Available online at http://iorajournal.org/indx.php/orics /index

Operations Research: International Conference Series

e-ISSN: 2722-0974
p-ISSN: 2723-1739

Vol. 6, No. 1, pp. 28-38, 2025

A Study of Public Opinion on the 2024 Regional Elections Using Cosine Similarity and TF-IDF Algorithms

Ari Riyanto Hidayat[1*], Arie Qurania[2], M. Iqbal Suriansyah [3]

[1,2,3]*Department of Computer Science, Universitas Pakuan, Jl. Pakuan PO Box 452, 16143 Bogor, Jawa Barat, Indonesia*

[*]*Corresponding author e-mail address*: *aririyanto123@gmail.com*

Abstract

The organization of general and regional head elections is an essential aspect of implementing an indirect democracy system. The primary objective of regional elections is to ensure that leaders are elected democratically and act on behalf of the people. The simultaneous holding of regional head elections has become a major topic of public discussion, giving rise to diverse opinions, particularly among Twitter users. This study aims to classify public opinion regarding the 2024 regional head elections using TF-IDF weighting, followed by a classification process with the Cosine Similarity algorithm. Of the 1,000 data points successfully scraped, 34.9% were classified as positive sentiment, 23.5% as negative sentiment, and 37.1% as neutral sentiment.

Keywords: Cosine Similarity, Regional Elections, TF-IDF Weighting.

1. Introduction

Regional head elections are a democratic process that requires citizens to elect regional leaders capable of advocating for their interests (Nurmandi, 2015). The implementation of simultaneous regional head elections began in 2015, and since then, the number of provinces holding these elections has continued to increase. Campaigns are actively promoted through both online and offline media, with social media serving as a key tool for digital campaigning. Twitter, in particular, has become a platform for candidates to communicate their aspirations and build personal branding. In various general elections across Asia in 2019, candidates extensively used Twitter and other social media channels to share slogans and policies, influence public perception, and mobilize support ahead of the campaign (Putra, 2023).

Sentiment analysis is a process that involves understanding, extracting, and processing textual data automatically to identify sentiment within opinion-based sentences (Pratama, 2019). It is used to determine an individual's opinion or stance on a particular issue, whether positive, negative, or neutral (Wu, 2022). Analyzing tweets from the Twitter social network and extracting opinions from them requires a text mining approach, preceded by a preprocessing stage. Previous research by Bening et al. (2018) successfully classified online news portal content using the TF-IDF and Cosine Similarity methods, which combine two key concepts: the frequency of data occurrence in a document and the inverse frequency value of an object containing the specified word. These numerical values are then used to calculate similarity between documents. In this study, 1,000 data points were successfully scraped from Twitter. The TF-IDF weighting method was applied before classifying the data into positive, negative, and neutral opinions using the Cosine Similarity algorithm. The objective of this classification process is to serve as a tool for assessing public opinion regarding the implementation of simultaneous regional elections in 2024.

2. Materials and Methods

2.1. Materials

The data used in this study is sourced from Twitter, consisting of sentiments expressed by users who will be voting in the 2024 regional elections. The collected data is categorized into three sentiment types: positive, negative, and neutral, totaling approximately 1,000 data points. The processed results will be visualized in a graph displayed on the web.

2.2. Methods

The classification of public opinion for the 2024 Pilkada was conducted using the Systems Development Life Cycle (SDLC) approach. The SDLC consists of a series of stages carried out by experts and users of information and communication systems to develop and implement applications (Resna, 2018). The complete stages of the SDLC approach are illustrated in Figure 1.

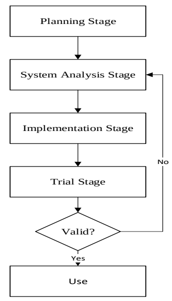

Figure 1: Systems Development Life Cycle (SDLC).

2.2.1. Planning Stage

The planning stage involves collecting data from Twitter user posts related to the 2024 regional elections. The objective is to determine whether these posts express positive, negative, or neutral sentiments. The planning stage involves collecting data from Twitter user posts related to the 2024 regional elections. The objective is to determine whether these posts express positive, negative, or neutral sentiments.

2.2.2. System Analysis Stage

The purpose of this stage is to process the collected information through data preprocessing, followed by the application of the TF-IDF weighting method. The classified data is then analyzed using the Cosine Similarity method.

