Sentiment Analysis of 2024 Presidential Candidates Election Using SVM Algorithm

**Michael Alfonso1\*), Dionisia Bhisetya Rarasati, S.Kom., M.T.I2**

1Program Studi Teknik Informatika, Fakultas Teknologi dan Desain, Universitas Bunda Mulia

2 Program Studi Teknik Informatika, Fakultas Teknologi dan Desain, Universitas Bunda Mulia

Email: 1[s32190039@student.ubm.ac.id](mailto:s32190039@student.ubm.ac.id), 2[l1783@lecturer.ubm.ac.id](mailto:l1783@lecturer.ubm.ac.id)

***Abstract*** Elections for presidential candidates are held every 5 years with various candidates, especially on Twitter, arguments about political matters often occur that many Twitter users participate in discussions about the election for presidential candidate. Therefore, this study focuses on sentiment analysis to infer user responses to the presidential election and validate it by looking for a correlation between electability survey results and Twitter sentiment data using Pearson Correlation. In sentiment analysis model, the 10-Fold Cross Validation method is used to find the best model from a dataset with a division of training data and test data with 90:10 split. Then the alphabetic data will be converted into numeric data using the TF-IDF weighting method. To validate the best model, Confusion Matrix is used to get the best f1-score. The model is using Support vector machine algorithm with the Gaussian RBF (Radial Basis Function) kernel. The results of the analysis are compared with the results of the news portal electability survey which contains the 3 candidates using Pearson Correlation. This study produces the best fold for each data on each presidential candidate with the f1-score to find the best model for each fold. In the Peason Correlation result, the higher positive sentiment of each presidential candidate, the higher electability survey data. For further research, research can be discuss about hyper tuning parameters and using other kernels on Support vector machine algorithm.

***Keywords – NLP, Pearson Correlation, Sentiment analysis, SVM, TF-IDF***

# **INTRODUCTION**

Elections occur every 5 years with the number of voters increasing each period, with an increase of 28.90% from 2004 to 2019 [1]. Social media such as Facebook, Twitter, LinkedIn, YouTube have become part of human life, it is proven that most interactions can be done with social media, until it becomes self-identity to the general public [2]. On social media, especially Twitter, there is a lot of dissemination of information, entertainment, and political opinion. This is not spared from the election which is a political activity that will be held in 2024, referred to as a democratic party, this is certainly a lot of discussion on social media. One way to convey opinions and through social media is to use certain hashtags in tweets posted on Twitter. Twitter itself is used because of the ease in replying to tweets with likes, comments and even replies, so that topics that are currently being discussed can be easily spread [3]. Therefore, further analysis is needed to see how much Twitter users agree and disagree with the 2024 presidential election, then data will be obtained from Twitter user tweets related to this topic, namely the 2024 general election. Sentiment analysis or Opinion Mining is a study that learn about opinions, sentiments, evaluations, behavior and emotions towards an entity [2], [3]. Sentiment analysis is used to find out which tweets are posted containing positive, negative or neutral words [4]. In this study, the Support vector machine (SVM) method will be used, which is one of the classification methods in machine learning (supervised learning) that predicts classes based on patterns from training results. Classification is done by creating a hyperplane between positive and negative classes [2]. The results of this study are the sentiments of Twitter users towards the presidential election in 2024.

# **Research Methodology**

This research consists of 6 processes in outline, namely data collection, preprocessing, Term Frequency-Inverse Document Frequency weighting, Support Vector Machine training using 10-Fold Cross Validation, Choosing the best fold to be used as a model, and testing as shown below.

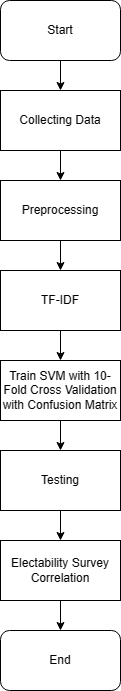


Figure 1. Research Flow

* 1. *Collection Data*

The first stage of this research is to scrape data sourced from Twitter with the keyword "(Name of Candidate) President" to narrow down the context of a Twitter user's Tweet. The data collection process is carried out with the help of the snscrape library by python. Data was collected from 1 January 2023 to 31 March 2023. This scraping process resulted in a total of 18,267 data with 7,810 data that were duplicate and irrelevant to the theme "2024 Presidential Candidate Election" resulting in a total of 10,457 clean data from 3 presidential candidates.

