SINOMICS JOURNAL
International Journal of Social Science, Education, Communication and Economics
ISSN (e): 2829-7350 | ISSN(p): 2963-9441

Exploring Sentimen Analysis Using Machine Learning: A Case Study on Partai Demokrasi Indonesia Perjuangan (PDIP) in the 2024 General Election

Fandi Kurniawan[1], Qois Al Qorni[2]
Universitas Muhammadiyah Kota Bumi, Lampung, Indonesia[1]
Universitas Multi Data Palembang, Palembang, Indonesia[2]
E-mail: fandi.kurniawan@umko.ac.id[1], qoisalqorni501@gmail.com[2]

Abstract

General elections have an important role in democratic systems in all countries. In this context, sentiment analysis plays an important role in revealing the public's views on political parties. This study analyzes sentiment towards the Partai Demokrasi Indonesia Perjuangan (PDIP) in the 2024 General Election using the Naive Bayes algorithm. This method classifies social media content related to PDIP into positive (support) or negative (criticism) categories. The analysis results show 100% precision in identifying positive sentiment, but recall is only 97.11%, indicating that some positive sentiment was missed. For negative sentiment, the precision was 94.33% with a recall of 100%, indicating the ability to recognize negative sentiment but little negative prediction error. This study provides an in-depth understanding of the PDIP's perception in the 2024 General Election, supports better political decision making and provides insight to the PDIP in understanding the public's views.

Keywords *General elections, Sentiment analysis, Naive Bayes algorithm, Perception.*

INTRODUCTION

Elections are an important democratic process in the political life of a country. Political parties are one of the key elements in elections, which play a role in representing the interests of society and nominating their candidates for political office. Elections are also defined as an arena for competition to fill political positions in government based on the formal choice of qualified citizens. Elections are the most important mechanism for the continuation of representative democracy, as a way for the people to choose their representatives (Pamungkas, 2009).

In the 2024 Election, the Partai Demokrasi Indonesia Perjuangan (PDIP) is one of the political parties that will participate in political contestation. PDIP is a political party that has a long history and significant influence in Indonesian politics. This party was founded by Indonesian struggle figure, Megawati Soekarnoputri, in 1999 and has become the largest party in several elections since then. PDIP is considered a political party that has strong attachments and links with the ideology of marhaenism. This is drawn from the history of the PDIP as the political relay of the PDI (Geraldy, 2019).

In the increasingly developing digital era and social media, sentiment analysis has become an important tool for understanding public perceptions and opinions on various things, including political parties. Sentiment analysis or opinion mining is the process of understanding, extracting and processing textual data automatically to obtain sentiment information contained in an opinion sentence. The magnitude of the influence and benefits of sentiment analysis has caused research and applications based on sentiment analysis to grow rapidly. Even in America there are around 20-30 companies that focus on sentiment

analysis services (Go & Huang, 2009). Sentiment analysis allows us to explore people's thoughts, feelings and responses to a particular entity, in this case, PDIP.

In this research, sentiment analysis will be carried out towards PDIP in the context of the 2024 Election. The data used comes from public opinion on Twitter social media. Social media, especially Twitter, has now become an effective and efficient promotion or campaign (Buntoro, 2017). Based on 2010 statistics, Twitter has 106 million accounts and as many as 180 million unique visitors each month. The number of Twitter users is said to continue to increase by 300,000 users every day (Yarrow, Clausen, & Robbins, 2010). The aim of this analysis is to gain a deeper understanding of how people respond, discuss and express opinions about PDIP on Twitter social media.

Through sentiment analysis, we can identify positive, negative or neutral trends related to PDIP. The results of this analysis can provide valuable insights for political parties in understanding public perception and support and enable them to formulate more effective strategies in their political campaigns.

This research will consist of several parts, starting with the sentiment analysis method used. Next, an exploration of data obtained from Twitter social media related to PDIP will be carried out. The results of the analysis will be described and analyzed to describe the general sentiment associated with PDIP. Finally, conclusions will be drawn based on the findings of the sentiment analysis and provide an outlook for the 2024 election.