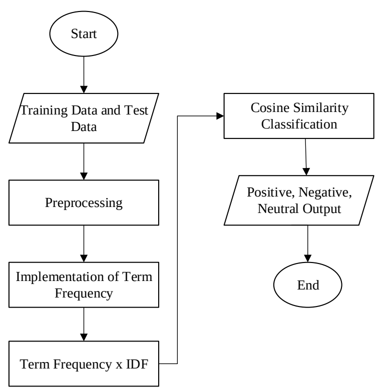

Figure 2: System Analysis Stage.

2.2.3. Design Stage

The design stage builds upon the analysis stage and focuses on developing the system's structure. This stage includes designing the user interface and the database architecture used in the system.

2.2.4. Implementation Stage

The implementation stage begins after the system design is completed. In this stage, the system is developed using the PHP programming language, with MySQL as the database for storing data.

2.2.5. Trial Phase

The system trial phase is conducted to assess whether the system functions properly. This phase includes three types of testing:

1. Structural Trial – Evaluates the alignment between system design and implementation.
2. Functional Trial – Ensures that all forms, buttons, and other interface elements perform their intended functions.
3. Validation Trial – Tests all data processing functions to verify their accuracy by comparing the system's output with manual and alternative processes.

3. Result and Discussion

3.1 Planning Stage

In the planning stage, the initial process involves identifying the software's requirements and limitations. This stage also includes gathering data on positive, negative, and neutral opinions from Twitter users regarding the 2024 regional elections. The objective is to ensure the effective application of the Cosine Similarity method to the collected data.

3.2 System Analysis Stage

3.2.1 Preprocessing

Before classifying the data collected from Twitter, preprocessing must be performed. During this stage, the data undergoes several steps to prepare it for analysis by the system. The preprocessing steps include case folding, tokenization, and filtering.

3.2.2. Case Folding

The Case folding stage is a process for adjusting the shape of the alphabet, and only alphabets from "a" to "z" are accepted for this process. Any characters other than these characters will be removed and treated as delimiters (Rusland, 2017). An example of the application of the case folding stage is shown in Figure 3.

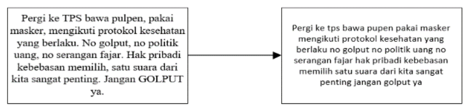

Figure 3: Implementation of Case Folding stage

3.2.3. Tokenizing

The tokenization stage involves breaking down a sequence of text into individual terms or words. The purpose of tokenization is to separate a collection of words in a paragraph, sentence, or page into distinct units (Yapinus, 2019). An example of tokenization is illustrated in Figure 4.

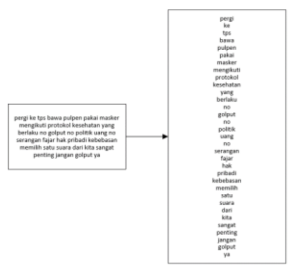

Figure 4: Implementation of Tokenizing stage

3.2.4. Filtering

Following the tokenization process, filtering is performed to remove less important words. The filtering stage utilizes the stoplist algorithm, which eliminates insignificant words while retaining essential terms. Stoplists, or stopwords, refer to non-descriptive words that can be excluded from the vocabulary. Examples of stopwords include *"dan"*, *"yang"*, and *"di"* (Rusland, 2017). An example of the filtering process is illustrated in Figure 5.

Figure 5: Implementation of Filtering

Table 1: Data Before Preprocessing

Code	Opinion	Class
	Training Data	
D1	Our votes are needed only during elections, after that we will remain silenced, and we never depend on you, and we do not need you at all~ Cheers	Negative
D2	Go to the polling station with a pen. And golf, and money politics, and dawn attacks. The personal right of freedom of	Positive

	choice, one voice from us is very important. Don't GOLPUT.	
D3	Begging for the voice of the community, it's your turn to forget about the community, DILAPIDATION!	Negative
D4	The 2024 presidential election will continue to be a success for the betterment of the nation.	Neutral
D5	Let's Succeed in the 2024 Simultaneous Regional Elections that are safe, peaceful, and cool.	Positive
	Testing Data	
D6	In the midst of a disaster, the National Police remains firm to secure a safe, peaceful, and conducive simultaneous regional elections on Wednesday, November 27, 2024	?
D7	Presidential Elections, Legislative Elections, Regional Elections, ask for the support of the people, appeal there to SNI so that there are many supporters votes. When in office & there are policies that the people do not agree with and even criticize harshly, considered enemies. dialogue is incapable of playing the blind midnight of the true traitor, you yes you, CORRUPT!	?

