Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 993-1000
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Sentiment Analysis on Platform X (Twitter) Towards The 2024 General Election Using The Probabilistic Neural Network Algorithm

Aaron Rumondor[1], Fitrah Rumaisa[2]
[1,2]Informatics Study Program, Faculty of Engineering, Universitas Widyatama, Indonesia
Email: aaron.lorenzo@widyatama.ac.id[1], fitrah12@gmail.com[2]

Abstract

Sentiment analysis is the process of analyzing an opinion or public opinion regarding a phenomenon that has occurred, is currently occurring or will occur. Sentiments that are commonly discussed are the public's assessment of an object positively, negatively or neutrally. Twitter is the most popular media used to express all forms of public emotions and opinions in the form of tweets or text. The issues discussed in it include many events, such as the 2024 general election which will be held in Indonesia. This media is easily accessible to many people to show other people's opinions regarding existing phenomena. This research discusses the topic after the 2024 general election with opinions based on three sentiment classes, namely positive, neutral and negative. The aim of this research is to build a sentiment analysis system by applying the Probabilistic Neural Network algorithm as a classification model. The method used is data collection, cleaning, preprocessing, TF-IDF word weighting, labeling, classification models, and evaluation of results. The data used amounted to 2002 data with a division of 1035 positive tweets, 693 neutral tweets and 274 negative tweets. The program was built using Google Colaboratory and the Python programming language. Testing was carried out with 3 (three) comparisons, namely 90:10, 70:30, and 50:50. By comparing 90% training data and 10% testing data, the greatest model accuracy was obtained with a value of 88.42% and taking into account the evaluation using the confusion matrix and precision parameters of 89%, recall of 88%, and f1-score of 88%. Evaluation was also carried out for website-based applications on new data with an accuracy value of 66%.

Keywords *- Sentiment Analysis, 2024 General Election, Probabilistic Neural Network, Confusion Matrix*

1. INTRODUCTION

General elections are a form of popular sovereignty to produce a democratic state in accordance with the state principles of Pancasila and the 1945 Constitution of the Republic of Indonesia. Freedom of expression in Indonesia is a right protected by the constitution[1].

In this modern, digital era, freedom of the press and opinion has become the main polemic for discussing public opinion on social media related to politics and elections. With this phenomenon, the public provides opinions and opinions textually through social media such as the Twitter application (now "X"). The X application (Twitter) is a platform that is often used by the public to convey their aspirations, therefore this platform is also widely used to analyze public sentiment regarding current issues.

Sentiment analysis was previously discussed by Ina Najiyah & Ifani Haryanti with the title "Covid-19 Sentiment Analysis using Probabilistic Neural Network and TF-IDF Methods". This research discusses sentiment analysis with a dataset of 1177 taken from Facebook, Twitter and Instagram with a division of 560 positive, 355 negative and 262 neutral. The accuracy obtained in the training used a

probabilistic neural network and produced an accuracy of 86% and classified errors of 11% as seen from the confusion matrix [2]. Furthermore, previous research discussed analyzing public sentiment on Twitter regarding the 2024 election using the Naïve Bayes algorithm (Salim Puad et al, 2023). The level of accuracy of sentiment analysis in the 2024 General Election using the Naïve Bayes algorithm was tested using the Split data method, by dividing the data into four models, 90:10, 80:20, 70:30, and 60:40[3].

This research uses the Probabilistic Neural Network (PNN) algorithm to classify data of 2,000 tweets in order to obtain a higher and better level of accuracy. PNN is an algorithm that uses a probability or chance function. PNN is often used in classification because it can map each input pattern to the optimal number of classifications, is faster and more accurate than other neural network models [4].

2. RESEARCH METHODOLOGY

2.1. Implementation Method

In this research, 2 (two) methods were used to obtain data and implementation as follows:

a) Data Collection Methods

1) Literature Study. Research data is collected using references from relevant journals or theses, thereby helping to write research more systematically according to the problems discussed.
2) Google Colab. Cloud-based platform for writing, running and processing Python code via a web browser online. This platform is used to crawl data from application X (Twitter) using the Twitter API.
3) Twitter API. A system that bridges communication between an application and a server. Thus, the data collection method is obtained by taking the Twitter API from a personal account and then running it on Python code to crawl the data.

b) Implementation Methods

In implementing the model proposed in this research, this research was carried out in several stages.