METHOD

This research aims to produce high accuracy in the classification of opinions regarding the 2024 Partai Demokrat Indonesia Perjuangan (PDIP) on Twitter social media using the Naive Bayes Algorithm.

Sentiment Analysis

Sentiment analysis is a method or technique used to evaluate and identify the sentiments or opinions contained in text, such as articles, product reviews, tweets or other social media posts. Sentiment analysis is a process that aims to determine whether the contents of a dataset in the form of text (documents, sentences, paragraphs, etc.) are positive, negative or neutral (Kontopoulos, Berberidis, Dergiades, & Bassiliades, 2013).

In this era of rapid technological development, companies really need an analyst to evaluate the products they are promoting, so that in recent years an analyst has really needed to analyze and apply new methods to maintain the existence of their products. Sentiment analysis is a fairly popular research field, because it can provide benefits for various aspects, ranging from sales prediction (Liu, Huang, An, & Yu, 2007), politics (Park, Ko, Kim, Liu, & Song, 2011), and investor decision making (Dergiades, 2012).

Text Mining

Text mining, also known as text analysis or text exploration, is the process of extracting valuable and meaningful information from documents or unstructured text. The main goal of text mining is to identify patterns, trends, insights, or new knowledge contained

SINOMICS JOURNAL
International Journal of Social Science, Education, Communication and Economics
ISSN (e): 2829-7350 | ISSN(p): 2963-9441

in the text. In the current era, the use of text mining is very necessary to visualize or evaluate knowledge from large collections of text documents (Deolika, Kusrini, & Luthfi, 2019).

With text mining, we will carry out the process of searching or extracting useful information from textual data (Han & Kamber, 2006). By applying the processes in text mining, data patterns, trends and extraction of potential knowledge from text data will be obtained (Kao & Poteet, 2005).

Public Figure Review

Opinions about public figures are numerous and easy to find in cyberspace, but turning these opinions into useful information is very difficult. To make it easier to observe public figures, the social media Twitter is used to collect opinions expressed by the public so that various opinions can be collected. In this research, the public figure appointed is the Partai Demokrasi Indonesia Perjuangan (PDIP) for the 2024 period.

Naïve Bayes Classifier

Naïve Bayes is a classification algorithm based on Bayes' theorem with simple assumptions called "naïve" or "plain". This algorithm is very popular in natural language processing, text analysis, and classification of data with discrete features. Basically, the Naïve Bayes algorithm uses conditional probability to classify data into appropriate categories. This algorithm assumes that all features or attributes used for classification are independent of each other, that is, there is no dependency between the features. The Naive Bayes Classifier method is a method that can be used in decision making to get better results in a classification problem (Harimurti & Riksakomara).

The Naive Bayes Classifier method takes two stages in the text classification process, namely the training stage and the classification stage. At the training stage, a process is carried out on data samples that can be as representative of the data as possible. Next is determining the prior probability for each category based on the data sample. At the classification stage, the category value of a data is determined based on the terms that appear in the classified data. Naïve Bayes' theorem can be stated in equation 1

$$P(Xk\,|Y) = P(Y|Xk) \,/\, \sum i\; P(Y|Xi) \quad \text{.....(1)}$$

Where, the Probability state of Xk in Y can be calculated from the Probability state of Y in Xk divided by the sum of all Y probabilities in all Xi. To be able to classify a tweet, in this research the author uses the Naive Bayes Classifier method for text classification, as done in equation 2.

$$P(V1|C = c) = CountTerms(v1, docsv(c)) \,/\, AllTerms(docs(c)) \quad \text{.....(2)}$$

Where v1 in this study is one particular word in a tweet, while CountTerms?v1 , docsv?c?? refers to the number of occurrences of a word labeled c ("positive" or "negative" or "neutral"). AllTerms1docs?c?2 refers to the number of all words labeled c in the dataset. To avoid zero values in the probability, Laplace (add-one) smoothing is applied. The purpose

of smoothing is to reduce the probability of the observed outcome/output, and at the same time increase/increase the probability of the unobserved outcome/output, so that the equation becomes as follows:

$$P(V1|C = c) = CountTerms(v1, docsv(c)) + 1 / AllTerms(docs(c)) + |V| \quad \text{.....(3)}$$

Where |V| refers to the number of all words in the tweets in the dataset.