Table 2: Data After Preprocessing

Code	Opinion	Class
	Training Data	
D1	Need to be silenced, no need to toast	Negative
D2	Freedom of choice is important	Positive
D3	Smiling and Forgetting.	Negative
D4	Successful progress	Neutral
D5	Success, Peace, Peace, Cold	Positive
	Testing Data	
D6	Steadfast, safe, peaceful, conducive	?
D7	No, Criticizing, enemies, blind, traitors, depraved	?

3.2.5 TF-IDF Weighting

The Term Frequency-Inverse Document Frequency (TF-IDF) stage is used to determine the relevance of a term (word) within a dataset by assigning a weight to each word. In the TF-IDF process, two key concepts are combined: the frequency of a word's occurrence in a dataset and the inverse frequency of documents containing that word (Sihombing, 2024). To perform TF-IDF weighting, the TF score for each word is first calculated, where the initial weight of each word is set to one. The IDF score is then formulated as follows (Melita et al., 2018):

$$IDF(Word) = log\frac{td}{df} \quad (1)$$

$$W(t,d) = Wtf(t,d) * idf\ t \quad (2)$$

- (t) = Term Frequency-Invers Document Frequency weighting
- (t,) = (W) t f (t, d) weighting
- IDF = Inverse value of df t

After completing the preprocessing stage, the data is weighted using the Term Frequency-Inverse Document Frequency (TF-IDF) method. Term Frequency (TF) represents the frequency of a word's occurrence, which is then multiplied by Inverse Document Frequency (IDF) to determine its importance. Tokens refer to a collection of keywords in the dataset, represented in the chart as words D1 to D7. Each dataset entry containing a listed token is assigned a weight of 1.

Table 3: Implementation of TF

No	Token	dF							DF	D/dF
		D 1	D 2	D 3	D 4	D 5	D 6	D 7		
1.	Needed	1	0	0	0	0	0	0	1	7
2.	Silenced	1	0	0	0	0	0	0	1	7
3.	No	1	0	0	0	0	0	1	2	3.5
4.	Need	1	0	0	0	0	0	0	1	7
5.	Toast	1	0	0	0	0	0	0	1	7
6.	freedom	0	1	0	0	0	0	0	1	7
7.	choose	0	1	0	0	0	0	0	1	7
8.	important	0	1	0	0	0	0	0	1	7
9.	Squirt	0	0	1	0	0	0	0	1	7
10.	forgot	0	0	1	0	0	0	0	1	7
11.	dilapidated	0	0	1	0	0	0	1	2	3.5
12.	successful	0	0	0	1	1	0	0	2	3.5
13.	Progress	0	0	0	1	0	0	0	1	7
14.	safe	0	0	0	0	1	1	0	2	3.5
15.	peace	0	0	0	0	1	1	0	2	3.5
16.	Cold	0	0	0	0	1	0	0	1	7
17.	firm	0	0	0	0	0	1	0	1	7
18.	conducive	0	0	0	0	0	1	0	1	7
19.	Traitor	0	0	0	0	0	0	1	1	7
20.	Criticize	0	0	0	0	0	0	1	1	7
21.	Enemy	0	0	0	0	0	0	1	1	7
22.	blind	0	0	0	0	0	0	1	1	7

Table 4: The TF stage is multiplied by the IDF.

IDF.log D/dF	TF-IDF						
	D 1	D 2	D 3	D 4	D 5	D 6	D 7
0.847	0.847	0	0	0	0	0	0
0.847	0.847	0	0	0	0	0	0
0.243	0.243	0	0	0	0	0	0.243
0.847	0.847	0	0	0	0	0	0
0.847	0.847	0	0	0	0	0	0
0.847	0	0.847	0	0	0	0	0
0.847	0	0.847	0	0	0	0	0
0.847	0	0.847	0	0	0	0	0
0.847	0	0	0.847	0	0	0	0
0.847	0	0	0.847	0	0	0	0
0.847	0	0	0.847	0	0	0	0.847
0.243	0	0	0	0.243	0.243	0	0
0.847	0	0	0	0.847	0	0	0
0.243	0	0	0	0	0.243	0.243	0
0.243	0	0	0	0	0.243	0.243	0
0.847	0	0	0	0	0.847	0	0
0.847	0	0	0	0	0	0,847	0
0.847	0	0	0	0	0	0.847	0
0.847	0	0	0	0	0	0.847	0
0.847	0	0	0	0	0	0	0.847
0.847	0	0	0	0	0	0	0.847
0.847	0	0	0	0	0	0	0.847
Total	3.631	2.541	2.541	1.09	1.576	3.027	3.631

