tweet data is retrieved and saved in the form of a comma separated value (CSV) file.

4.2 Dataset

The study collected 728 tweet data from the Makassar mayor election from 2019-2020, augmented with searches from 2019-2020, and pre-processed it to 254 tweet data points, illustrating the structure of tweet data. Several lines of tweet data from the Makassar mayor election, as shown in *Table 1*.

Table 1 Sample dataset

No.	Tweet
1	Dari 3 calon pasangan walikota makassar yg maju ini, siapaami nanti yg punya program tuk urus public transport di makassar?
2	Momen ter-SAVAGE di debat #Pilkada2020 calon Walikota Makassar beberapa waktu lalu. https://t.co/Y56Al9InAs
3	Alhamdulillah sdh terpakai tenda tenda bantuan dr kandidat calon walikota Makassar APPI ARB https://t.co/4opqhdmzQ6
4	Gerindra resmi mendukung Mohammad Ramdhan Pomanto (@DP_dannypomanto menjadi Calon Walikota Makassar di Pilkada 2020.
5	Lima pelaku penusukan salah satu pendukung pasangan calon Walikota Makassar di tangkap pihak kepolisian Polda Metro Jaya. Karena sakit jatung Satu orang tersangka (S) meninggal dunia saat akan di lakukan penangkapan, Jumat (13/11) . (Eds) https://t.co/W27KiS1Tp5

4.3Cleaning and pre-processing data

The data should be subjected to a process of cleaning and pre-processing in order to establish consistency and eliminate any extraneous or noisy information. The processing of text data involves several essential activities, including tokenization, elimination of stopwords, stemming/lemmatization, and handling of special characters.

a.Data cleaning process

Figure 4 shows the stages of the attribute cleaning process using the replace operator in the RapidMiner tool to clean username, hashtag, uniform resource locator (URL) mention, punctuation and symbol attributes. This process is carried out to remove inconsistent or relevant data.

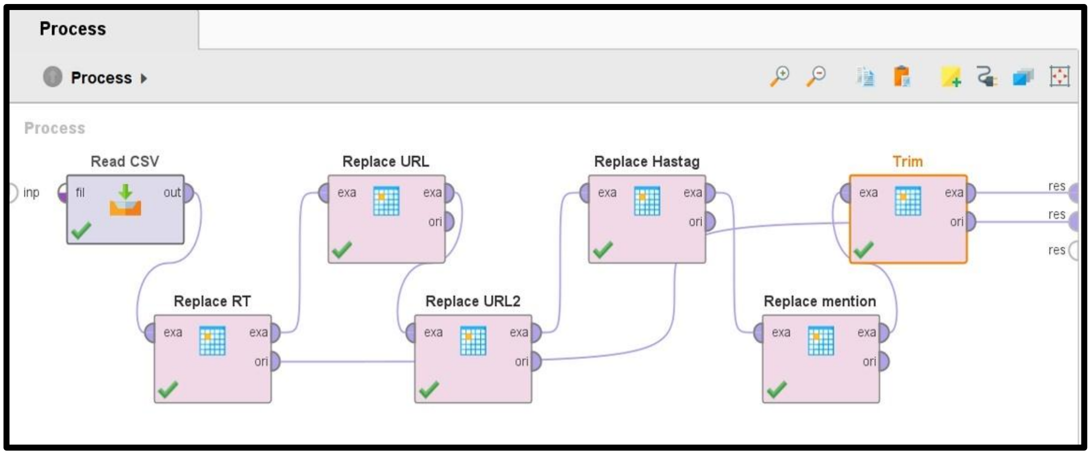


Figure 4 Design operator replace

b.Pre-processing data

In the data pre-processing stage, the tokenization stage is carried out to separate each word of a sentence into word units, followed by the transformation stage for each letter so that the input data changes uniformly into lowercase format. The next stage is to filter stop words by eliminating common words found in the text, can be seen in *Figure 5*. Finally, filter tokens based on the number of characters of each word to produce a more effective text representation. An illustration of this process can be seen in *Figure 6*.

4.4Tweet data sentiment analysis process using naïve Bayes

This study will employ machine learning methodologies, with a specific focus on the naïve Bayes algorithm, which will be executed through the utilization of the RapidMiner program. The first phase involves developing a classification model to classify tweets into positive, negative, or neutral categories, thereby determining the expressed attitude. The first methodology was the utilization of Twitter data that had been manually classified. *Table 2* displays examples of sentiment analysis data used for the classification model.

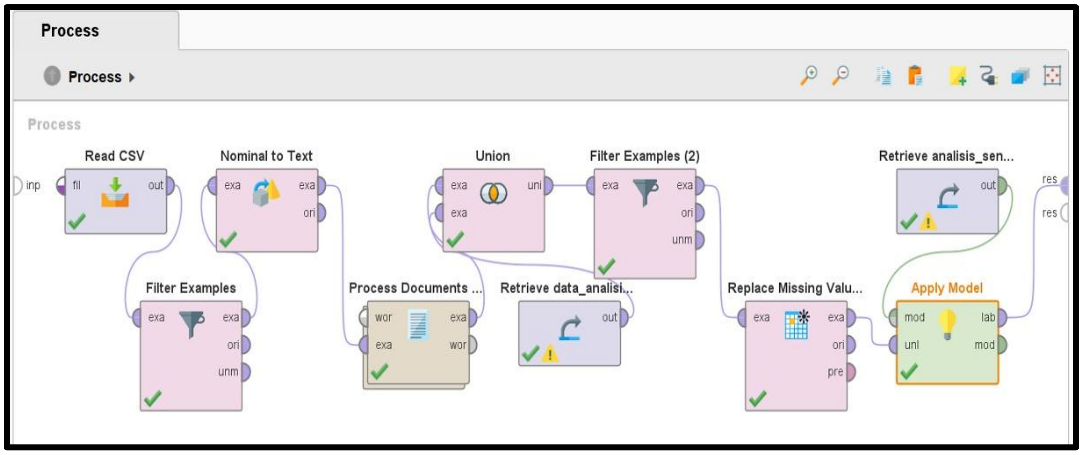


Figure 5 Process pre-processing data

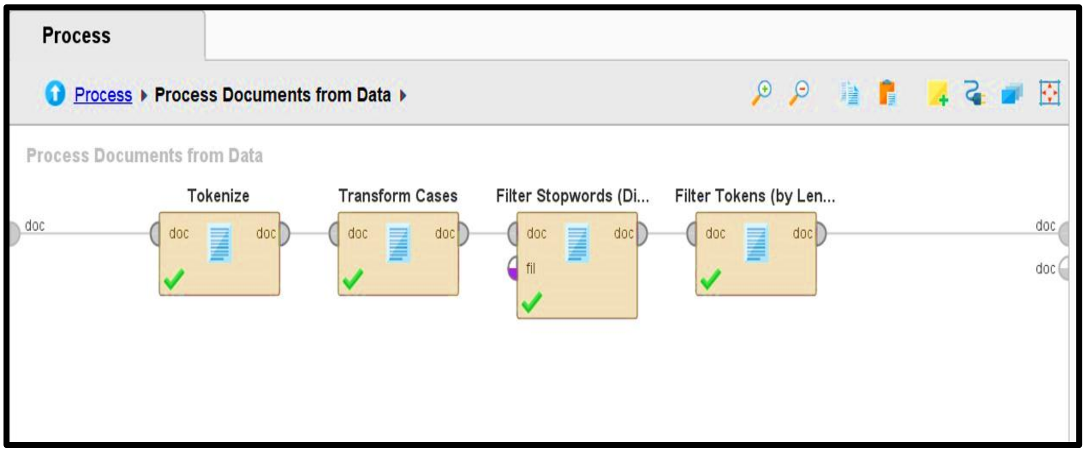


Figure 6 Sub-process pre-processing data

Table 2 Sample classification models data

Date;"Tweet"	Sentiment
2020-07-01 06:38:09+00:00";"Apa harapan kalian kepada calon Walikota Makassar yang baru nanti?	Positif
2020-12-05 15:41:18+00:00";"Bukankahkah dia calon Walikota MakassarPasti dia punya dongeng bagus."	Negatif
2020-11-08 01:56:12+00:00;"Penikaman di luar gedung Kompas TV yg terekam kamera pengawasPelaku berdiri mengawasi korbanmenikamlalu berlari naik ke motor temannyaIni terjadi saat debat calon walikota/wakil walikota makassar	Negatif
2020-11-09 12:13:12+00:00;"Momen ter-SAVAGE di debat"	Neutral
2020-08-31 15:46:44+00:00;"dari 3 calon pasangan walikota makassar yg maju inisiapaami nanti yg punya program tuk urus public transport di makassar?"	Positif
2020-11-17 12:25:46+00:00;"bahas calon walikota Makassar"	Neutral
2020-11-24 13:21:06+00:00;"bagus ini kalo bikin cek fakta terkait penyampaian2 setiap pasangan calon walikota makassar di tiap debat."	Negatif
2020-02-23 06:38:29+00:00;"sy mendukung ir.h.danny pomanto jadi calon walikota makassar (2021-2024krn dia adik kelas sy di fakultas teknik universitas hasanuddin	Positif
2020-10-21 04:17:46+00:00;"Video viral salah satu pendukung calon walikota Makassar	Neutral
2020-10-14 11:58:25+00:00;"Hasil survey menunjukan Pasangan ADAMA unggul jauh dengan pasangan calon walikota makassar lainnya"	Positif
2020-08-24 06:17:10+00:00;"Alhamdulillah sdh terpakai tenda tenda bantuan dr kandidat calon walikota Makassar APPI ARB"	Positif
2020-07-01 15:58:33+00:00;"Gerindra resmi mendukung Mohammad Ramdhan Pomanto (menjadi Calon Walikota Makassar di Pilkada 2020."	Positif