* 1. *Preprocessing*

The pre-processing stage functions to clean and transform data so that it can be processed by the system easily and accurately. Several stages of pre-processing in this study include:

1. *Lowercasing*

Lowercasing is a step to change all letters to lowercase so that the word being processed becomes the same form. Table 1 shows an example of the lowercasing process from some of Anies Baswedan's tweets:

Table 1. Lowercasing Result

| No | Kalimat Awal | | Kalimat Hasil |
| --- | --- | --- | --- |
| 1 | PKS Mendukung Anies Baswedan Menjadi Calon Presiden 2024 | pks mendukung anies baswedan menjadi calon presiden 2024 | |
| 2 | Anies Baswedan Presiden Indonesia | anies baswedan presiden indonesia | |
| 3 | Walau sekarang didukung Demokrat dan PKS, ada empat alasan Anies Baswedan gak bakal bisa menang jadi presiden !! | walau sekarang didukung demokrat dan pks, ada empat alasan anies baswedan gak bakal bisa menang jadi presiden !! | |
| 4 | Anies Baswedan lh yg pantas jadi presiden | anies baswedan lh yg pantas jadi presiden | |
| 5 | Pak Anies Baswedan calon presiden 2024 | pak anies baswedan calon presiden 2024 | |

1. *Remove Special Character and Number*

This stage functions to produce sentences without punctuation, emoticons, and numbers. Table 2 shows an example of the Remove Special Character and Number process:

Table 2. Remove Special Character and Number Result

| No | Kalimat Awal | | Kalimat Hasil |
| --- | --- | --- | --- |
| 1 | pks mendukung anies baswedan menjadi calon presiden 2024 | pks mendukung anies baswedan menjadi calon presiden | |
| 2 | anies baswedan presiden indonesia | anies baswedan presiden indonesia | |
| 3 | walau sekarang didukung demokrat dan pks, ada empat alasan anies baswedan gak bakal bisa menang jadi presiden !! | walau sekarang didukung demokrat dan pks ada empat alasan anies baswedan gak bakal bisa menang jadi presiden | |
| 4 | anies baswedan lh yg pantas jadi presiden | anies baswedan lh yg pantas jadi presiden | |
| 5 | pak anies baswedan calon presiden 2024 | pak anies baswedan calon presiden | |

1. *Stemming*

At this stage, all affixed words will be processed to become basic words, as shown in table 3 below:

Table 3. Stemming Result

| No | Kalimat Awal | | Kalimat Hasil |
| --- | --- | --- | --- |
| 1 | pks mendukung anies baswedan menjadi calon presiden | pks dukung anies baswedan jadi calon presiden | |
| 2 | anies baswedan presiden indonesia | anies baswedan presiden indonesia | |
| 3 | walau sekarang didukung demokrat dan pks ada empat alasan anies baswedan gak bakal bisa menang jadi presiden | walau sekarang dukung demokrat dan pks ada empat alasan anies baswedan gak bakal bisa menang jadi presiden | |
| 4 | anies baswedan lh yg pantas jadi presiden | anies baswedan lh yg pantas jadi presiden | |
| 5 | pak anies baswedan calon presiden | pak anies baswedan calon presiden | |

1. *Tokenize*

This stage break sentences into tokens per word. This process is carried out with the aim that the next process can be applied at the word level. The process is shown in table 4 below:

Table 4. Tokenize Result

| No | Kalimat Awal | | Kalimat Hasil |
| --- | --- | --- | --- |
| 1 | pks dukung anies baswedan jadi calon presiden | [[pks] [dukung] [anies] [baswedan] [jadi] [calon] [presiden]] | |
| 2 | anies baswedan presiden indonesia | [[anies] [baswedan] [presiden] [Indonesia]] | |
| 3 | walau sekarang dukung demokrat dan pks ada empat alasan anies baswedan gak bakal bisa menang jadi presiden | [[walau] [sekarang] [dukung] [demokrat] [dan] [pks] [ada] [empat] [alasan] [anies] [baswedan] [gak] [bakal] [bisa] [menang] [jadi] [presiden]] | |
| 4 | anies baswedan lh yg pantas jadi presiden | [[anies] [baswedan] [lh] [yg] [pantas] [jadi] [presiden]] | |
| 5 | pak anies baswedan calon presiden | [[pak] [anies] [baswedan] [calon] [presiden]] | |

1. *Remove Stopwords*

Stopwords are for removing words that don't have an impact on sentiment, but often appear in sentences. Table 5 shows the Remove Stopwords process as follows:

Table 5. Remove Stopwords Result

| No | Kalimat Awal | | Kalimat Hasil |
| --- | --- | --- | --- |
| 1 | [[pks] [dukung] [anies] [baswedan] [jadi] [calon] [presiden]] | [[pks] [dukung] [anies] [baswedan] [calon] [presiden]] | |
| 2 | [[anies] [baswedan] [presiden] [Indonesia]] | [[anies] [baswedan] [presiden] [Indonesia]] | |
| 3 | [[walau] [sekarang] [dukung] [demokrat] [dan] [pks] [ada] [empat] [alasan] [anies] [baswedan] [gak] [bakal] [bisa] [menang] [jadi] [presiden]] | [[sekarang] [dukung] [demokrat] [pks] [empat] [alasan] [anies] [baswedan] [gak] [bakal] [menang] [presiden]] | |
| 4 | [[anies] [baswedan] [lh] [yg] [pantas] [jadi] [presiden]] | [[anies] [baswedan] [lh] [pantas] [presiden]] | |
| 5 | [[pak] [anies] [baswedan] [calon] [presiden]] | [[pak] [anies] [baswedan] [calon] [presiden]] | |

1. *Slangwords Filtering*

This stage removes slang words into standard words. The slang dictionary is taken from the github repository posted by [louisowen6](https://github.com/louisowen6) which is shown in Table 6 as follows:

Table 6. Slangwords Result

| No | Kalimat Awal | | Kalimat Hasil |
| --- | --- | --- | --- |
| 1 | [[pks] [dukung] [anies] [baswedan] [calon] [presiden]] | [[pks] [dukung] [anies] [baswedan] [calon] [presiden]] | |
| 2 | [[anies] [baswedan] [presiden] [Indonesia]] | [[anies] [baswedan] [presiden] [Indonesia]] | |
| 3 | [[sekarang] [dukung] [demokrat] [pks] [empat] [alasan] [anies] [baswedan] [gak] [bakal] [menang] [presiden]] | [[sekarang] [dukung] [demokrat] [pks] [empat] [alasan] [anies] [baswedan] [tidak] [akan] [menang] [presiden]] | |
| 4 | [[anies] [baswedan] [lh] [pantas] [presiden]] | [[anies] [baswedan] [pantas] [presiden]] | |
| 5 | [[pak] [anies] [baswedan] [calon] [presiden]] | [[pak] [anies] [baswedan] [calon] [presiden]] | |

1. *Remove Synonim*

At this stage, the writer removes synonyms in the same sentence if any. Table 7 shows the Remove Synonim process as follows:

Table 7. Remove Synonim Result

| No | Kalimat Awal | | Kalimat Hasil |
| --- | --- | --- | --- |
| 1 | [[pks] [dukung] [anies] [baswedan] [calon] [presiden]] | [[pks] [dukung] [anies] [baswedan] [calon] [presiden]] | |
| 2 | [[anies] [baswedan] [presiden] [Indonesia]] | [[anies] [baswedan] [presiden] [Indonesia]] | |
| 3 | [[sekarang] [dukung] [demokrat] [pks] [empat] [alasan] [anies] [baswedan] [tidak] [akan] [menang] [presiden]] | [[sekarang] [dukung] [demokrat] [pks] [empat] [alasan] [anies] [baswedan] [tidak] [akan] [menang] [presiden]] | |
| 4 | [[alhamdulillah] [mudah] [mudah] [anies] [baswedan] [takdir] [allah] [swt] [presiden] [aamiin] [yra] | [[alhamdulillah] [mudah] [anies] [baswedan] [takdir] [allah] [swt] [presiden] [aamiin] [yra] | |
| 5 | [[pak] [anies] [baswedan] [calon] [presiden]] | [[pak] [anies] [baswedan] [calon] [presiden]] | |

* 1. *TF-IDF*

TF-IDF weighting is a method of assessing the weight of each word in a document, TF-IDF is a strong feature weighting method that measures the frequency and distribution of words to determine their weight. TF-IDF consists of TF (Term Frequency), which is the frequency of occurrence of words in documents and IDF (Inverse Document Frequency), which is the opposite of the frequency of documents, the TF-IDF value can be obtained by using the following equation:

1. The value of Term Frequency (TF) is obtained by the value of the frequency of appearance of the t feature in document d[5]

(1)

1. The Inverse Document Frequency (IDF) value is obtained from the logarithm of the number of documents n divided by df documents containing t features[6]

(2)

1. The value of the Term Frequency Inverse Document Frequency (TF-IDF) is obtained by multiplying the TF and IDF values[7]

(3)

TF-IDF functions to convert data into numerical form so that the features that have been processed can be loaded into the equation. So this method was chosen to extract the Twitter data.