3.2.6. Cosine Similarity

The similarity between documents is calculated using a similarity measure function. To represent the numerical values of documents and determine their closeness, TF-IDF weighting is applied. A higher similarity value indicates greater similarity between two documents, while a lower value suggests less similarity. This measurement enables documents to be evaluated based on their relevance to a given query. The effectiveness of the similarity function directly impacts the quality of the retrieved results (Bening et al., 2018).

$$Similarity(x,y) = \frac{x.y}{||x||||y||} = \frac{\sum_{i=1}^{n} xi.yi}{\sqrt{\sum_{i=1}^{n}(xi)^2}.\sqrt{\sum_{i=1}^{n}(yi)^2}} \quad (3)$$

The cosine similarity classification stage is conducted after the TF-IDF weighting stage. In this phase, the cosine similarity method is applied to measure the similarity between documents. Cosine similarity assigns a numerical value to each document, enabling the calculation of their similarity. This stage involves computing the vector length and the similarity of data D6 and D7, along with all classified data (D1, D2, D3, D4, and D5).

Table 5: Positive Cosine Similarity Classification Stage.

Weight D 6*weight di					Vector Length					
D 1*D 6	D 2*D 6	D 3*D 6	D 4*D 6	D 5*D 6	D 1	D 2	D 3	D 4	D 5	D 6
0	0	0	0	0	0.717	0	0	0	0	0
0	0	0	0	0	0.717	0	0	0	0	0
0	0	0	0	0	0.059	0	0	0	0	0
0	0	0	0	0	0.717	0	0	0	0	0
0	0	0	0	0	0.717	0	0	0	0	0
0	0	0	0	0	0	0.013	0	0.013	0.013	0
0	0	0	0	0	0	0.013	0	0	0	0
0	0	0	0	0	0	0.717	0	0	0	0
0	0	0	0	0	0	0.717	0	0	0	0
0	0	0	0	0	0	0.717	0	0	0	0
0	0	0	0	0	0	0	0.717	0	0	0
0	0	0	0	0	0	0	0.717	0	0	0
0	0	0	0	0	0	0	0.059	0	0	0
0	0	0	0	0	0	0	0	0.059	0.059	0
0	0	0	0	0	0	0	0	0.717	0	0
0	0	0	0	0.059	0	0	0	0	0.059	0.059
0	0	0	0	0.059	0	0	0	0	0.059	0.059
0	0	0	0	0	0	0	0	0	0.059	0
0	0	0	0	0	0	0	0	0	0	0.717
0	0	0	0	0	0	0	0	0	0	0.717
0	0	0	0	0.118	2.927	2.177	1.493	0.789	0.249	1.552
					1.711	1.475	1.222	0.888	0.499	1.246

After multiplying the weights, the vector length is calculated by finding the square root of the sum of the squared values for each data point in the vector length table. Next, the similarity calculation is performed by comparing data 1, 2, 3, 4, and 5 with data 6 (positive words), as shown below. The same process is then repeated to calculate the similarity between Document 6 and Documents 1, 2, 3, 4, and 5.

Cos (D 6, D 1) = 0/ (1,246 * 1.711) = 0

Cos (D 6, D 2) = 0 / (1.246 *1.475) = 0

Cos (D 6, D 3) = 0 / (1.246 * 1.222) = 0

Cos (D 6, D 4) = 0 / (1.246 * 0.888) = 0

Cos (D 6, D 5) = 0.118 / (1.246 * 0.499) = 0.19

The results of the Cosine Similarity calculation are as follows:

Table 6: The results of the Cosine Similarity calculation.

D 1	D 2	D 3	D 4	D 5
0	0	0	0	0.300

Then, sort the results of the Cosine Similarity calculation.

Table 7: Similarity Level Rating.