1) Analysis. Analysis of a case is carried out after the data has been successfully collected through data collection or crawling X (Twitter) data on Google Colab. The data will be sorted and analyzed whether the data is suitable for processing or needs to be collected again.
2) Implementation. The implementation process is carried out in several stages such as pre-processing, sentiment labels into training and testing data, and confusion matrix.
3) Testing. Testing is a discussion to obtain the results of the data mining process on sentiment analysis on X (Twitter) regarding the 2024 general election. From these results, the accuracy of the algorithm in classifying data will be obtained.

c) General Election

According to Law no. 15 of 2011, Article 1 paragraph (1) concerning the holding of general elections states and explains the meaning of general elections, hereinafter referred to as elections, as: A means of implementing people's sovereignty which is held directly, publicly, freely, confidentially, honestly and fairly in a Unitary State The Republic of Indonesia is based on Pancasila and the 1945 Constitution of the Republic of Indonesia[5].

d) Sentiment Analysis

Sentiment analysis, or also called opinion mining, is a Natural Language Processing (NLP) technique for analyzing emotions from a text [6]. Sentiment analysis is also useful for understanding consumer preferences and behavior towards a product. This can help businesses to improve the service quality of their business products. This analysis also helps businesses monitor brand image on social media so that their image is positive, neutral or negative and can be improved to take action for the progress of a company's business.

e) X (Twitter)

Twitter is a communication medium that allows users to interact and discuss topics that are trending or hotly discussed among the general public in the form of text, images or videos [7].Twitter, renamed X in July 2023[8], is an online social media and social networking service operated by the American company, X Corp., the successor to Twitter, Inc. This name change was initiated by Twitter CEO Elon Musk, who is also CEO of X Corp.

2.2. Text Preprocessing Data

Data preprocessing is a process for processing raw data into quality data (good input in data mining). Usually at this stage data will be eliminated that is not in accordance with research needs [9]. Text preprocessing is a process for selecting text data to make it more structured by going through a series of stages which include case folding, tokenizing, filtering and stemming stages.

2.3. Probabilistic Neural Network

Probabilistic Neural Network (PNN) is a Bayes theorem method used for conditional probability by estimating a function regarding the probability of random variables [10]. PNN is a Neural Network introduced by Donald Specht in 1990 and is used to perform non-liner calculations by estimating the probability density function of a dataset using parzen probability density estimation. The PNN algorithm has several layers used, including the input layer, pattern layer, summation layer, and output layer. The architecture of the PNN network can be seen in Figure 2.3 as follows [3].

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 993-1000
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

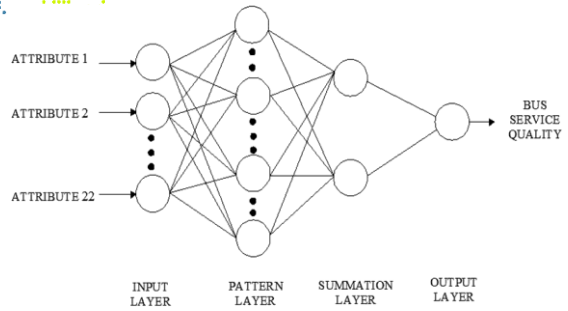

Figure 1. Probabilistic Neural Network

a) Input Layer
b) Pattern Layer. The Pattern Layer will calculate the probability of the distance between the input and the data stored in the pattern layer neurons.

$$f(x) = \exp(-\frac{(x-x_{ij})^T-(x-x_{ij})}{2\sigma^2}) \quad (1)$$

c) Summation Layer. This stage receives input from each pattern layer neuron and adds them together to obtain several possibilities for input x to fall into a group.

$$p(x) = \frac{1}{(2\pi^{k/2}\sigma^k t)} - \frac{1}{n^t}\sum_{j=1}^{t} f(x) \quad (2)$$

d) Output Layer

2.4. Confusion Matrix

Confusion Matrix is a method that can be used to measure the performance of a classification method in calculating accuracy in data mining concepts. Confusion matrix is also one way of visualizing system learning results, the visualization displayed contains two or more categories (Rahman, et al., 2017).

Klasifikasi		Kelas Sebenarnya	
		Kelas = Ya	Kelas = No
Kelas Prediksi	Kelas = Ya	True Positive (TP)	False Negative (FN)
	Kelas = No	False Positive (FP)	True Negative (TN)

Figure 2. Confusion Matrix

2.5. Research Method

The research methodology in this final assignment is divided into several plans and flows that are adapted to the needs in writing the final assignment.