Date;"Tweet"	Sentiment
2020-11-13 09:49:33+00:00;"Lima pelaku penusukan salah satu pendukung pasangan calon Walikota Makassar di tangkap pihak kepolisian Polda Metro JayaKarena sakit jatung Satu orang tersangka (Smeninggal dunia saat akan di lakukan penangkapanJumat (13/11(Eds"	Negatif
2020-08-22 04:20:43+00:00;"Alhamdulillah Deklarasi APPI ARB calon walikota Makassar berjalan lancar sesuai dg protokol covi 19"	Positif
2020-11-07 12:34:33+00:00;"Reply twit ini dengan nama calon walikota makassar andalanmu"	Positif
2019-10-14 03:21:27+00:00;"Calon Pemimpin Cerdas Siap Bertarung di Pilwali Kota Makassar 2020"	Positif
2020-11-07 12:08:59+00:00;"Komandan Kompi 2 Batalyon A Pelopor Satbrimob Polda Sulsel IPTU Muh SyukriS.Sos mengecek Kesiapan Personel Back Up Polrestabes dalam rangka Debat Kandidat Pemilihan Calon Walikota Makassar"	Positif
2020-11-21 14:06:28+00:00;"Rocky Gerung kampanyekan Calon Walikota Makassar Appi-Rahman"	Positif
2019-10-14 05:31:19+00:00;"Siang hari ini saya mengembalikan formulir pendaftaran bakal calon Walikota Makassar di Partai Perindo Kota Makassar."	Positif
2019-10-10 04:46:01+00:00;"Alhamdulillahsiang ini saya yg diwakili tim pemenangan mengambil fomulir pendaftaran bakal calon walikota Makassar di kantor PKB Makassarjalan A.R Hakim."	Positif
2020-10-14 11:54:26+00:00;"Popularitas fatmawati rusdi melejitdibanding tiga calon walikota makassar lainnya"	Neutral

a.Design operator classification models

Figure 7 shows a classification model construction process using RapidMiner's "Read Csv" operator, "Examples Filter" operator for selective data extraction, and "Nominal to Text" operator for textual conversion. The data presented in *Table 3* represents the findings obtained through the application of the TF-IDF processing approach. The extraction procedure yielded a set of labeled data, namely 39 tweet data points. Additionally, the analysis identified 224 significant word occurrences, which serve as characteristics in the dataset.

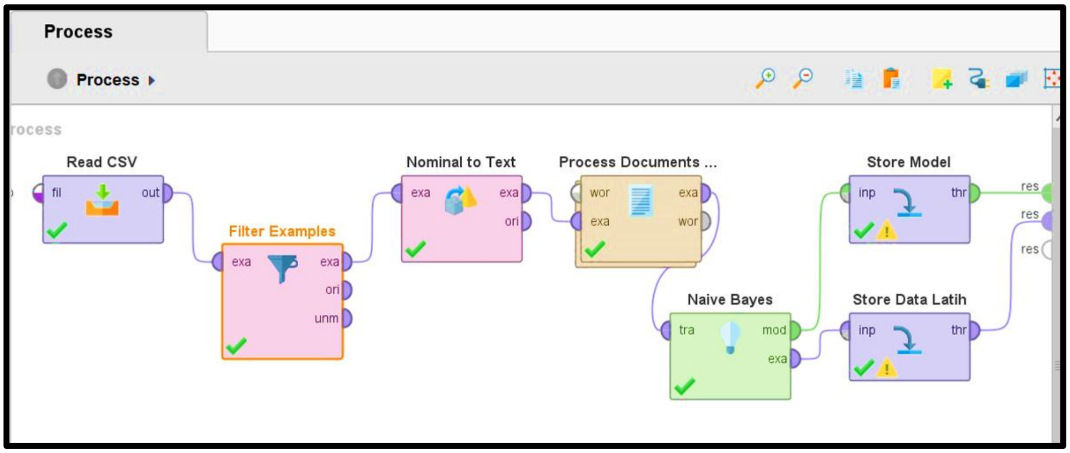


Figure 7 Design operator classification model

Table 3 TF-IDF result

Row. No.	Sentimen	Text	adama	adik	aksa	aksi	alhamdu	appi
1	Positif	Penikaman di luar gedung Kompas TV yg terekam kamera Ini terjadi saat debat calon walikota/wakil walikota makassar	0	0	0	0	0	0
2	Negatif	harapan calon walikota makassar	0	0	0	0	0	0
3	Negatif	bukankahkah calon walikota makassarpasti dongeng bagus	0	0	0	0	0	0
4	Neutral	momen savage debat	0	0	0	0	0	0
5	Positif	calon pasangan walikota makassar maju inisiapaami program urus public transport makassar	0	0	0	0	0	0

Row. No.	Sentimen	Text	adama	adik	aksa	aksi	alhamdu	appi
6	Neutral	bahas calon walikota makassar	0	0	0	0	0	0
7	Negatif	bagus kalo bikin fakta terkait penyampaian pasangan calon walikota makassar debat	0	0	0	0	0	0
8	Positif	mendukung danny pomanto calon walikota makassar adik kelas fakultas teknik universitas hasanuddin mohon teman jurnalis bantu pemberitaannya kania sutisnawinata avimalik andi fifi aleyda fitri metrotv	0	0.219	0	0	0	0
9	Neutral	video viral salah pendukung calon walikota makassar memburu mahasiswa unras mengeluarkan senjata tajamdibawa jembatan over tindakan melanggar hukum	0	0	0	0	0	0
10	Positif	hasil survey menunjukan pasangan adama unggul pasangan calon walikota makassar	0.384	0	0	0	0	0
11	Positif	alhamdulillah terpakai tenda tenda bantuan kandidat calon walikota makassar appi	0	0	0	0	0	0.306
12	Positif	gerindra resmi mendukung mohammad ramdhan pomanto calon walikota makassar pilkada	0	0	0	0	0	0
13	Negatif	pelaku penusukan salah pendukung pasangan calon walikota makassar tangkap kepolisian polda metro jayakarena sakit jatung orang tersangka smeninggal dunia lakukan penangkapanjumat	0	0	0	0	0	0

b. Sentiment analysis model design

The design depicted in *Figure 8* is a model for sentiment analysis. The current stage involves the implementation of the naïve Bayes classification model that was previously implemented. Currently, the application employs a methodology to ascertain sentiment from unlabelled Twitter data. This process involves the automated labelling of data using a naïve Bayes model that has been trained with existing data. *Table 4* presents the outcomes of sentiment analysis conducted using the naïve bayes algorithm in the RapidMiner program. This analysis was performed on Twitter data in order to determine the sentiment expressed in the tweets. The results of this analysis include the prediction values for sentiment data.

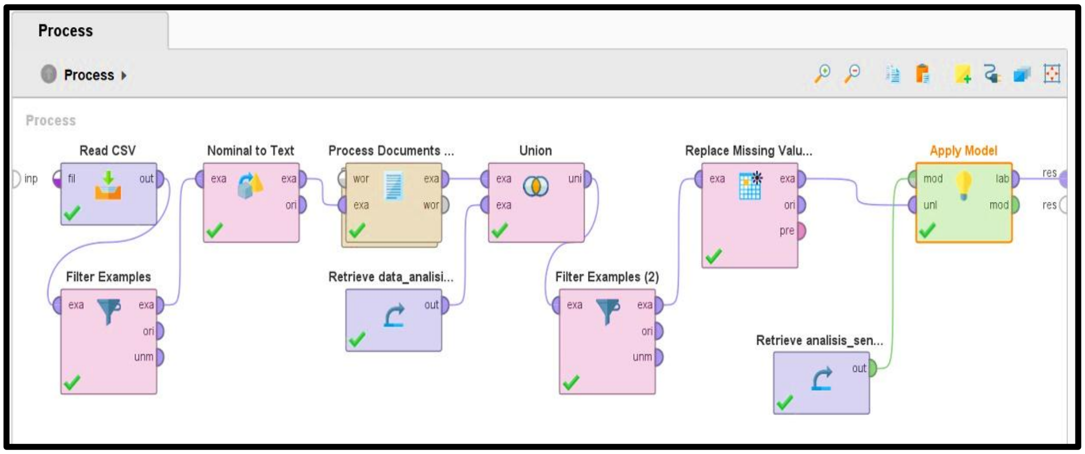


Figure 8 Sentiment analysis model design

Table 4 Results of naïve Bayes sentiment analysis

Row No	Sentiment	Prediction (sentiment)	Confidence (positif)	Confidence (negatif)	Confidence (neutral)	Text	abdillah	acara
1	?	Positif	1	0	0	dampingi appiirfan darmawan panggilan nurani makassar calon walikota makassarmunafri arifuddin appikembali bertarung kandidat calon wali kota makassar mendatangjumat	0	0
2	?	Positif	1	0	0	penanggulangan kemacetan masuk daftar program calon walikota makassarandi mustaman	0	0
3	?	Positif	1	0	0	abdillah natsir golkar terbuka calon walikota makassar	0.519	0
4	?	Positif	1	0	0	penentuan nomor urut calon walikota makassarkomphotel harper rame politik ketimbang protokol kesehatan politik	0	0
5	?	Positif	1	0	0	hidupkan budaya lokalbegini rancangan program calon walikota makassarandi mustaman	0	0
6	?	Positif	1	0	0	calon walikota makassar mendaftar partai golkar makassar terkini	0	0
7	?	Positif	1	0	0	sukriansyah slatief kandidat walikota makassar	0	0
8	?	Negatif	0	1	0	ewako appi calon walikota makassar	0	0
9	?	Neutral	0	0	1	mencari calon walikota makassar menyelesaikan parkir liarmiris gerai mini market parkir liar	0	0
10	?	Neutral	0	0	1	figur bermunculan calon walikota makassar	0	0
11	?	Neutral	0	0	1	Cerita	0	0
12	?	Neutral	0	0	1	calon walikota makassar rupanyahehhehe	0	0
13	?	Negatif	0	1	0	kallasalah jaringan bisnis munafri arifuddin appi appi	0	0

Row No	Sentiment	Prediction (sentiment)	Confidence (positif)	Confidence (negatif)	Confidence (neutral)	Text	abdillah	acara
						calon walikota makassar menantu aksa mahmudjika appi walikotadia cepat kaya kalla group wapres		

Table 5 shows the nominal values from a sentiment analysis process, which involved combining training and test data. With 39 instances of training and 215 instances of test, 254 results were obtained. Imputing missing values, a prediction dataset of 194 tweet data instances was created. Out of these, 137 were classified as positive, 28 as negative, and 28 as neutral.