1	2	3	4	5
D5	D2	D4	D3	D1

In the classification results using the Cosine Similarity method for positive sentiment, among all the data compared with D6, D5 has the highest similarity, as its cosine similarity value is closest to 1. A cosine similarity value closer to 1 indicates a smaller angle between vectors, signifying a higher degree of similarity between the data points (Han et al., 2012).

Table 8: Negative Cosine Similarity Classification Stage.

Weight D 7*weight di					Vector Length					
D 1*D 7	D 2*D 7	D 3*D 7	D 4*D 7	D 5*D 7	D 1	D 2	D 3	D 4	D 5	D 7
0	0	0	0	0	0.717	0	0	0	0	0
0	0	0	0	0	0.717	0	0	0	0	0
0	0	0	0	0	0.059	0	0	0	0	0.059
0	0	0	0	0	0.717	0	0	0	0	0
0	0	0	0	0	0.717	0	0	0	0	0
0	0	0	0	0	0	0.717	0	0	0	0
0	0	0	0	0	0	0.717	0	0	0	0
0	0	0	0	0	0	0.717	0	0	0	0
0	0	0	0	0	0	0	0.717	0	0	0
0	0	0	0	0	0	0	0.717	0	0	0
0	0	0.00348	0	0	0	0	0.059	0	0	0.059
0	0	0	0	0	0	0	0	0.059	0.059	0
0	0	0	0	0	0	0	0	0.605	0	0
0	0	0	0	0	0	0	0	0	0.059	0
0	0	0	0	0	0	0	0	0	0.059	0
0	0	0	0	0	0	0	0	0	0.717	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0.717
0	0	0	0	0	0	0	0	0	0	0.717
0	0	0	0	0	0	0	0	0	0	0.717
0	0	0	0	0	0	0	0	0	0	0.717
0	0	0.00348	0	0	2.927	2.151	1.493	0.664	0.894	2.986
					1.711	1.467	1.222	0.815	0.956	1.728

After performing the weight multiplication, the vector length is calculated by finding the square root of the sum of the squared values for each data point in the vector length table. Next, the similarity calculation is performed by comparing data 1, 2, 3, 4, and 5 with data 7 (negative words), as shown below. The same process is then repeated to calculate the similarity between Document 7 and Documents 1, 2, 3, 4, and 5.

Cos (D7, D1) = 0/(1.728*1.711) = 0

Cos (D7, D2) = 0/(1.728*1.467) = 0

Cos (D7, D3) = 0.00348/(1.728*1.222) = 0.00164

Cos (D7, D4) = 0/(1.728*0.815) = 0

Cos (D7, D5) = 0/(1.728*0.956) = 0

Table 9: Negative Cosine Similarity Result.

D 1	D 2	D 3	D 4	D 5
0	0	0.162	0	0

Then, sort the results of the Cosine Similarity calculation.

Table 10: Similarity Level Rating

1	2	3	4	5
D 3	D 1	D 2	D 4	D 5

In the classification results using the Cosine Similarity method for negative sentiment, among all the data compared with D7, D3 has the highest similarity, as its cosine similarity value is closest to 1. A cosine similarity value closer to 1 indicates a smaller angle between vectors, signifying a higher degree of similarity between the data points (Han et al., 2012).

3.3 Implementation Stage

3.3.1. Homepage

The home page displays the main information generated by the system on this website. An illustration of the home page is shown in Figure 6.

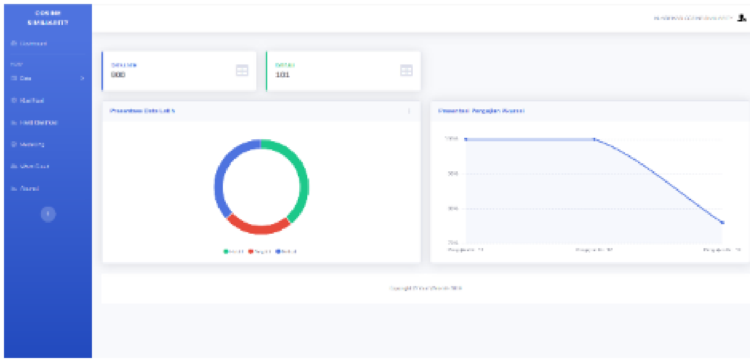

Figure 6: Homepage

3.3.2. Training Data Page

The training data page shows the data taken from the scrapping results that will be used to apply the classification using the Cosine Similarity method. The training data page is shown in Figure 7.

