JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 5 No. 1 (2024)
https://doi.org/10.35877/454RI.jinav2775

Sentiment Analysis Against Rohingya Immigrants On Twitter Using The Support Vector Machine Method

Lailatul Husna Aulia* & Sriani

Departement of Computer Science, Faculty of Science and Tecnology, Universitas Islam Negeri Sumatera Utara, Medan, 20236, Indonesia

Abstract

The influx of Rohingya migrants into Indonesia has sparked diverse reactions from the public, ranging from both positive and negative perspectives. These viewpoints have surfaced prominently on Twitter, where netizens express conflicting opinions that often lead to division and discord. This study seeks to examine the sentiment of public opinion regarding Rohingya immigrants on Twitter, employing a Support Vector Machine with RBF kernel implemented in Python as its analytical method. The sentiment resulting from the crawling process on Twitter was 1347 pieces of data. In the analysis, the comparison between training data and test data was 8: 2. The dataset after preprocessing consisted of 1321 data, 1056 of which were training data while 265 were test data. The results of sentiment analysis show that the SVM method can be used to analyze sentiment, the accuracy value obtained is 72%, precision is 100%, recall is 2%, and f1-Score is 3%.

Keywords: sentiment analysis, rohingya, twitter, support vector machine.

Received: 10 April 2024 Revised: 6 July 2024 Accepted: 20 July 2024

1. Introduction

Indonesia serves as a destination for the Rohingya ethnic group seeking refuge. The Rohingya have inhabited Myanmar and Bangladesh for centuries. Their recent arrival as refugees in Aceh, traveling by boat, surprised both the Indonesian public and government, as well as international and regional communities and various humanitarian organizations (Yulian Azhari, 2022). The Rohingya are considered illegal immigrants since they entered the country without authorization and lack valid identification documents. A UNHCR survey conducted among roughly 1,543 people suggests that since mid-November 2023, a total of 1,543 Rohingya refugees have arrived in Aceh. Additionally, there were already 140 Rohingya refugees residing in the Mina Raya Foundation complex, Pidie Regency. This brings the total number of Rohingya refugees in Aceh to 1,683. This influx of refugees has sparked discussions on social media platforms like Twitter, highlighting the ongoing debate about how Indonesia should respond to the situation (Zulkarnain, 2020).The arrival of Rohingya refugees has sparked a heated debate on Twitter, with netizens expressing strong opinions on both sides. Some posts highlight potential negative consequences if the situation isn't managed effectively, calling for a coordinated national and international response. Others, driven by compassion, express sympathy for the refugees' plight and the violence they've endured (Dhea Nada, 2021).

In this case, it is necessary to design a sentiment analysis system model for public opinion on Twitter to understand people's reactions, views and anxiety towards the Rohingya so that it is more relevant. Information obtained from Twitter is not completely structured and often experiences noise. The system model was built using the non-linear kernel Support Vector Machine algorithm. This algorithm is used to classify public comments in the form of text on Twitter. So it will be easier to find out the accuracy, precision, recall and F1-Score of the algorithm used. The advantage of this algorithm is its good performance for classifying data and producing a high level of accuracy (Ariansyah & Kusmira, 2021).

* Corresponding author.
E-mail address: lailatulhusnaaulia@gmail.com

2. Research Methodology

This research outlines a step-by-step plan (framework) for conducting the study. It will employ a quantitative approach, which relies on numerical data analysis. The following sections detail the various stages involved in this research process.

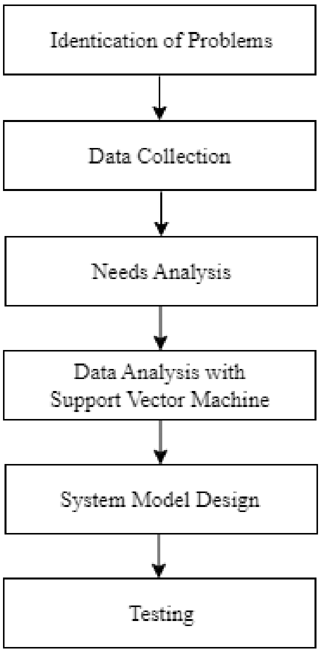

Figure 1. Research Framework.