Table 5 Nominal value

Index	Nominal value	Absolute count	Fraction
1	Positif	137	0.706
2	Negatif	29	0.149
3	Neutral	28	0.144

Figure 9 displays the results of sentiment analysis performed on Twitter data related to the Makassar mayoral race. The sentiment analysis results are visually depicted using graphs, which display the distribution of positive sentiment in 137 tweets, negative sentiment in 29 tweets, and neutral sentiment in 28 tweets.

4.5Tweet data clustering process using k-means

The k-means method was employed to cluster Twitter data related to the Makassar city election using the RapidMiner program. The tweet data was pre-processed and saved in CSV format to facilitate efficient mining. The k-means technique partitions data into clusters based on center distance. The RapidMiner software implemented the k-means Design, which uses data clustering to design the k-means model, several examples of clustering using k-means data used for the design model k-means are shown in *Table 6*.

a. Design model k-means

Figure 10 depicts the k-means model implemented in a rapid mining tool, which encompasses multiple steps in the grouping procedure. The process begins by gathering 255 tweets. Next, the nominal to text operator is utilized to transform nominal data into text. Finally, the attributes of single and select attribute data are filtered.

b. Cluster model

This *Figure11* represents the results of a clustering model applied to a dataset, showing many items are in each cluster. Here is a breakdown of the cluster:

- Cluster 0: contains 39 items.
- Cluster 1: contains 129 items.
- Cluster 2: contains 26 items.
- Cluster 3: contains 61 items.

The total number of items in all clusters is 255, which is the sum of the items in each cluster and *Table 7* is an example of data used as a source to determine each cluster.

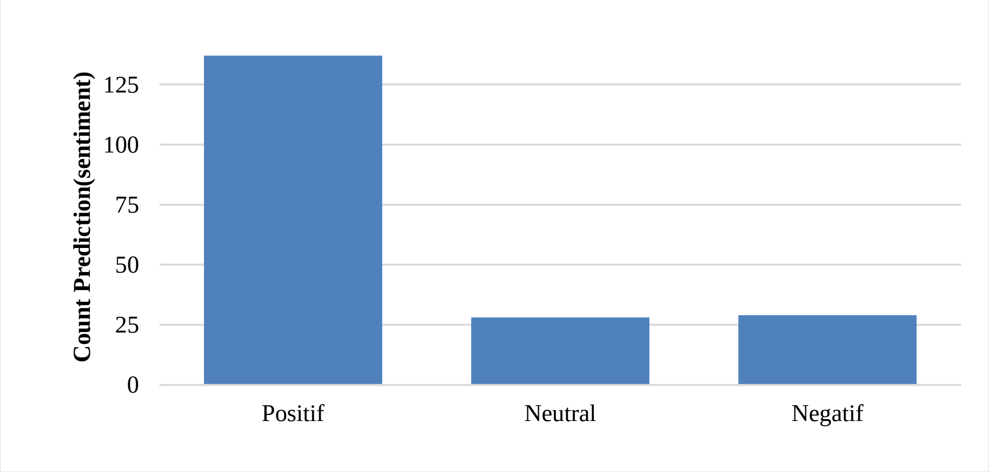


Figure 9 Sentiment analysis results graph

Table 6 Sample clustering data

No.	Date";"Tweet
1	2020-11-08 01:56:12+00:00";"Penikaman di luar gedung Kompas TV yg terekam kamera pengawasPelaku berdiri mengawasi korbanmenikamlalu berlari naik ke motor temannyaIni terjadi saat debat calon walikota/wakil walikota makassar,
2	2020-07-01 06:38:09+00:00";"Apa harapan kalian kepada calon Walikota Makassar yang baru nanti?,
3	2020-12-05 15:41:18+00:00";"Bukankahkah dia calon Walikota MakassarPasti dia punya dongeng bagus.
4	2020-11-09 12:13:12+00:00";"Momen ter-SAVAGE di debat
5	2020-08-31 15:46:44+00:00";"dari 3 calon pasangan walikota makassar yg maju inisiapaami nanti yg punya program tuk urus public transport di makassar?
6	2020-11-17 12:25:46+00:00";"bahas calon walikota Makassar
7	2020-11-24 13:21:06+00:00";"bagus ini kalo bikin cek fakta terkait penyampaian2 setiap pasangan calon walikota makassar di tiap debat.
8	2020-02-23 06:38:29+00:00";"sy mendukung ir.h.danny pomanto jadi calon walikota makassar (2021-2024krn dia adik kelas sy di fakultas teknik universitas hasanuddin,Mohon teman jurnalis bantu pemberitaannya w/ kania sutisnawinata avimalik andi fifi aleyda fitri_metrotv
9	2020-10-21 04:17:46+00:00";"Video viral salah satu pendukung calon walikota Makassar,' memburu mahasiswa unras dengan mengeluarkan senjata Tajamdibawa jembatan fly over Apakah Tindakan semacam ini tidak melanggar HUKUM ?
10	2020-10-14 11:58:25+00:00";"Hasil survey menunjukan Pasangan ADAMA unggul jauh dengan pasangan calon walikota makassar lainnya
11	2020-08-24 06:17:10+00:00";"Alhamdulillah sdh terpakai tenda tenda bantuan dr kandidat calon walikota Makassar APPI ARB
12	2020-07-01 15:58:33+00:00";"Gerindra resmi mendukung Mohammad Ramdhan Pomanto (menjadi Calon Walikota Makassar di Pilkada 2020.
13	2020-11-13 09:49:33+00:00";"Lima pelaku penusukan salah satu pendukung pasangan calon Walikota Makassar di tangkap pihak kepolisian Polda Metro JayaKarena sakit jatung Satu orang tersangka (Smeninggal dunia saat akan di lakukan penangkapanJumat (13/11(Eds

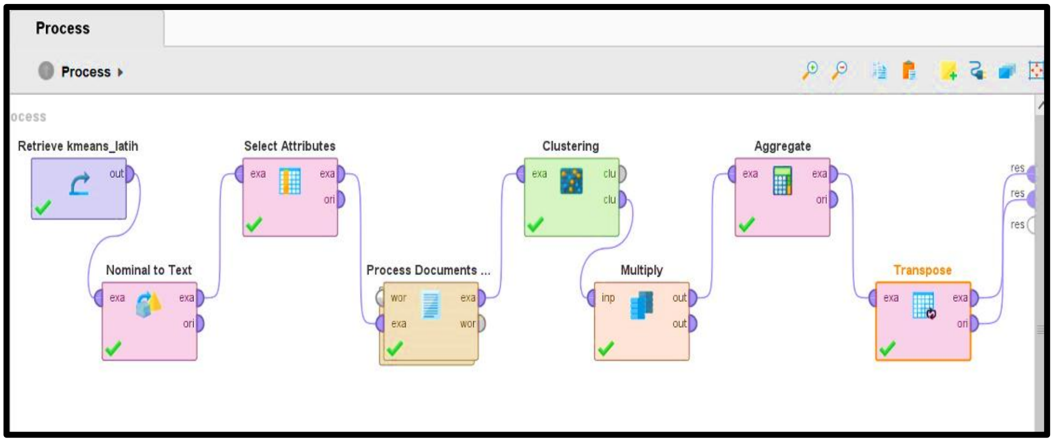
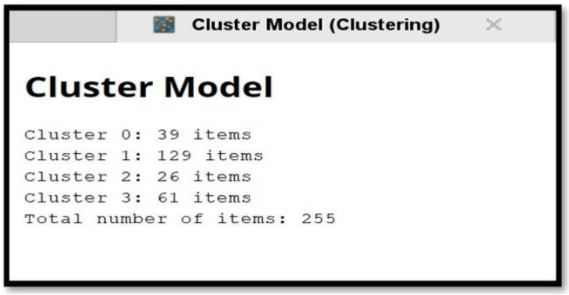


Figure 10 Design model k-means



Cluster Model

```
Cluster 0: 39 items
Cluster 1: 129 items
Cluster 2: 26 items
Cluster 3: 61 items
Total number of items: 255
```

Figure 11 Cluster result

Table 7 Centroid value final results

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
arifuddin	0	0	0.076923	0
been	0	0.090909	0	0
cluster	0	0	0.076923	0
danny	0	0	0.076923	0
debat	0	0	0	0.333333
diskominfo	0	0.090909	0	0
dukung	0	0	0.076923	0

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
fatma	0	0	0.076923	0
kandidat	0	0	0	0.333333
kosong	0	0.090909	0	0
kota	0	0	0.076923	0
kotak	0	0.090909	0	0
layanan	0.333333	0	0	0
maju	0	0	0.076923	0
masyarakat	0	0.090909	0	0

Figure 12 and *Figure 13* demonstrates the varying cluster value findings due to iteration number, illustrating the k-means algorithm's graphical line and x and y axis representation of clustering outcomes.

c. Cluster model visualizer

Figure 14 displays the classification of data into four distinct groups, highlighting the performance of keywords. The data consists of 255 tweets, with the first cluster including 39 instances, the second cluster containing 129 instances, the third cluster containing 26 instances, and the fourth cluster containing 61 examples.

Figure 15 and *Figure 16* shows is graphical line x and y axis the data being grouped into four clusters, with 255 tweets categorized into each cluster. The 30 keywords are distributed evenly, with cluster 0 having 39 keywords, cluster 1 having 129 items, cluster 2 having 26 items, and cluster 3 having 61 items.