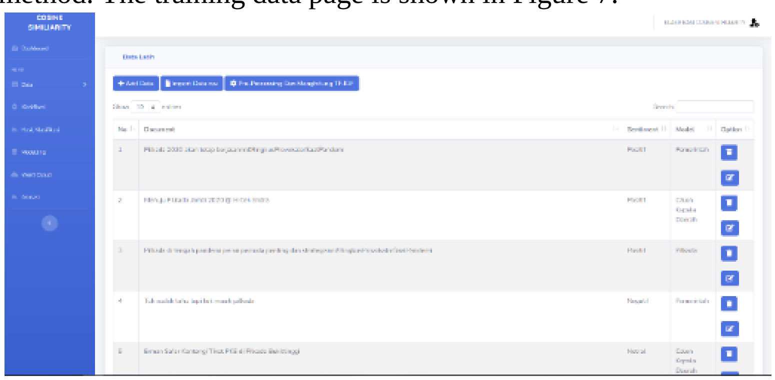

Figure 7: Training Data Page

3.3.3. Test Data Page

The test data page displays the retrieved data used for classification with the Cosine Similarity method. An illustration of the test data page is shown in Figure 8.

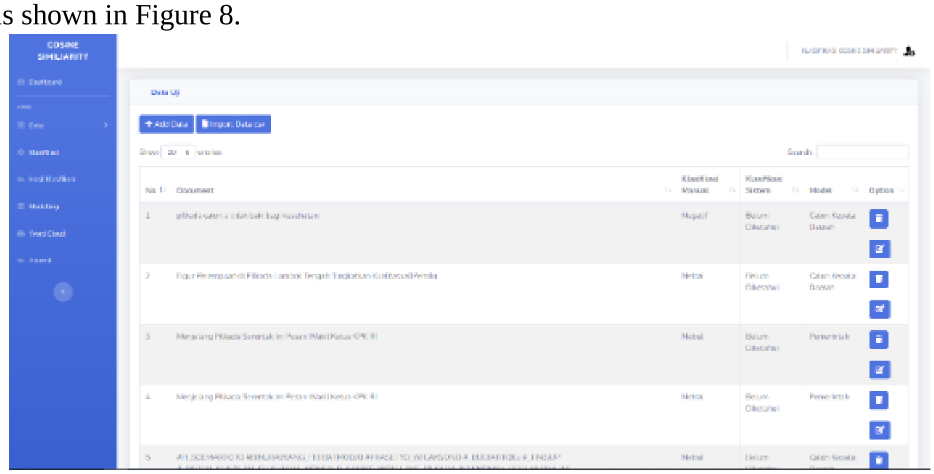

Figure 8: Test Data Page

3.3.4. Filtering Result Page

The filtering results page displays the filtered data used in classification calculations with TF-IDF weighting. An illustration of the filtering results page is shown in Figure 9.

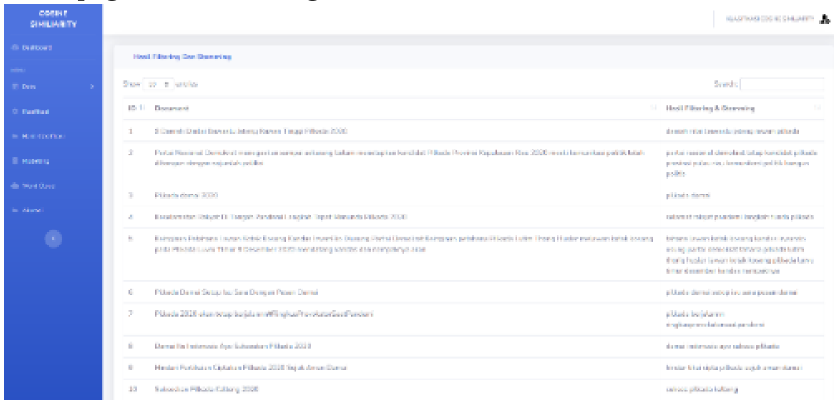

Figure 9: Filtering Result Page.

3.4 Trial Phase

3.4.1. Accuracy Results Page

The accuracy results page displays the accuracy values of the manual classification performed by the system. An overview of the accuracy results page is shown in Figure 10.