Confusion Matrix

Confusion Matrix is a table with four different combinations of predicted and actual values. This performance evaluation uses F1 Score and accuracy, F1 Score is a comparison of recall and precision, while accuracy describes how accurate the model is in classifying correctly. Confusion matrix is a method that is usually used to calculate accuracy in data mining concepts (Rahman, Alamsah, Darmawidjadja, & Nurma, 2017). An example of a Confusion Matrix is in Table 1.

Table 1 Confusion Matrix

Predicted Class	Actual Class	
	Negative	Positive
Positive	True Positive (TP)	False Positive (FP)
Negative	True Positive (TP)	True Negative (TN)

Four output values are obtained based on Table 1:

Recall

Calculation to find out what percentage of cases are correctly identified as correct.

$$Recall = TP / TP+FN \quad \text{.....(4)}$$

Accuracy

Accuracy results are classified with other data to determine the accuracy of the model used.

$$Accuracy = TP+TN / TP+FP+FN+TN \quad \text{.....(5)}$$

Precision

The comparison between the TP classification results is classified as positive with all positive predictions.

$$Precision = TP/ TP+FP \quad \text{.....(6)}$$

F1 Score

Comparison of weighted average precision and recall.

$$F1\ Score = 2*recall*precision / recall+precision \quad \text{.....(7)}$$

SINOMICS JOURNAL
International Journal of Social Science, Education, Communication and Economics
ISSN (e): 2829-7350 | ISSN(p): 2963-9441

RESULTS AND DISCUSSION

At this stage it explains starting from data collection, data processing, data sharing and data testing using the Naïve Bayes method.

Project Stages

The flow in creating an analysis program is carried out in several stages as shown in Figure 1 below.

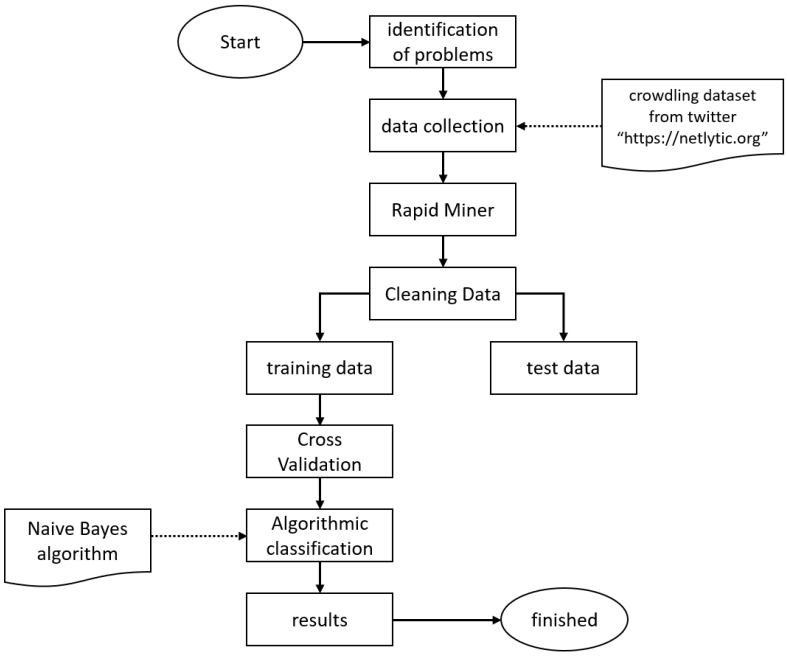


Figure 1 project flow

Crawling Data

Opinion data collection from Twitter is carried out with the help of the website https://netlytic.org/ which provides tools for crawling datasets. With the Netlytic website, we can retrieve Twitter data with a certain time range, certain keywords or export it in a certain format. In this study, 2500 opinions were used with positive and negative values which were manually labeled. For examples of tagged tweets, see the table below.