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 993-1000
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

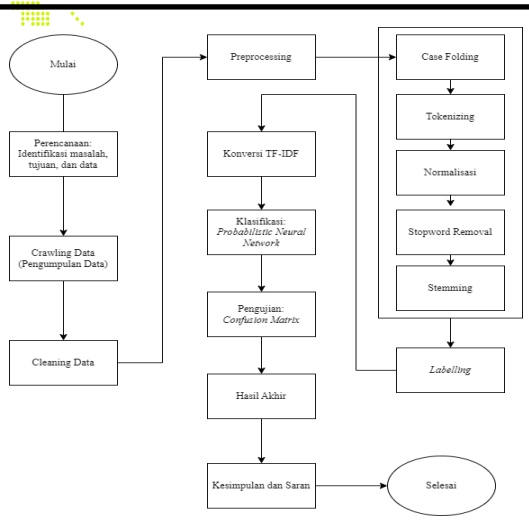

Figure 3. Sentiment Analysis Flowchart

3. RESULT AND DISCUSSION

3.1. Crawling Data

At this stage, a cloud-based platform, namely Google Colaboratory, is used to run the data crawling program from the X application (Twitter). Previously, an API was obtained from Twitter to bridge the X application (Twitter) with Google Colaboratory so that Google could obtain raw data from Twitter and then process it to become the data needed to write this research with predetermined limitations. The dataset stage will display 2024 election data stored in Comma Separated Values (CSV) format and totaling 2,002 tweets.

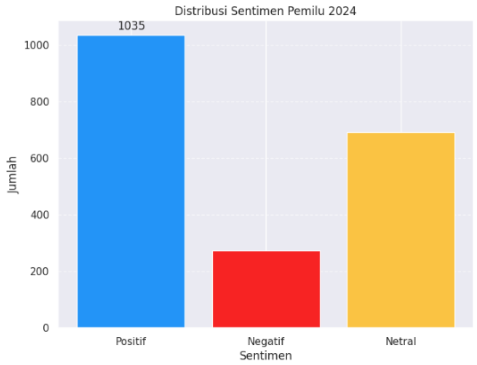

Figure 4. Distribution Of Election Sentiment

3.2. Preprocessing Data

Data preprocessing or data pre-processing stage is the stage for processing raw data. Data Preprocessing is one of the stages in data mining.

Table 1. Preprocessing Stages

Word Example:	@godhongcoffee Dukung pemilu sebagai bentuk dukungan terhadap peningkatan akses air bersih dan sanitasi #PemiluJujur
Preprocessing	**Result**

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 993-1000
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Cleaning	Dukung pemilu sebagai bentuk dukungan terhadap peningkatan akses air bersih dan sanitasi
Case Folding	dukung pemilu sebagai bentuk dukungan terhadap peningkatan akses air bersih dan sanitasi
Tokenizing	'dukung', 'pemilu', 'sebagai', 'bentuk', 'dukungan', 'terhadap', 'peningkatan', 'akses', 'air', 'bersih', 'dan', 'sanitasi'
Normalization	dukung pemilihan umum sebagai bentuk dukungan terhadap peningkatan akses air bersih dan sanitasi
Stopwords Removal	'dukung', 'pemilihan', 'umum', 'bentuk', 'dukungan', 'peningkatan', 'akses', 'air', 'bersih', 'sanitasi'
Stemming	'dukung', 'pilih', 'umum', 'bentuk', 'dukung', 'tingkat', 'akses', 'air', 'bersih', 'sanitasi'

3.3. Labelling Data

The labeling process is classify the message or meaning of a sentence based on the emotional expression contained in it. These signs or categories are usually divided into 3 (three) sentiments, namely positive, neutral and negative.

Table 2. Labelling Process

No	Text	Label
1	dukung pemilihan umum sebagai bentuk dukungan terhadap peningkatan akses air bersih dan sanitasi	Positif
2	anjim bacain komentar bikin naik pitam parah padahal kalau janjinya ditepatin juga mereka sendiri yang untung	Negatif
3	pemilihan umum 2024	Netral

3.4. Conversion TF-IDF

In general, TF-IDF is a statistical measure that describes the importance of a term in a document in the form of a collection. The frequency with which a word appears in a given document indicates how important it is in that document.

$$tf = 0{,}5 + 0{,}5 \times \frac{tf}{\max(tf)}$$

$$idf_t = \log\left(\frac{D}{df_t}\right)$$

$$W_{d,t} = tf_{d,t} \times IDF_{d,t}$$

Formula Explanation:

Tf = the amount of data searched for in a document;
max(tf) = the highest number of occurrences of a term in the same document
value D = total documents;
df_t = number of documents containing the term t;
IDF = *Inversed Document Frequency* ($\log(D/df_t)$);
d = d document;
t = t word of the keyword;
W = weight of the d document against the t word;

3.5. Classification Process

The research will use 3 comparisons with different ratios, namely 90:10, 70:30, and 50:50, where 90%, 70%, and 50% will be used as training data and 10%, 30%, and 50% used as testing data. Accuracy results using the PNN algorithm will show the Final Results with the percentage of accuracy obtained for each ratio comparison.