* 1. *Support Vector Machine*

Support Vector Mahine was created by Boser, Guyon, and Vapnik [8] which is a classification method for finding the best hyperplane to separate 2 classes [9]. This algorithm will produce a hyperplane that separates a plane into 2 parts in each class which can be seen in figure 2.

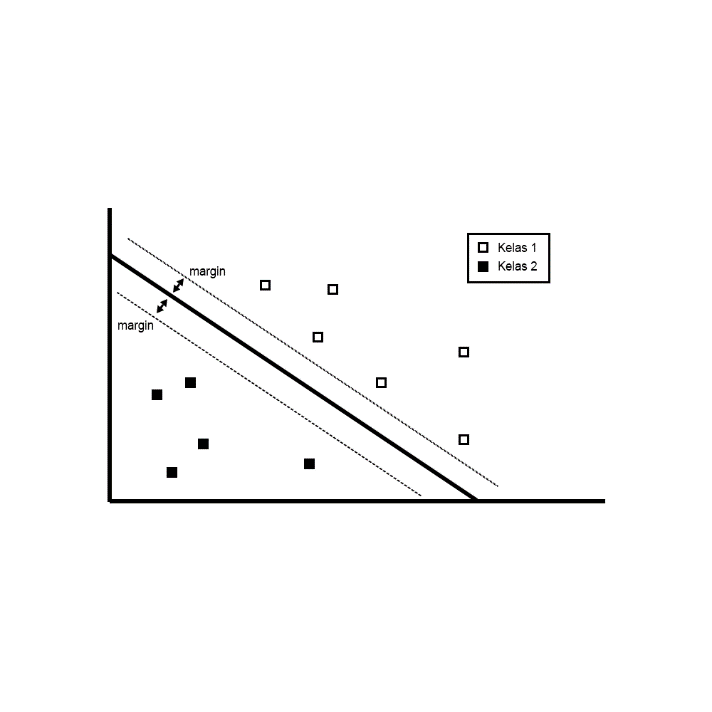


Figure 2. Hyperplane Support Vector Machine Example

Several kernels that can be applied to the SVM algorithm:

1. Linear Kernel is a kernel function that separates 2 classes linearly with the following equation:

(4)

Description of the following equation is xi as the train data, followed by i as ammount of data (ex: 1,2, …, n), then yi as the label of train data.

1. Polynomial Kernel is a kernel function that is not linearly separated, the equation is as follows:

(5)

Description of the following equation is xi, xj as the train data, followed by c and d as constant.

1. Radial Basis Function (RBF) The kernel is a function for analyzing data that is not linearly separated and requires 2 types of parameters (Gamma and Cost).

(6)

Description of the following equation is as euclidean distance and as independent parameter which define substraction level of RBF toward 0.

To use the Support vector machine algorithm, it is necessary to find the best to separate each class. Normally with data that is not spread out, the hyperplane is either linear or polynomial. But if it comes to the author's case, the data distribution is very wide and the feature dimensions are unlimited, then it is necessary to use a kernel trick, the implementation that the writer has done to overcome this problem is to use a Gaussian RBF (Radial Basis Function) which is able to overcome the problem above.

* 1. *Table Description*

Tables are sequentially numbered with the table title and number above the table. Tables should be centered in the column OR on the page. Tables should be followed by a line space (8pt). Elements of a table should be single-spaced, however double spacing can be used to show groupings of data or to separate parts within the table. Table headings should be in 8pt bold. Tables are referred in the text by the table number without location (above, below, following, etc.), e.g.: Table 1. Do not show vertical line in the table.

Table 1. Table Description

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

* 1. *Equations*

Equations should be numbered serially within parentheses as shown in Equation (1). Equation should be prepared using MS Equation Editor (not in image format). The equation number is to be placed at the extreme right side. Explanation of the symbols in the equation should be presented as a descriptif paragraph after the equation.

(1)

# **Results And Discussion** (10 pt,Capital, Bold)

Results should be clear and concise. The results should summarize (scientific) findings rather than providing data in detail. Please highlight differences between your results or findings and the previous publications by other researchers. The discussion should explore the significance of the results of the work, not repeat them. A combined Results and Discussion section is often appropriate. Avoid extensive citations and discussion of published literature.

In discussion, it is the most important section of your article. Here you get the chance to sell your data. Make the discussion corresponding to the results, but do not reiterate the results. Often should begin with a brief summary of the main scientific findings (not experimental results).