Table 2 Twitter Opinion

Text	Sentimen
@__AnakKolong Kalau Mbk Puan jadi calon. Terus PDIP menolak terus RUU perampasaan Aset... 2024 fix PDIP nyungsep	Negatif
Sy pemilih PDIP sejak ORBA hingga kini. Sy pemilih yg rasional. Dan kini sy pesimis dngn partai lama. Tahun 2024, suara untuk @psi_id. Sy sdh mulai bergerilya di WAG alumni dll. Terlihat bnyk jg yg tadinya PDIP, mulai lirik ke PSI. Semoga 2024, suara PSI melonjak dngn signifikan.	Positif

A positive label means that the person making the tweet agrees or supports the public figure being discussed, and vice versa, if the label is negative then the community does not agree or does not support the public figure being discussed.

Pre-processing

Before entering the pre-processing process, this research carried out a sub-process where in the Rapidminer application the operator that can be used is subprocess. In the subprocess operator, the steps to go through are removing the Uniform Resource Locator or better known as the URL by using the Remove URL operator and configuring it by using the Regular Expression (Regexp) http\S+|\S+co\S+ then all URLs will be generated on every tweet will disappear.

The next sub-process in this study is to delete emoticons and change them to a word that is in each tweet so that the opinions conveyed by the public through Twitter social media become easier to analyze.

After the sub-process is complete, the next stage, the data that has been prepared is then pre-processed where existing opinions will get tokenizing, stopwords and steming processes. In the rapid miner application, the pre-processing process is carried out by the document processing operator. The following is a pre-processing opinion on this project

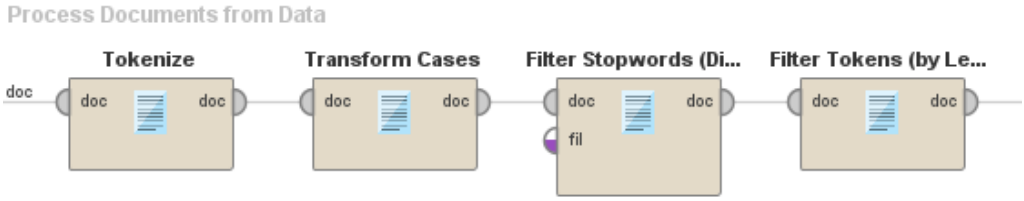


Figure 2 Pre-Pocessing Data

The Transform Case operator is useful for changing all existing text into lowercase or non-capital letters, then the tokenize operator is useful for breaking each sentence into words for further processing by the stopwords filter operator, where the stopwords filter operator is useful for removing words that are not needed in the next process. Below are examples of words that are included in the stopwords file.

The list of unneeded words is in one file which must be opened first using the open file stopword operator.

The next process in pre-processing is stemming or in other words removing affixes in each word that has been previously processed so that the output from this pre-processing process can be used for calculations using the naïve Bayes classifier algorithm or support vector machine. For the stemming process in this study using the Rapidminner application

SINOMICS JOURNAL
International Journal of Social Science, Education, Communication and Economics
ISSN (e): 2829-7350 | ISSN(p): 2963-9441