Table 3. Accuracy Result Based On Data Testing And Training

Data Training	Data Testing	Accuracy
90	10	88.42%
70	30	86.48%
50	50	82.29%

3.6. Confusion Matrix Evaluation

Evaluation of the performance of the model used in this research will use the Confusion Matrix with the parameters accuracy, precision, recall and f1-score. Model testing results on data with a ratio of 90:10 with data division 1800 tweets for training data and 200 tweets for testing data. Model testing results on data with a ratio of 70:30 with data division 1400 tweets for training data and 600 tweets for testing data. Model testing results on data with a ratio of 50:50 with data division 1000 tweets for training data and 1000 tweets for testing data.

Table 4. Classification Result

Data Comparison		Data Amount		Accuracy	Sentiment	*Precision*	*Recall*	F1-Score
Data *Training*	Data *Testing*	Data *Training*	Data *Testing*					
90	10	1800	200	88.42%	Positive	0.88	0.78	0.83
					Negative	0.95	0.97	0.96
					Neutral	0.82	0.90	0.86
					Average	88%	88%	88%
70	30	1400	600	86.48%	Positive	0.86	0.77	0.81
					Negative	0.92	0.97	0.95
					Neutral	0.81	0.85	0.83
					Average	86%	86%	86%
50	50	1000	1000	82.29%	Positive	0.77	0.76	0.76
					Negative	0.90	0.95	0.92
					Neutral	0.79	0.75	0.77
					Average	82%	82%	82%

3.7. Web-Based Application

The results of the comparison for data input via the web which has been built based on the PNN classification model show appropriate accuracy values if the data has been tested. This test was carried out to assess the accuracy results of the PNN model on data newly input by the user on the website. Each parameter in the confusion matrix for new data obtained an average accuracy value of 0.66 or 66%. Testing on this website-based application uses all test data, namely 15 tweets which are input and the model is evaluated.

4. CONCLUSION

From the results of this research for sentiment analysis, several conclusions can be drawn as based on this classification model, it can be concluded that the most sentiment shows a label of positive sentiment with a total of 1035 tweets. The PNN model classification results with the highest accuracy were obtained at a comparison ratio between training data and testing data, namely 90:10 with an accuracy value of 88.42%. Respectively, the data comparison ratio is 70:30 with an accuracy of 86.48% and a comparison ratio of 50:50 with an accuracy of 82.29%. These results show that the website received an average score of 66% for accuracy. However, the average accuracy results based on text or sentences are quite large, namely above 80% and even reaching 99%.

REFERENCES

[1] Pratama, Rizky. 2022. Hate Speech: Deviations from the ITE Law, Freedom of Opinion and the Values of Dignified Justice. Pelita Harapan University. Tangerang.
[2] Najiyah, Ina. 2021. Sentiment Analysis of Covid-19 using Probabilistic Neural Network and TF-IDF Methods. Accessed 21 February 2024 from https://ejurnal.ars.ac.id/index.php/jti/article/view/488.
[3] Puad, Salim. 2023. Analysis of Public Sentiment on Twitter Regarding the 2024 General Election Using the Naïve Bayes Algorithm. Accessed 21 February 2024 from https://ejournal.itn.ac.id/index.php/jati/article/download/6920/4114/.
[4] Yasin, H., & Ispriyanti, D. 2017. Classification of Baby Birth Weight Data Using Weighted Probabilistic Neural Network (WPNN) (Case Study at Sultan Agung Islamic Hospital Semarang). Statistics Media. Semarang.
[5] Indonesia. Law Number 15 of 2011 concerning General Election Organizers. State Secretariat. Jakarta.
[6] Revoupedia Editorial Team. 2024. What is Sentiment Analysis. Revoupedia.com. Jakarta.
[7] Scientific Journal Center Editorial Team. 2021. Definition, History and Benefits of Twitter for Millennial Young People. Center for Scientific Journals. Medan.
[8] Davis, Wes. 2023. Twitter is being rebranded as X (Twitter changed its name to X). Retrieved February 24, 2024 from https://www.theverge.com/2023/7/23/23804629/twitters-rebrand-to-x-may-actually-be happening-soon.
[9] Suripto. 2022. Pre-Processing and Classification Techniques in Data Science. Binus University. Jakarta
[10] Anzir, Oumfa. 2019. Application of the Probabilistic Neural Network Method for Classifying the Pekanbaru City Family Hope Program. Sultan Syarif Islamic University. Pekanbaru.