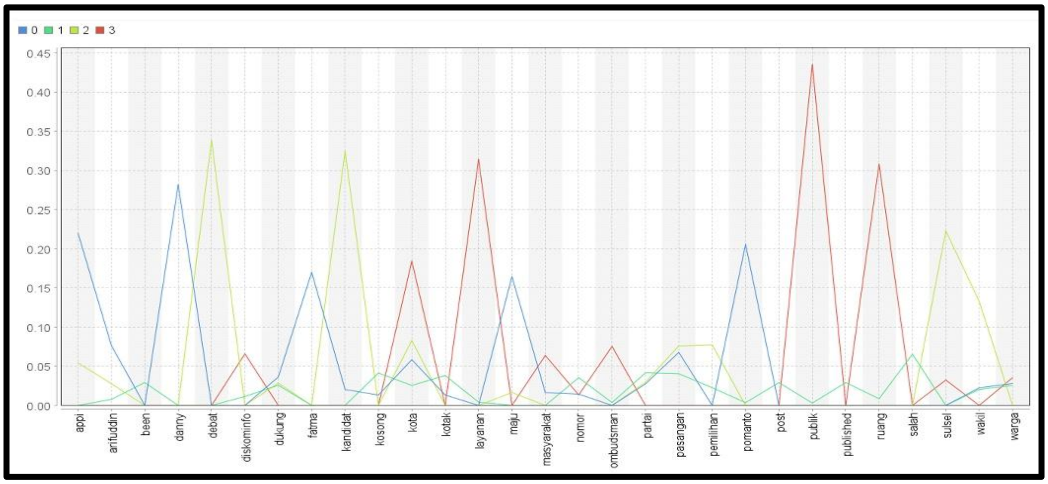


Figure 12 Clustering Results k-means line graph

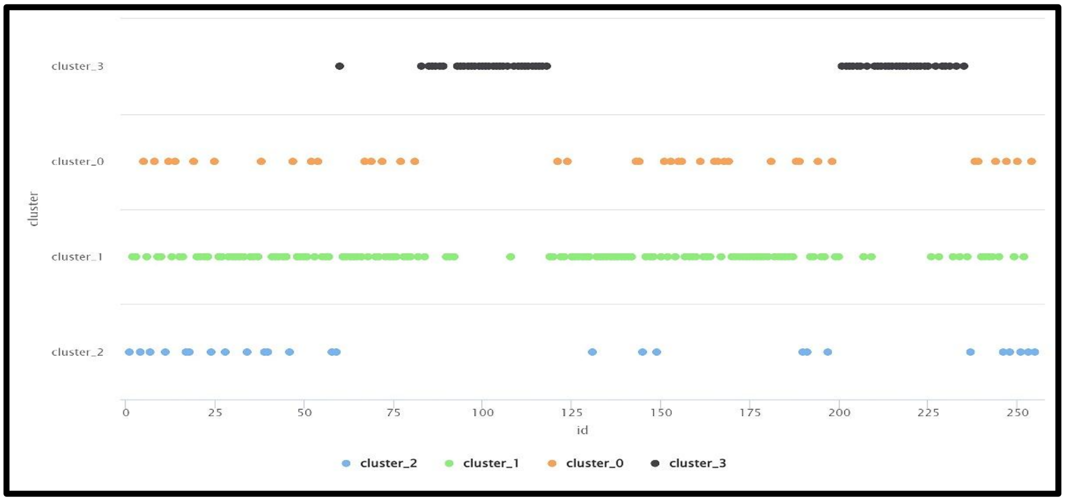
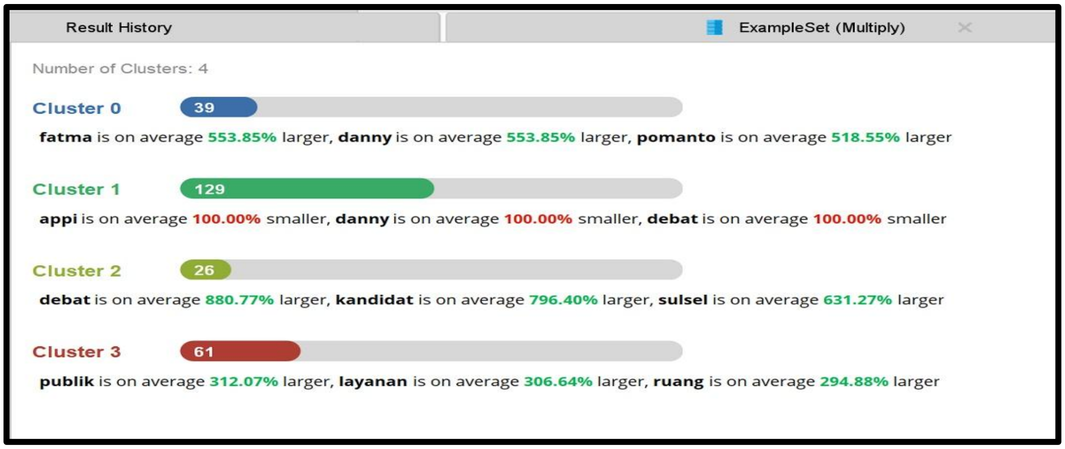


Figure 13 Clustering Results k-means X and Y Axis



Number of Clusters: 4

Cluster 0 39

fatma is on average **553.85%** larger, **danny** is on average **553.85%** larger, **pomanto** is on average **518.55%** larger

Cluster 1 129

appi is on average **100.00%** smaller, **danny** is on average **100.00%** smaller, **debat** is on average **100.00%** smaller

Cluster 2 26

debat is on average **880.77%** larger, **kandidat** is on average **796.40%** larger, **sulsel** is on average **631.27%** larger

Cluster 3 61

publik is on average **312.07%** larger, **layanan** is on average **306.64%** larger, **ruang** is on average **294.88%** larger

Figure 14 Results of visualizer model clustering

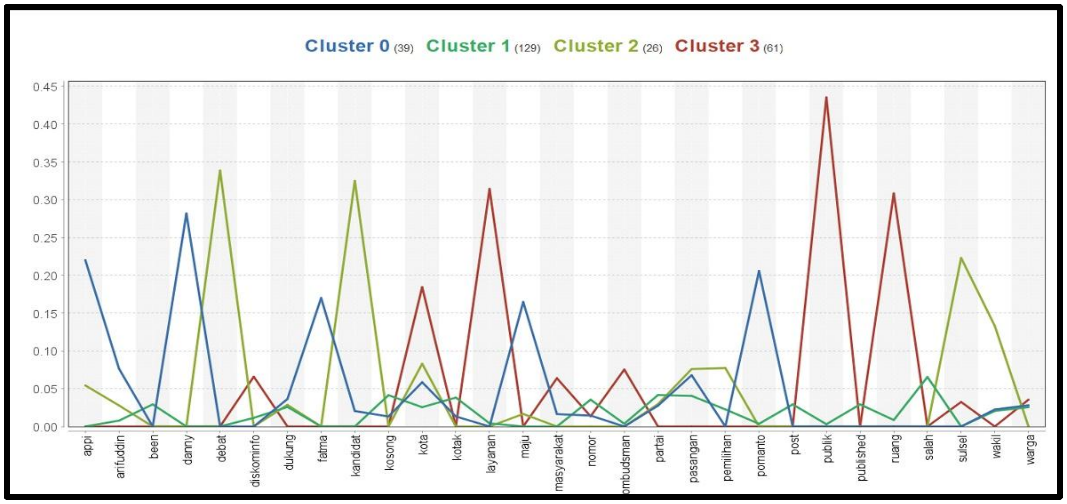


Figure 15 Clustering model visualizer result graph

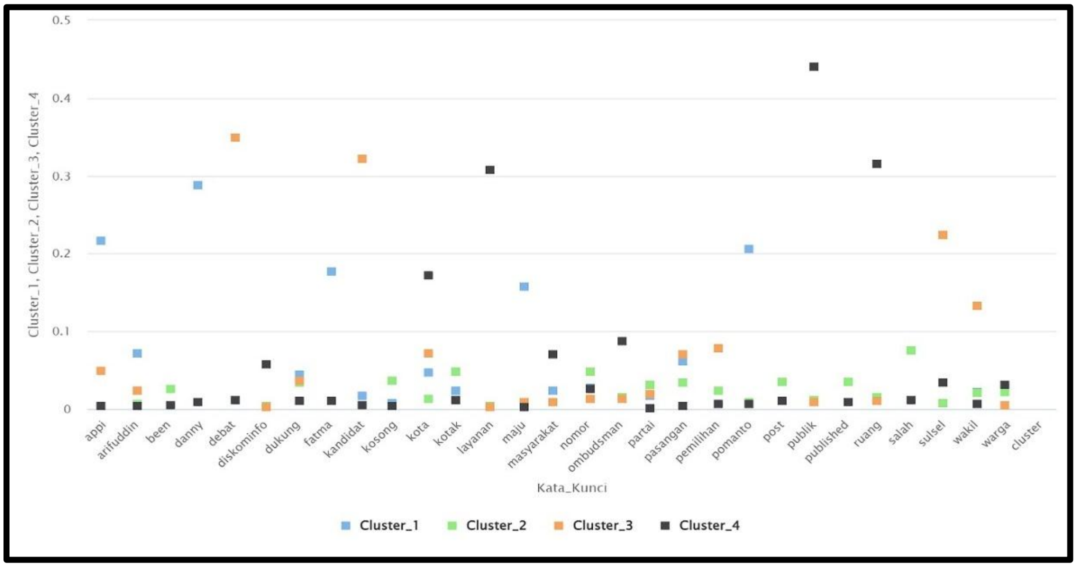


Figure 16 Clustering model visualizer result x and y axis

4.6Classification process with SVM

The study focuses on hand labelling using the SVM technique, a method that is not well-suited for handling polynomial attributes. The model is compared to the naïve Bayes algorithm for sentiment analysis, which categorizes outputs into negative, positive, and neutral categories. The SVM method uses binomial characteristics for labelling, enhancing its effectiveness [63].

The SVM model is designed in *Figure 17* by importing twitter data using the read CSV operator, assigning a role to polarity labeling with the choose attribute target role label operator, and transforming the data from nominal format to text format.