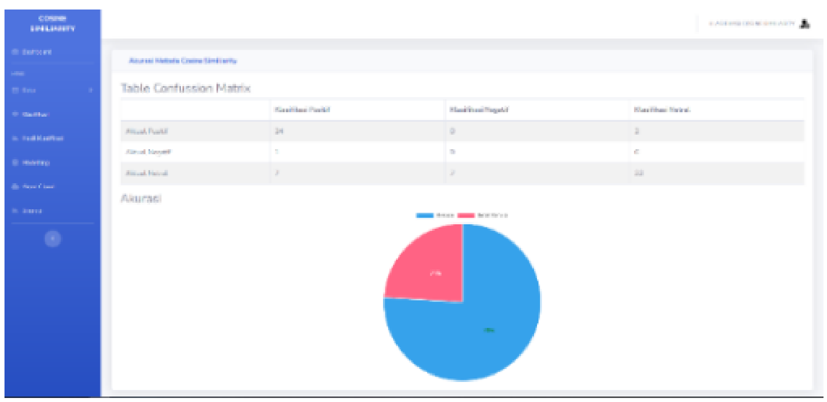

Figure 10: Accuracy Calculation Results Page

3.4.2 Sentiment Analysis Result Page

The sentiment analysis results of Twitter data from tweets about the 2024 simultaneous regional elections produce three variable values: positive, negative, and neutral. The percentage results are shown in Figure 11.

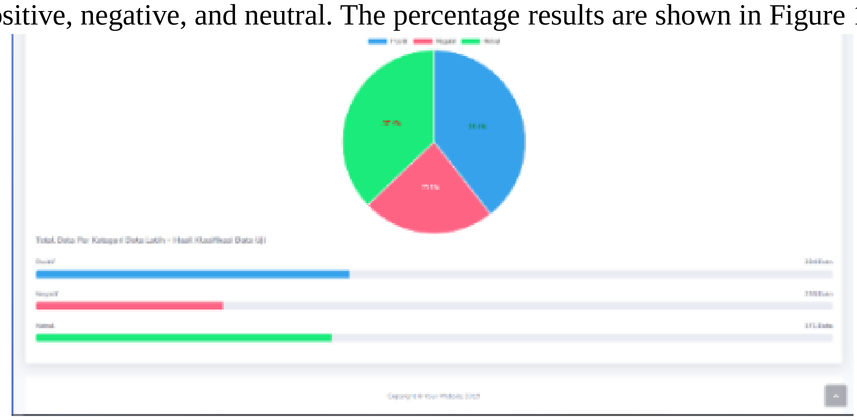

Figure 11: Sentiment Analysis Percentage Results

The sentiment analysis results show that 39.4% of tweets were classified as positive, 23.5% as negative, and 37.1% as neutral. A sentiment is classified as neutral if it cannot be distinctly categorized as positive or negative. Additionally, the system presents the sentiments expressed by the public on Twitter regarding the 2024 regional elections. An illustration of the public sentiment results is shown in Figure 12.

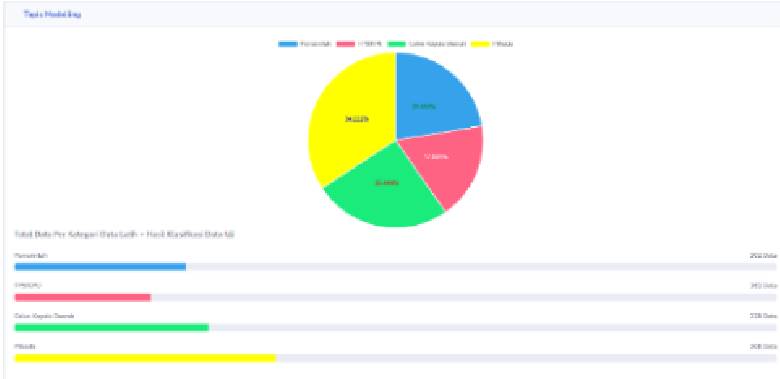

Figure 12: Public Sentiment Results.

It can be seen from the results of the sentiments that people expressed on Twitter, which were directed at the Government at 22.444%, TPS/KPU 17.889%, Regional Head Candidates 25.444%, and Regional Elections 34.222%.

3.4.3. Accuracy Results Page

The accuracy results are derived from a comparison between manual classification and system-generated classification. These results are used to evaluate the correctness of the manual classification. The system's classification results are presented in the confusion matrix table, while the overall accuracy results are shown in Table 11.