The following components should be covered in discussion: (a) How do your results relate to the original question or objectives outlined in the Introduction section? What is your finding of research? (what/how)? (b) Do you provide interpretation scientifically for each of your results or findings presented? This scientific interpretation must be supported by valid analysis and characterization (why)? (c) Are your results consistent with what other investigators have reported (what else)? Or are there any differences?

.

# **Conclusion** (10 pt,Capital, Bold)

Conclusions should only answer the objectives of the research. Tells how your work advances the field from the present state of knowledge. Without clear Conclusions, reviewers and readers will find it difficult to judge the work, and whether or not it merits publication in the journal. Do not repeat the Abstract, or just list experimental results. Provide a clear scientific justification for your work, and indicate possible applications and extensions. This conclusion should be provided as a paragraph. You should also suggest future experiments and/or point out those that are underway.

# **References** (10 pt,Capital, Bold)

Cite the main scientific publications on which your work is based. Cite only items that you have read. Do not inflate the manuscript with too many references. Avoid excessive self‐citations. Avoid excessive citations of publications from the same region. Check each reference against the original source (authors name, volume, issue, year, DOI Number). Please use Reference Manager Applications like EndNote, Mendeley, Zotero, etc. Use other published articles in the same journal as models. **The minimum number of references should be 15 references**.

All publications cited in the text should be included as a list of references. **References are sequentially numbered as they appear in the text**. **Reference numbers are indicated in square brackets**. Please ensure that every reference cited in the text is also present in the reference list (and vice versa). Any references cited in the abstract must be given in full. Unpublished results and personal communications are not recommended in the reference list, but may be mentioned in the text. If these references are included in the reference list they should follow the standard reference style of the journal and should include a substitution of the publication date with either “Unpublished results” or “Personal communication”. Citation of a reference as 'in press' implies that the item has been accepted for publication.

As a minimum, the full URL should be given and the date when the reference was last accessed. Any further information, if known (DOI, author names, dates, a reference to a source publication, etc.), should also be given. Web references can be listed separately (e.g., after the reference list) under a different heading if desired, or can be included in the reference list.

**References list MUST follow American Psychological Association (APA) Version 6.0.**

1. Authors Last Name, Initial(s). (year). *Book Title*. Edition*.* Place: Publisher. ← *Book*
2. Engelmore, R., Morgan, A. (1986). *Blackboard Systems*. Edition. Reading, Mass.: Addison-Wesley. ← *Book*
3. Authors Last Name, Initial(s). (year). Article Title. *Journal Title*. Volume (Issue Number: optional), page number start-page number end. DOI information. ← *Journal*
4. Robinson, A.L. (1980). New Ways to Make Microcircuits Smaller. *Science*, 208, 1019-1026. DOI: 10.1126/science.208.4447.1019. ← *Journal*
5. Bhavsar, D.S., Saraf, K.B. (2002). Morphology of PbI2 Crystals Grown by Gel Method. *Crystal Research and Technology*, 37, 51–55. DOI: 10.1002/1521-4079(200202)37:1%3C51::AID-CRAT51%3E3.0.CO;2-N ←*Journal*
6. Li, M., Xing, S., Yang, L., Fu, J., Lv, P., Wang, Z., Yuan, Z. (2019). Nickel-loaded ZSM-5 catalysed hydrogenation of oleic acid: The game between acid sites and metal centres. *Applied Catalysis A: General.*587, 117112. DOI: 10.1016/j.apcata.2019.117112. ←*Journal*
7. Clancey, W.J. (1983). Communication, Simulation, and Intelligent Agents: Implications of Personal Intelligent Machines for Medical Education. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 556-560. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence, Inc. ←*Conferences*
8. Amin, N.A.S., Istadi, I. (2012). Different Tools on Multiobjective Optimization of a Hybrid Artificial Neural Network – Genetic Algorithm for Plasma Chemical Reactor Modelling. In Olympia Roeva (Editor) *Real-World Applications of Genetic Algorithms*. Croatia: InTech Publisher. ← *Book Chapter*
9. Rice, J. (1986). Polygon: A System for Parallel Problem Solving, *Technical Report*, KSL-86-19, Dept. of Computer Science, Stanford Univ. ←*Report*
10. Clancey, W.J. (1979). Transfer of Rule-Based Expertise through a Tutorial Dialogue. *Ph.D. Dissertation*, Department of Computer Science, Stanford University. ←*Thesis*
11. Ivey, K.C. (2 September 1996). *Citing Internet sources* URL http://www.eei- alex.com/eye/utw/96aug.html. ←*Website (but should be avoided)*