Once the model design is complete, data for training and testing is created, with 15% of the data allocated for training and 85% for testing. Cross validation is employed as a component in the SVM method, encompassing sub-processes that involve datasets and testing data, as depicted in *Figure 18.*

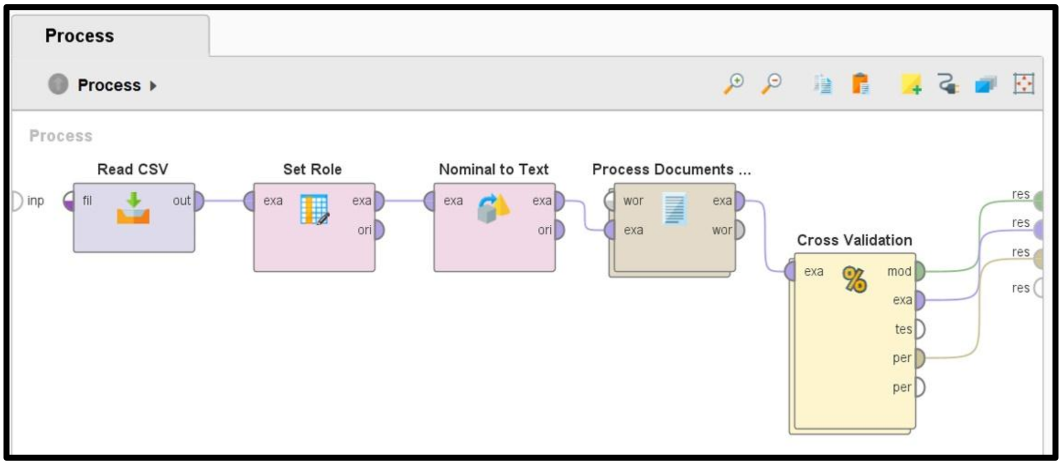
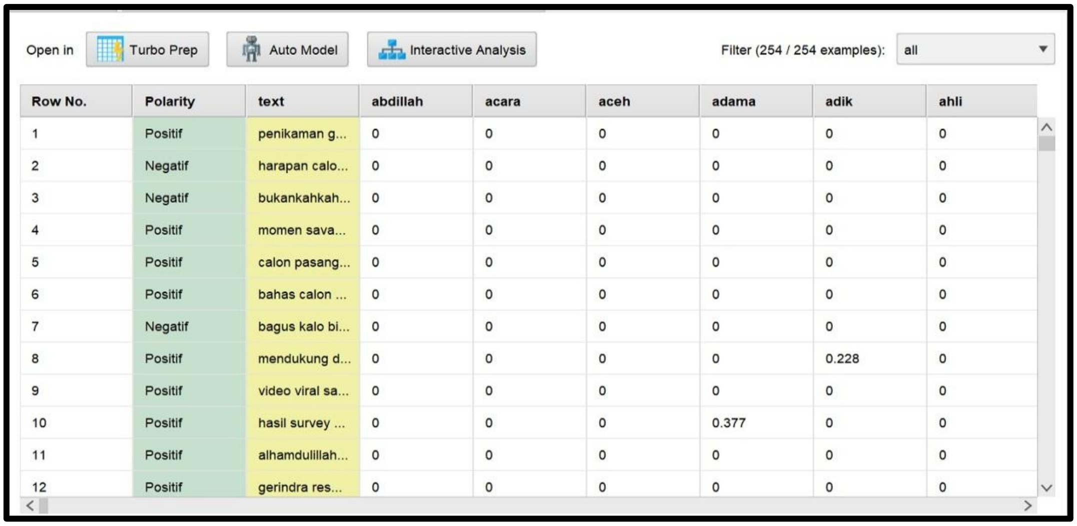


Figure 17 Design model SVM



Row No.	Polarity	text	abdillah	acara	aceh	adama	adik	ahli
1	Positif	penikaman g...	0	0	0	0	0	0
2	Negatif	harapan calo...	0	0	0	0	0	0
3	Negatif	bukankahkah...	0	0	0	0	0	0
4	Positif	momen sava...	0	0	0	0	0	0
5	Positif	calon pasang...	0	0	0	0	0	0
6	Positif	bahas calon ...	0	0	0	0	0	0
7	Negatif	bagus kalo bi...	0	0	0	0	0	0
8	Positif	mendukung d...	0	0	0	0	0.228	0
9	Positif	video viral sa...	0	0	0	0	0	0
10	Positif	hasil survey ...	0	0	0	0.377	0	0
11	Positif	alhamdulillah...	0	0	0	0	0	0
12	Positif	gerindra res...	0	0	0	0	0	0

Figure 18 Data testing result

4.7Validation

Validation is the evaluation of a trained model's performance, typically done using cross-validation methodology. This approach assesses the prediction precision of a model using techniques such as holdout, k-fold cross-validation, stratified k-fold cross-validation, and leave-p-out cross-validation. K-fold cross-validation partitions a dataset into homogeneous groups of equal size, where a portion is allocated for testing and the remaining is designated for training. This process is repeated multiple times [64].

The study evaluates two strategies: naïve Bayes and SVM are assessed using confusion matrix values, while k-means quality is evaluated using cluster distance performance. The sentiment analysis several data employed for evaluating the using confusion matrix testing, as shown *Table 8.*

Table 8 Sample confusion matrix testing Naïve Bayes and SVM

Date";"Tweet	**Polarity**
2020-11-08 01:56:12+00:00";"Penikaman di luar gedung Kompas TV yg terekam kamera pengawasPelaku berdiri mengawasi korbanmenikamlalu berlari naik ke motor temannyaIni terjadi saat debat calon walikota/wakil walikota makassar,	Positif
2020-07-01 06:38:09+00:00";"Apa harapan kalian kepada calon Walikota Makassar yang baru nanti?,	Negatif
2020-12-05 15:41:18+00:00";"Bukankahkah dia calon Walikota MakassarPasti dia punya dongeng bagus.	Negatif
2020-11-09 12:13:12+00:00";"Momen ter-SAVAGE di debat	Positif
2020-08-31 15:46:44+00:00";"dari 3 calon pasangan walikota makassar yg maju inisiapaami nanti yg punya program tuk urus public transport di makassar?	Positif
2020-11-17 12:25:46+00:00";"bahas calon walikota Makassar	Positif
2020-11-24 13:21:06+00:00";"bagus ini kalo bikin cek fakta terkait penyampaian2 setiap pasangan calon walikota makassar di tiap debat.	Negatif
2020-02-23 06:38:29+00:00";"sy mendukung ir.h.danny pomanto jadi calon walikota makassar (2021-2024krn dia adik kelas sy di fakultas teknik universitas hasanuddin,Mohon teman jurnalis bantu pemberitaannya w/ kania sutisnawinata avimalik andi fifi aleyda fitri_metrotv	Positif
2020-10-21 04:17:46+00:00";"Video viral salah satu pendukung calon walikota Makassar,' memburu mahasiswa unras dengan mengeluarkan senjata Tajamdibawa jembatan fly over Apakah Tindakan semacam ini tidak melanggar HUKUM ?	Positif
2020-10-14 11:58:25+00:00";"Hasil survey menunjukan Pasangan ADAMA unggul jauh dengan pasangan calon walikota makassar lainnya	Positif
2020-08-24 06:17:10+00:00";"Alhamdulillah sdh terpakai tenda tenda bantuan dr kandidat calon walikota Makassar APPI ARB	Positif
2020-07-01 15:58:33+00:00";"Gerindra resmi mendukung Mohammad Ramdhan Pomanto (menjadi Calon Walikota Makassar di Pilkada 2020.	Positif
2020-11-13 09:49:33+00:00";"Lima pelaku penusukan salah satu pendukung pasangan calon Walikota Makassar di tangkap pihak kepolisian Polda Metro JayaKarena sakit jatung Satu orang tersangka (Smeninggal dunia saat akan di lakukan penangkapanJumat (13/11(Eds	Negatif
2020-08-22 04:20:43+00:00";"Alhamdulillah Deklarasi APPI ARB calon walikota Makassar berjalan lancar sesuai dg protokol covi 19	Positif
2020-11-07 12:34:33+00:00";"Reply twit ini dengan nama calon walikota makassar andalanmu	Positif
2019-10-14 03:21:27+00:00";"Calon Pemimpin Cerdas,Siap Bertarung di Pilwali Kota Makassar 2020	Positif
2020-11-07 12:18:59+00:00";"Personel Batalyon A Pelopor Satbrimob Polda Sulsel melaksanakan Pengamanan Debat Kandidat Calon Walikota Makassar	Positif

4.7.1 K-fold cross validation naïve Bayes

Cross-validation techniques are used for confusion matrix evaluations using the naïve Bayes method, followed by the use of models and performance operators for accurate results, as shown in *Figure 19.* The naïve Bayes method achieved an accuracy of 64.96%, indicating its ability to classify data with a margin of error of around +- 5.85%, based on 244 instances of positive emotion, 37 instances of negative sentiment, and 16 instances of neutral sentiment, as shown is *Figure 20.*

4.7.2 K-fold cross validation SVM

The precision metric measures the agreement between predicted outcomes and desired data, with accuracy rates of 82.20% for positive, 25.00% for negative, and 2.86% for neutral predictions. The recall value measures the model's effectiveness in recovering information, with positive data accounting for 78.11%, negative data 18.92%, and neutral data 6.25%, as seen is *Figure 21.* The confusion matrix and SVM are used to evaluate the performance of naïve Bayes models. The SVM technique achieved an accuracy rating of 85.43%, indicating accurate classification of data. The margin of error is 1.90%, and 217 instances of good emotion and 37 instances of negative sentiment were observed. The precision rate for positive predictions is 85.43%, while negative predictions have a 0.00% accuracy rate. The recall value measures the model's efficacy in recovering information, with a 100% accuracy rate on positive data and a 0.00% accuracy rate on negative data, as shown is *Figure 22.*