Table 11: Confusion Matrix.

		Positive Prediction	Negative Prediction	Neutral Prediction
Positive Value	Actual	34	0	3
Negative Value	Actual	1	9	6
Neutral Value	Actual	7	7	33

Out of 37 positive sentiment data points, 34 were correctly classified as positive, while 3 were classified as neutral. Among the 16 negative sentiment data points, 9 were correctly classified as negative, 1 was misclassified as positive, and 6 were classified as neutral. For the 47 neutral sentiment data points, 33 were correctly classified as neutral, while 7 were misclassified as positive and 7 as negative. The accuracy results based on the confusion matrix are shown in Figure 13.

Figure 13: Accuracy Result

It can be seen from the accuracy results page above that the manual classification achieved an accuracy of 76%, while 24% of the classifications were incorrect. In this study, positive, negative, and neutral sentiments were successfully classified by applying the TF-IDF process and utilizing the Cosine Similarity classification algorithm. In manual calculations, a higher similarity score indicates a greater resemblance between the two evaluated objects, and vice versa

(Bening, 2018). This study effectively classified sentiment analysis results from people's tweets on Twitter into positive, negative, and neutral categories based on similarity.

4. Conclusion

From the results of the study on public opinion classification regarding the 2024 regional elections using Cosine Similarity and TF-IDF, it can be concluded that the Cosine Similarity method is effective in automatically classifying sentiments expressed by the public on Twitter using the developed system. Sentiment analysis was conducted by collecting data (mining), preprocessing the data to extract relevant words, calculating word weights using TF-IDF, and applying Cosine Similarity to rank the collected data. A total of approximately 1,000 data points were used in this study, consisting of 900 training data and 100 test data. The sentiment classification results indicate that 39.4% of the sentiments expressed were positive, 23.5% were negative, and 37.1% were neutral. These findings suggest strong public support for conducting the regional elections safely.

References

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques.* Massachusetts: Elsevier Inc.

Melita, R., Amrizal, V., Suseno, H., & Dirjam, T. (2018). Application of Term Frequency Inverse Document Frequency (TF-IDF) and Cosine Similarity Methods in Information Retrieval Systems to Know Web-Based Hadith Syarah (Case Study: Syarah Umdatil Ahkam). *J. Tek. Inform*, *11*(2), 149-164.

Nurmandi, A., Purnomo, E. P., Prianto, A. L., & Jaenuri, M. (2015). The current practice of citizen participation in Indonesia netizen and selected cases of local democracy in Indonesia. Citizen Participation, 21, 119-167.

Pratama, Y., Roberto Tampubolon, A., Diantri Sianturi, L., Diana Manalu, R., & Frietz Pangaribuan, D. (2019, March). Implementation of sentiment analysis on Twitter using Naïve Bayes algorithm to know the people responses to debate of DKI Jakarta governor election. In *Journal of Physics: Conference Series* (Vol. 1175, p. 012102). IOP Publishing.

Putra, B. A., & Afdal, M. (2023, December). Sentiment Analysis of Presidential Candidates of the Republic of Indonesia Using Naïve Bayes Classifier and Support Vector Machine. In *2023 7th International Conference on New Media Studies (CONMEDIA)* (pp. 263-268). IEEE.

Resna, R. W., Lazuardi, L., Werdati, S., & Rochmah, W. (2019). Development of Detection Instrument Models for Mobility Impairment in The Older Adults Based on A Mobile Health Nursing Application in A Public Health Center. *Jurnal Ners*, *14*(3), 86.

Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017, August). Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets. In *IOP conference series: materials science and engineering* (Vol. 226, No. 1, p. 012091). IOP Publishing.

Sihombing, J. J., Arnita, A., Al Idrus, S. I., & Niska, D. Y. (2024). Implementation of text summarization on indonesian scientific articles using textrank algorithm with TF-IDF web-based. *Journal of Soft Computing Exploration*, *5*(3), 310-319.

Wu, W., & Deng, Z. (2022). The analysis of public opinion in colleges and universities oriented to wireless networks under the application of intelligent data mining. Wireless Communications and Mobile Computing, 2022(1), 7597366.

Yapinus, G., Erwin, A., Galinium, M., & Muliady, W. (2014, October). Automatic multi-document summarization for Indonesian documents using hybrid abstractive-extractive summarization technique. In *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 1-5). IEEE.