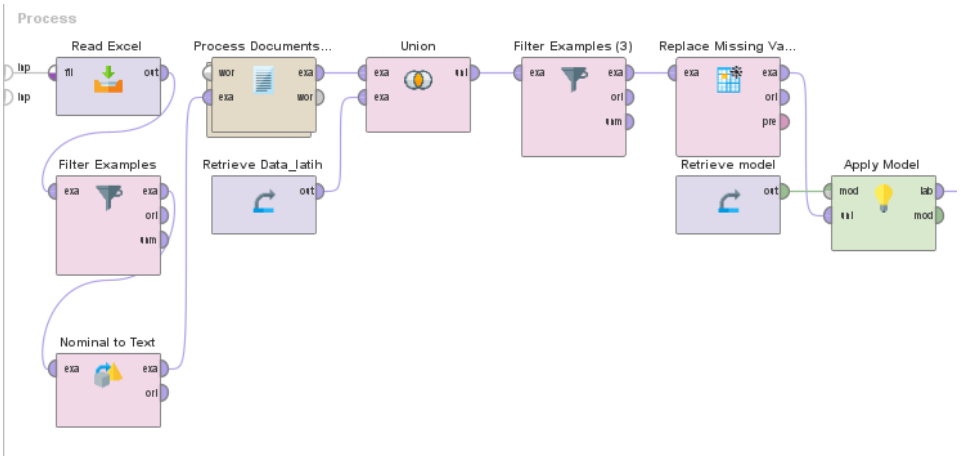
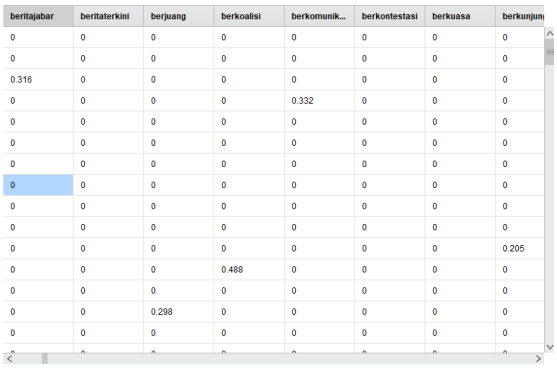


Figure 3 Rapidminer Process Suite

TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a method used to measure the importance of a word in a document in the context of a larger collection of documents. This method is often used in natural language processing, information retrieval, and text mining.

Figures 4 and 5 are the results of the TF-IDF that has been carried out.



beritajabar	beritaterkini	berjuang	berkoalisi	berkomunik...	berkontestasi	berkuasa	berkunjun
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0.316	0	0	0	0	0	0	0
0	0	0	0	0.332	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	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.205
0	0	0	0.488	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0.298	0	0	0	0	0
0	0	0	0	0	0	0	0

Figure 4 Result TF-IDF

Row No.	word	in documents	total
1	pdip	2085	2446
2	prabowo	382	912
3	pilpres	765	798
4	radikal	220	660
5	jokowi	462	584
6	koalisi	535	561
7	pemilu	544	547
8	capres	507	544
9	partai	460	490
10	httpstco	488	489

Figure 5 Result All TF-IDF

Word Cloud

Word Cloud is a visual representation of a collection of words in text displayed in the form of a word cloud, where the words that appear more frequently in the text have a larger size and are more prominent in the visual display, data visualization is displayed to make it easier for someone to see and interpret the meaning and results obtained.

Figure 6 is a continuation of the TF-IDF process where in the large image that has PDIP written on it is the amount of text that often appears compared to other text/words.

Figure 6 Word Cloud Analysis Results

Naïve Bayes Classifier

In this study using one algorithm, namely the Naïve Bayes Classifier. This is used to see the level of accuracy of the algorithm to carry out the sentiment analysis process. With the help of the rapidminer application, the Naïve Bayes Classifier operator is used for the experiment as shown below

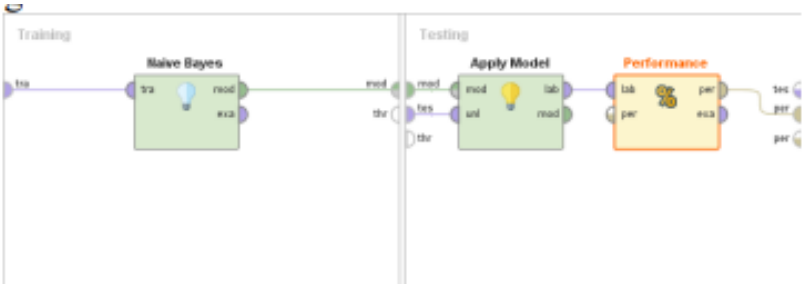


Figure 7 Naïve Bayes Classifier Operator

In the Naïve Bayes Classifier algorithm, the dataset used is 588 in total which have been cleaned or filtered in the previous process, then the data is divided into 70% training data and 30% test data by dividing the training data used 410 which have been given sentiment and 178 test data which not given sentiment. The resulting accuracy value is 98.05% as in the image below.