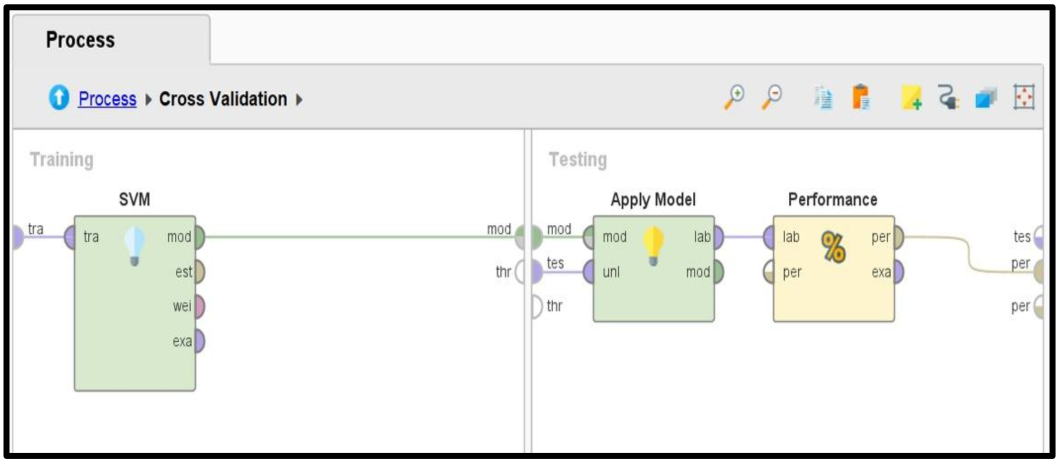
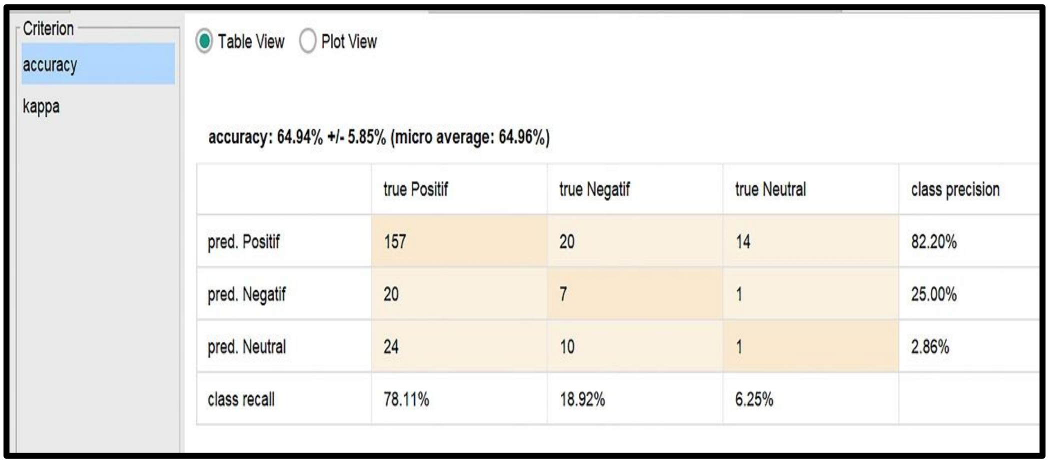


Figure 19 Naïve bayes validation model design



accuracy: 64.94% +/- 5.85% (micro average: 64.96%)

	true Positif	true Negatif	true Neutral	class precision
pred. Positif	157	20	14	82.20%
pred. Negatif	20	7	1	25.00%
pred. Neutral	24	10	1	2.86%
class recall	78.11%	18.92%	6.25%	

Figure 20 Confusion matrix naïve bayes

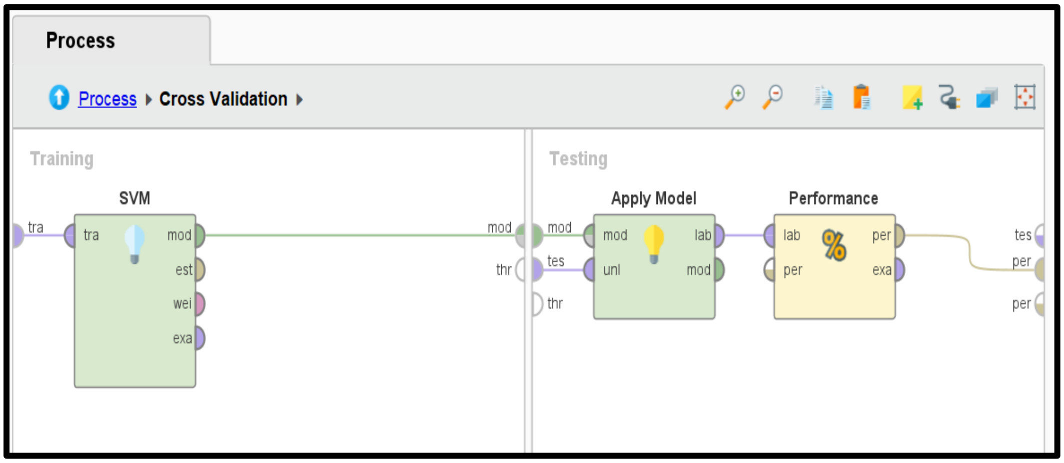


Figure 21 SVM validation model design

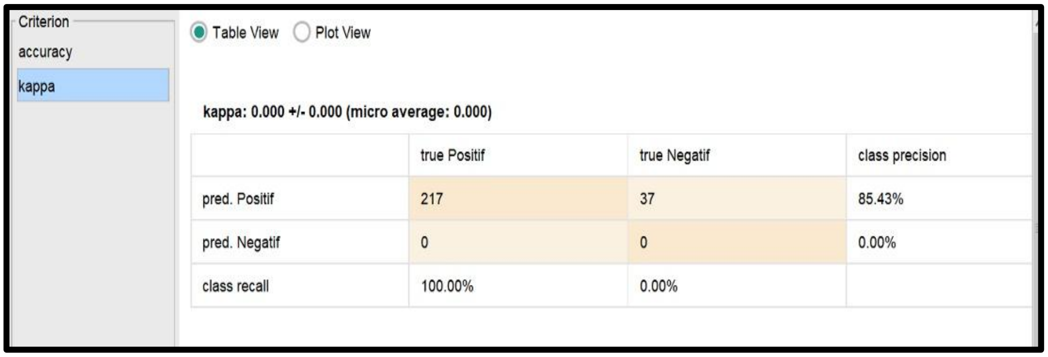
4.7.3 Comparison performance of Naïve Bayes and SVM in precision

a. Naïve Bayes:

The accuracy of the naïve Bayes algorithm was reported at 64.96%, indicating that it could classify data correctly with a margin of error of ±5.85%. This demonstrates its capability to identify the sentiment of tweets with a reasonable degree of precision. However, the document does not explicitly state the precision values for naïve Bayes across different sentiment categories (positive, negative, and neutral).

b. SVM:

The precision for SVM is detailed as follows: 82.20% for positive predictions, 25.00% for negative predictions, and 2.86% for neutral predictions. This suggests that SVM is significantly more effective at correctly identifying positive sentiments compared to negative and neutral sentiments. The accuracy of the SVM algorithm was reported at 85.43%, with a margin of error of ±1.90%. This higher accuracy rate compared to naïve bayes indicates that SVM is generally more effective in classifying the sentiment of tweets accurately.



kappa: 0.000 +/- 0.000 (micro average: 0.000)

	true Positif	true Negatif	class precision
pred. Positif	217	37	85.43%
pred. Negatif	0	0	0.00%
class recall	100.00%	0.00%	

Figure 22 Confusion matrix naïve bayes

5.Discussion

The data was gathered by extracting Twitter tweet data spanning from 2019 to 2020, supplemented with additional searches conducted during the same time period. The data was collected both prior to and following the mayoral election in Makassar. The data collecting procedure involves the retrieval of tweet data using Twitter crawling, which serves as a reference for analysis. Data pertaining to the Makassar mayoral election is acquired automatically through the use of relevant keywords. The process of collecting data using the Twitter API encompasses two methods: employing the RapidMiner tool and conducting data crawling with Python. From the results of crawling data, it can be concluded as follows:

- A total of 728 tweets related to the Makassar mayoral election in 2019 and 2020 were collected. Additional searches were conducted to supplement the dataset for the same time period. Subsequently, the data underwent a data pre-processing procedure resulting in a total of 254 tweet data.
- Sentiment analysis algorithms can misclassify tweets due to various factors, including linguistic nuances like irony and sarcasm, context and context, emoji and slang usage, preprocessing and feature extraction techniques, algorithmic restrictions, training data bias, and class imbalance. Factors like irony and sarcasm, ambiguity and context, emoji and slang usage, preprocessing techniques, algorithmic restrictions, training data bias, and class imbalance can all contribute to misclassification. Algorithms like naïve Bayes and SVM may struggle with sentiment classes with overlap, while training data bias can affect the effectiveness of the algorithm on diverse real-world datasets. Class imbalance can also result in misclassification of minority classes.
- The TF-IDF processing technique was used to extract data, resulting in 39 labelled tweet data and 224 significant word occurrences.
- The sentiment analysis process involved multiple stages, combining training data and test data. The training data consisted of 39 samples, while the test data had 215 samples. By using 15% of the data for training and 85% for testing, a total of 254 results were obtained. Additionally, a merger process was conducted to handle missing values, resulting in 194 tweet data with predicted sentiments. Out of these, 137 were predicted as positive sentiments, 29 as negative sentiments, and 28 as neutral sentiments.

- The accuracy of the naïve Bayes method was determined to be 64.96%. This indicates that the naïve Bayes model is capable of correctly classifying data within a margin of error of ± 5.85%. There was a total of 244 instances of positive sentiments, 37 instances of negative sentiments, and 16 instances of neutral sentiments. In addition, the SVM algorithm achieved an accuracy score of 85.43%, indicating that 85.43% of the SVM model accurately classifies the data. The margin of error indicates that the sample error is plus or minus 1.90%. There are a total of 217 instances of positive sentiments and 37 instances of negative sentiments.
- This study's primary goal is to use text-mining algorithms to explore the topic of political communication analysis. Researchers specifically looked at the use of k-means, SVM, and Naive Bayes algorithms. Many academics have employed k-means and the Naive Bayes algorithm in past work, sometimes in conjunction with other techniques. Particularly in the context of political communication, the SVM method will be applied to categorize words or sentences into favourable, negative, or neutral categories. Next, phrases or statements are categorized using the k-Means clustering algorithm according to criteria such party affiliation, candidate associations, black campaign activities, public spaces, public facilities, and public services. Using SVM and the naive Bayes algorithm, a confusion matrix is used to carry out the validation process.
- Naïve Bayes is faster and easier for baseline models and sentiment investigations, while SVM performs more accurately in nonlinear situations and can model intricate patterns. SVM and Naïve Bayes are better suited for sentiment analysis jobs where labels are required, unlike k-means, which is unsupervised. Naïve Bayes is optimal for quick decisions on large datasets, while SVM is recommended for high-complexity data and k-Means for pattern identification. Each algorithm's effectiveness depends on the study's specific demands, accuracy, computational efficiency, and analytical goals.

5.1Limitation

The study on political communication during the Makassar mayoral elections faced several limitations, including limited access to data, a focus on a specific portion of Twitter users in the Makassar city region, and a restricted sample size. This limited scope may not provide a comprehensive understanding of social media voter opinion or political communication trends. The study's findings may not apply to other cities, regions, or nations due to their limited reach. The research design also limited the accuracy and generalizability of the findings, potentially compromising the reliability of the conclusions. Future studies should broaden their focus, use diverse data sources, and employ techniques to capture a more comprehensive and nuanced understanding of political communication on social media.

A complete list of abbreviations is listed in *Appendix I*.

6.Conclusion and future work

The study showcased the substantial impact of Twitter and other social media platforms on political discourse. Allow for the quick spread of information and promote active communication between political entities and the public. The utilization of hybrid analytical techniques, which involve the combination of k-means clustering, naïve Bayes, and SVM, yielded a strong and reliable framework for conducting sentiment analysis and text categorization. This methodological improvement facilitated a more precise and nuanced examination of political sentiments conveyed on social media. The investigation yielded crucial insights into prevailing popular opinion and mood patterns, which are of immense use to political strategists and campaign managers. Gaining insight into these emotions can aid in formulating more impactful campaign messages and methods that really connect with the voting population. The study emphasized the significance of employing strategic social media tactics in political campaigns, indicating that future campaigns can greatly profit from utilizing comprehensive sentiment analysis and engagement data to customize communication strategies. The research emphasized the need for additional studies to investigate the incorporation of advanced data analytics methods and broaden the range of data sources to enhance our comprehension of the influence of social media on political communication.

6.1 Future work

This study explores the limitations of political communication data in the context of mayoral elections, focusing on Twitter users. The limited data availability and scope of the study pose challenges for future research, as it requires a comprehensive analysis of social media's impact on political dynamics. The study also highlights the importance of space for researchers, as offline encounters still influence individuals' beliefs and political outlooks.

Technological advancements, such as new social media platforms, advanced data analysis methodologies, and innovative communication tools, are transforming the political communication landscape. The study also explores the potential for ethical considerations, such as privacy issues, the digital divide, and disinformation dissemination. Data analysis commonly uses sentiment analysis, classification, and clustering techniques, but their flexibility and inventiveness are crucial for addressing issues related to retrieving material from social media platforms. Integrating methodologies from linguistics, sociology, and psychology with machine learning techniques can provide a more comprehensive understanding of complex phenomena, such as political speech on social media platforms. Future campaign techniques entail developing a compelling storyline that deeply connects with the emotions of voters, effectively conveying the societal ramifications of the policy, and actively participating in regular conversations to establish trust. Anecdotes and personal endorsements establish a connection between the candidate's experience and the voter's experience, so adding a human touch to the campaign. An assertive call to action is employed to mobilize supporters and enhance participation. Humour is employed judiciously to guarantee its enduring effect. Disinformation monitoring is conducted to uphold and safeguard one's reputation. Data-driven analysis is employed to customize messaging for various voter segments and conduct experiments with alternative content forms. Communication is customized to suit social media platforms for the most effective online interaction.

Acknowledgment

Many thanks to all who assisted in preparing this research, particularly the Chairman of the Dipanegara Foundation and the Chancellor of Dipa Makassar University in Makassar, South Sulawesi, Indonesia, who provided funding for this project.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability

This research data was collected from a social media platform, specifically Twitter. The collection period spanned from 2019 to 2020 and included conversations related to the Makassar mayoral election, both pre-election and post-election. The data is publicly accessible and was gathered using specific commands and a structured approach to data collection, preprocessing, and analysis.

Author’s contribution statement

Jufri: Creating the study framework, playing a significant part in analyzing the findings and authoring the article, doing research, gathering data, writing, editing, and reviewing the work. They also oversaw the project and helped build text-mining algorithms. Gathered and sorted social media information, carried out preliminary data processing, and made contributions to the creation of the text mining program utilized in this study. **Aedah Binti Abd. Rahman:** Created the methodology, putting particular emphasis on sentiment analysis model construction and the use of natural language processing (NLP) tools. made a substantial contribution to the data analysis section, carried out statistical analysis, and made sure the conclusions were reliable. Contribute to the manuscript's critical revision for key intellectual content. **H. Suarga**: Interpretation and analysis of data, concentrating on finding patterns and trends in the data from social media. performs specialized case analysis and is crucial in applying the text mining paradigm to the Makassar Mayoral Election case study. Participate in the literature review to help place this study in the context of political influence analysis as a whole.

References

[1] Rizki JW. Social media as tools of communication and learning. QALAMUNA: Jurnal Pendidikan, Sosial, Dan Agama. 2023; 15(1):391-404.

[2] Widjayanto F, Naim S, Mokodenseho S. Maruf Amin’s political communication strategy in the 2019 election campaign: a lesson for anti-hoax politics. JWP (Jurnal Wacana Politik). 2022; 7(2):108-19.

[3] Hadma AM, Anggoro JD. Political communication in the age of social media. COMMICAST. 2022; 3(1):1-7.

[4] Budiana M. Use of social media in political communication. Science Info Journal: Informatics and Science. 2022; 12(1):18-24.

[5] Andriyendi DO, Nurman S, Dewi SF. Social media and its influence on the political participation of first-time voters in the regional elections. Journal of Education, Culture and Politics. 2023; 3(1):101-11.

[6] Akili RH, Achmad W. The role of political parties in the implementation of democratic general elections in the Indonesian state administration system. Journal of Law and Sustainable Development. 2023; 11(4):1-13.

[7] Olusola SA, Oluseye OO, Udoh MS, Iember KJ, Ayomiposi DO. A content analysis of the vision and mission statements of top ten leading universities in Africa. Cogent Education. 2022; 9(1):1-16.

[8] Acharya A, Grillo E, Sugaya T, Turkel E. Electoral campaigns as dynamic contests. Journal of the European Economic Association. 2024: 1-45.

[9] Laila AF, Muslimin K, Hakim L. The political communication tactics of the national democratic party (NasDem) for winning the legislative election. MUHARRIK: Jurnal Dakwah Dan Sosial. 2022; 5(1):27-44.

[10] Afrinal A, Yunus NR, Esfandiari F. Bawaslu institution and its contribution in resolving election

disputes. SALAM: Syar-i Social and Cultural Journal. 2023; 10(5):1697-716.

[11] Keith D. The impact of social media on political activism. International Journal of Humanity and Social Sciences. 2023; 1(1):16-29.

[12] https://www.alexanderthamm.com/en/blog/text-mining-basics-methods-and-application-cases/. Accessed 19 March 2024.

[13] Castanho SB, Proksch SO. Politicians unleashed? political communication on twitter and in parliament in western Europe. Political Science Research & Methods. 2022; 10(4).

[14] Purwanti S, Krisdinanto N, Budiman B, Rezky R. The usage of social media and political branding of public official during Covid-19. The Journal of Society and Media. 2022; 6(2):286-308.

[15] Nguyen TA, Bui TC, Sokolovskiy K. Social media and political communication. Journal of Ethnic and Cultural Studies. 2022; 9(4):187-200.

[16] Ellis JT, Reichel MP. Twitter trends in# Parasitology determined by text mining and topic modelling. Current Research in Parasitology & Vector-Borne Diseases. 2023; 4:100138.

[17] Mahoney J, Le LK, Lawson S, Bertel D, Ambrosetti E. Ethical considerations in social media analytics in the context of migration: lessons learned from a horizon 2020 project. Research Ethics. 2022; 18(3):226-40.

[18] Spinder S, Frasincar F, Matsiiako V, Boekestijn D, Brandt T. A text mining approach to identifying sustainability in the private sector. Computers in Industry. 2023; 149:103932.

[19] Cantú F, Carreras M. Presidential debates and electoral preferences in weakly institutionalised democracies: evidence from 32 Latin American elections. Journal of Politics in Latin America. 2023; 15(3):239-61.

[20] Singh L, Ahmad TA. Examining the impact of social media on youth and its future for history learning. Paramita: Historical Studies Journal. 2022; 32(2):253-62.

[21] Bringula R, Ulfa SA, Miranda JP, Atienza FA. Text mining analysis on students' expectations and anxieties towards data analytics course. Cogent Engineering. 2022; 9(1):2127469.

[22] Samuelsson R. A mixed methods approach to analyzing embodied interaction: the potentials of integrated mixed methods analysis of video interaction data. Journal of Mixed Methods Research. 2023: 1-17.

[23] Xu P, Ye Y, Zhang M. Exploring the effects of traditional media, social media, and foreign media on hierarchical levels of political trust in China. Global media and China. 2022; 7(3):357-77.

[24] Moekahar F, Ayuningtyas F, Hardianti F. Social media political campaign model of local elections in Pelalawan Regency Riau. Jurnal Kajian Komunikasi. 2022; 10(02):242-52.

[25] Saputro ER. Social media as a means of communication for candidates in the 2020 regional head election in makassar city. Journal of Prophetic Politics. 2022; 10(1):61-78.

[26] Attiah SJ, Alhassan I. Turning workplace gossip into a springboard for productive behaviour. Voice of the Publisher. 2022; 8(3):65-82.

[27] Sianturi K, Megasari A. The effectiveness of communication messages in politics. Journal of Social Research. 2023; 2(11):3988-96.