SINOMICS JOURNAL
International Journal of Social Science, Education, Communication and Economics
ISSN (e): 2829-7350 | ISSN(p): 2963-9441

accuracy: 98.05%

	true positif	true negatif	class precision
pred. positif	269	0	100.00%
pred. negatif	8	133	94.33%
class recall	97.11%	100.00%	

Figure 8 Naïve Bayes Classifier Accuracy

CONCLUSION

These results show that the system or model has perfect precision (100%) in classifying the positive class, which means all the positive predictions given are correct. However, the recall for the positive class is 97.11%, which means that there are some positive cases that are not detected by the system or model.

For the negative class, the precision is 94.33%, which means that a fraction of the negative predictions may be wrong. However, the recall for the negative class is 100%, meaning all negative cases are correctly identified.

REFERENCES

Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. INTEGER: Journal of Information Technology, 2(1).

Deolika, A., Kusrini, K., & Luthfi, E. T. (2019). Analisis Pembobotan Kata Pada Klasifikasi Text Mining. (JurTI) Jurnal Teknologi Informasi, 3(2), 179-184.

Dergiades, T. (2012). Do investors' sentiment dynamics affect stock returns? Evidence from the US economy. Economics Letters, 116(3), 404–407. doi:10.1016/j.econlet.2012.04.018

Geraldy, G. (2019). Ideologi dan Partai Politik: Menakar Ideologi Politik Marhaenisme di PDIP, Sosialisme Demokrasi di PSI dan Islam Fundamentalisme di PKS. Ideology and Political Parties: Measuring the Political Ideology of Marhaenism in PDIP, Democratic Socialism in PSI and Islamic Fundamentalism in PKS..

Go, A., Huang, L., & Bhayani, R. (2009). Twitter Sentiment Analysis. Final Project Report, Stanford University, Department of Computer Science.

Han, J., & Kamber, M. (2006). Classification and prediction. Data mining: Concepts and techniques, 2006, 347-50.

Harimurti, F. A., & Riksakomara, E. (2017). Klasifikasi Penerimaan Beasiswa Menggunakan Metode Naïve Bayes Classifier (Studi Kasus Universitas Trunojoyo Madura) (Doctoral dissertation, Doctoral dissertation, Institut Teknologi Sepuluh Nopember).

Kao, A., & Poteet, S. (2005). Text mining and natural language processing: introduction for the special issue. ACM SIGKDD Explorations Newsletter, 7(1), 1-2.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. Expert Systems with Applications, 40(10), 4065–4074. doi:10.1016/j.eswa.2013.01.001

Liu, Y., Huang, X., An, A., & Yu, X. (2007). ARSA: A SentimentAware Model for Predicting Sales Performance Using Blogs. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07 (p. 607). New York, New York, USA: ACM Press. doi:10.1145/1277741.1277845

Marian Radke Yarrow, John A. Clausen and Paul R. Robbins (2010). The Social Meaning of Mental Illness. Journal of Social Issues. Volume 11, Issue 4, pages 33–48, Fall 1955.

Pamungkas, S. (2009). Perihal pemilu. Laboratorium Jurusan Ilmu Pemerintahan dan Jurusan Ilmu Pemerintahan, Universitas Gadjah Mada.

Park, S., Ko, M., Kim, J., Liu, Y., & Song, J. (2011). The Politics of Comments: Predicting Political orientation of News Stories with Commenters ' Sentiment Patterns.

Rahman, M. F., Alamsah, D., Darmawidjadja, M. I., & Nurma, I. (2017). Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN). Jurnal Informatika, 11(1), 36.