[28] Chang CW, Chang SH. The impact of digital disruption: influences of digital media and social networks on forming digital natives' attitude. SAGE Open. 2023; 13(3):21582440231191741.

[29] Laksana MO, Nurhaliza N. The impact of communication ethics on the communication quality in interpersonal relationships. Eduvest-Journal of Universal Studies. 2023; 3(5):989-95.

[30] Liu Z. The impact of interpersonal relationships on micro-influencer marketing. In SHS web of conferences 2024 (pp. 1-5). EDP Sciences.

[31] Yang B, Zhang R, Cheng X, Zhao C. Exploring information dissemination effect on social media: an empirical investigation. Personal and Ubiquitous Computing. 2023; 27(4):1469-82.

[32] Ong B, Toh DJ. Digital dominance and social media platforms: are competition authorities up to the task? IIC-International Review of Intellectual Property and Competition Law. 2023; 54(4):527-72.

[33] Stieglitz S, Dang-xuan L. Social media and political communication: a social media analytics framework. Social Network Analysis and Mining. 2013; 3:1277-91.

[34] https://info.populix.co/articles/data-mining-adalah/. Accessed 01 February 2024

[35] Shu X, Ye Y. Knowledge discovery: methods from data mining and machine learning. Social Science Research. 2023; 110:102817.

[36] Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications. 2023; 82(3):3713-44.

[37] Abdullah MH, Aziz N, Abdulkadir SJ, Alhussian HS, Talpur N. Systematic literature review of information extraction from textual data: recent methods, applications, trends, and challenges. IEEE Access. 2023; 11:10535-62.

[38] Putra SJ. Optimizing text categorization for Indonesian text using clustering label technique. Turkish Journal of Computer and Mathematics Education. 2021; 12(3):1483-91.

[39] Sudianto S, Sripamuji AD, Ramadhanti IR, Amalia RR, Saputra J, Prihatnowo B. Application of support vector machine and multi-layer perceptron algorithms in news topic classification. National Journal of Informatics Engineering Education: JANAPATI. 2022; 11(2):84-91.

[40] Kunam SR, Babu MR, Kumar PN. A sophisticated semantic analysis framework using an intelligent tweet data clustering and classification methodologies. Microprocessors and Microsystems. 2023; 98:104793.

[41] Ji D, Zhang P. Text data processing and classification algorithm based on data fusion and granular computing. Journal of Sensors. 2022; 2022:1-3.

[42] Raut P, Rathod R, Tidke R, Pande R, Rathod N, Kulkarni N. Sentiment analysis of twitter. International Journal for Research in Applied Science and Engineering Technology. 2022; 10(12):621-7.

[43] Cano-marin E, Mora-cantallops M, Sánchez-alonso S. Twitter as a predictive system: a systematic literature review. Journal of Business Research. 2023; 157:113561.

[44] https://www.questionpro.com/blog/conceptual-research/. Accessed 31 January 2024.

[45] Al-obaydy WI, Hashim HA, Najm YA, Jalal AA. Document classification using term frequency-inverse document frequency and K-means clustering. Indonesian Journal of Electrical Engineering and Computer Science. 2022; 27(3):1517-24.

[46] Rani R, Lobiyal DK. Performance evaluation of text-mining models with Hindi stopwords lists. Journal of King Saud University-Computer and Information Sciences. 2022; 34(6):2771-86.

[47] Uthirapathy SE, Sandanam D. Topic modelling and opinion analysis on climate change twitter data using LDA and BERT model. Procedia Computer Science. 2023; 218:908-17.

[48] Putra CK, Alamsyah A. Increase accuracy of naïve bayes classifier algorithm with k-means clustering for prediction of potential blood donors. Journal of Advances in Information Systems and Technology. 2022; 4(1):42-9.

[49] Subarkah P, Damayanti WR, Permana RA. Comparison of correlated algorithm accuracy naive bayes classifier and naive bayes classifier for heart failure classification. ILKOM Jurnal Ilmiah. 2022; 14:120-5.

[50] Bárcenas R, Gonzalez-lima M, Ortega J, Quiroz A. On subsampling procedures for support vector machines. Mathematics. 2022; 10(20):1-27.

[51] Sağlam F, Yıldırım E, Cengiz MA. Clustered Bayesian classification for within-class separation. Expert Systems with Applications. 2022; 208:118152.

[52] Tufail S, Riggs H, Tariq M, Sarwat AI. Advancements and challenges in machine learning: a comprehensive review of models, libraries, applications, and algorithms. Electronics. 2023; 12(8):1-43.

[53] Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Information Sciences. 2023; 622:178-210.

[54] Chaudhry M, Shafi I, Mahnoor M, Vargas DL, Thompson EB, Ashraf I. A systematic literature review on identifying patterns using unsupervised clustering algorithms: a data mining perspective. Symmetry. 2023; 15(9):1-44.

[55] Mishra A, Jatti VS, Messele SE, Jatti AV, Sisay AD, Khedkar NK, et al. Machine learning-assisted pattern recognition algorithms for estimating ultimate tensile strength in fused deposition modelled polylactic acid specimens. Materials Technology. 2024; 39(1):2295089.

[56] Maulana DJ, Saadah S, Yunanto PE. Kmeans-SMOTE integration for handling imbalance data in classifying financial distress companies using SVM and naïve bayes. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi). 2024; 8(1):54-61.

[57] Gumilar A, Prasetiyowati SS, Sibaroni Y. Performance analysis of hybrid machine learning methods on imbalanced data (rainfall classification). Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi). 2022; 6(3):481-90.

[58] Lo A, Pifarré AH, Renshon J, Liang S. The polarization of politics and public opinion and their effects on racial inequality in COVID mortality. Plos One. 2022; 17(9):e0274580.

[59] Dev DG, Bhatnagar V, Bhati BS, Gupta M, Nanthaamornphong A. LSTMCNN: a hybrid machine learning model to unmask fake news. Heliyon. 2024; 10(3):1-12.

[60] Liu L, Jones BF, Uzzi B, Wang D. Data, measurement and empirical methods in the science of science. Nature Human Behaviour. 2023; 7(7):1046-58.

[61] Perifanis NA, Kitsios F. Investigating the influence of artificial intelligence on business value in the digital era of strategy: a literature review. Information. 2023; 14(2):1-42.

[62] Li G, Li C, Wang C, Wang Z. Suboptimal capability of individual machine learning algorithms in modeling small-scale imbalanced clinical data of local hospital. Plos One. 2024; 19(2):e0298328.

[63] Janan F, Ghosh SK. Prediction of student’s performance using support vector machine classifier. Proceedings of the 11th annual international conference on industrial engineering and operations management 2021 (pp. 7078-88). IEOM Society International.

[64] Faisal M, Rahman TK. Determining rural development priorities using a hybrid clustering approach: a case study of South Sulawesi, Indonesia. International Journal of Advanced Technology and Engineering Exploration. 2023; 10(103):696-719.

Jufri, obtained a Bachelor of Computer Science degree in Information Systems from STMIK Dipanegara (Dipa Makassar University) Makassar Indonesia in 2004. Master's degree in Electro at Hasanuddin University Makassar Indonesia 2013. Currently a lecturer at Dipa University Makassar, apart from teaching, he is currently a PhD candidate at Asia e University (AeU) Kuala Lumpur Malaysia in the field of Information Computer and Technology. Research interests include Software Engineering, Computer Science, and Computer Networking. 19 year’s vast experience of academic, management, and research.
Email: jufri.ldp@dipanegara.ac.id

Aedah Binti Abd. Rahman received her doctoral degree from Malaysia University of Technology (Universiti Teknologi Malaysia (UTM)) in the area of Computer Science with Specialization in Software Engineering in 2014. She received her M. Comp. Sc. from University of Malaya (UM) in 2002 and B. Comp. Sc (Hons.) from University of Malaya (UM) in 1998. Presently she is an Associate Professor at School of Science and Technology (SST) and appointed as Head of Information Communication Technology and Services (ICTS) and Head of Asian Centre of E-Learning (ACE) at Asia e University (AeU), Subang Jaya. She has 23 year's vast experience of academic, industry, management, research, publication and consultancy. Her research area includes Software Engineering, Computer Science; Software Process and Quality Assurance; Software Quality Engineering; Requirements Engineering & Management; Software Evolution and Configuration Management; Software Project Management; Real-Time & Embedded System; Artificial Intelligence; Intelligent System; Machine Learning; Internet of Things (IoT); Big Data and Analytics; Agile and DevOps; Open Distance Learning (ODL), E-Learning and Digital Learning; Extended Reality (XR): Augmented Reality (AR), Virtual Reality (VR), Mixed Reality and Metaverse; Cloud Computing and Blockchain. She is also actively involved in training, consultancy and research project collaboration with various research and industry partners at national and international levels.

Email: aedah.abdrahman@aeu.edu.my

H. Suarga is a retired permanent lecturer at the Faculty of Mathematics and Science, Hasanuddin University, Makassar Indonesia. He holds a degree in Digital Electronic Instrument Physics from Gadjah Mada University, Yogyakarta, Indonesia in 1977, a Master of Science in Computer Science from the Asian Institute of Technology Bangkok, Thailand in 1983, a Master of Mathematics in Computer Science from Waterloo University, Canada in 1992, and a Doctor of Philosophy in Information. Systems from McMaster University Canada in 1997. His expertise is information systems, numerical computing, modeling systems, databases and programming. He is currently a part-time lecturer at Dipa Makassar University.

Email: ssuarga@hotmail.com

Appendix I

S. No.	Abbreviation	Description
1	API	Application Programming Interface
2	CSV	Comma Separated Value
3	FN	False Negative
4	FP	False Positive
5	GB	Giga Byte
6	JSON	JavaScript Object Notation
7	MAX	Maximum
8	MIN	Minimum
9	MySQL	My Structured Query Language
10	NLP	Natural Language Processing
11	RAM	Random Access Memory
12	SSD	Solid State Drive
13	SVM	Support Vector Machines
14	TF-IDF	Term Frequency-Inverse Document Frequency
15	TN	True Negative
16	TP	True Positive
17	URL	Uniform Resource Locator