2.1. Data Collection

At the data collection stage, a data crawling technique was carried out on Twitter social media using tweet harvest and Twitter auth token. This is done through Google Colaboratory and the Pandas library. Crawling data was carried out to look for tweets or public comments regarding Rohingya immigrants with the keyword “Rohingya” in Indonesian. The data used in this research is the result of crawling 1347 data. The data obtained is presented in .csv format. Next, after going through the text preprocessing stage, the data is given a label or class for each sentiment.

2.2. Pre-processing data

Since the collected data was raw and unorganized (unstructured), it needed cleaning before analysis. This text pre-processing step aimed to remove irrelevant information and improve the data quality for better computational results (Hermawan & Bellaniar Ismiati, 2020).

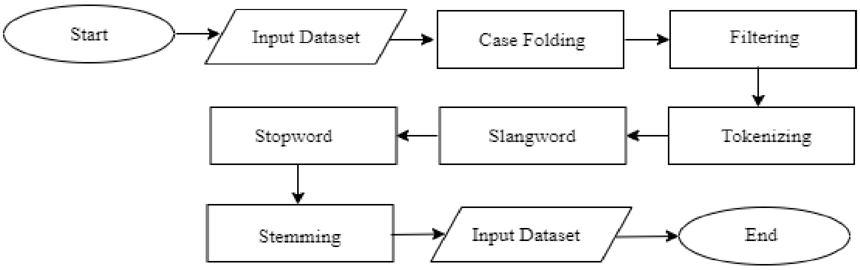

Figure 2. Preprocessing flow.

The cleaning process involved several techniques like To analyze the collected Twitter data, which was essentially raw and unorganized (unstructured), we performed several text cleaning steps. These steps are as follows:

a. Case Folding involves converting all letters to lowercase, regardless of whether they are uppercase or lowercase.
b. Filtering refers to the process of removing punctuation, symbols, and other unnecessary elements like URL from thedata.
c. Tokenizing is the process of dividing text into individual words; for example, "I go to school" is split into "I," "go," and "to".
d. Slangword in the context of sentiment analysis Refers to words or phrases used in informal conversations and often have special meanings or connotations in certain circles. slangword to change an abbreviated word into a basic word.
e. Stopword Removal consists of eliminating conjunctions and common words such as "to," "in," "and," "he," "us," "me," and "my."
f. Stemming Stemming is a morphological process in natural language processing that involves removing the beginning or end of a word to produce the basic form or root of the word.

2.3. TF-IDF Vectorization

In this study, TF-IDF (Term Frequency-Inverse Document Frequency) is employed to evaluate the significance of words within documents, such as tweets. This technique calculates weights based on how frequently a word appears within a document (term frequency) and how uncommon it is across all documents in the collection (inverse document frequency) (Que et al., 2020). It aims to pinpoint the most relevant words specific to each document. Equation (1) provides the following TF-IDF algorithm:

$$w_{ij} = tf_{ij} \; x \; \ln\left(\frac{D+1}{df_i+1}\right) + 1 \qquad (1)$$

2.4. Support Vector Machine

The Support Vector Machine (SVM) algorithm is an algorithm that is commonly used for classification and is included in the supervised learning category. The working concept of SVM is to find the optimal hyperplane or boundary (Rosyida et al., 2024). The research uses text preprocessing to prepare the data for analysis. Then, a machine learning method called Support Vector Machine (SVM) is used to classify the data into different categories, like positive, negative, or neutral sentiment in this case. To train the SVM model effectively, the data is split into two parts: a larger training set (80%) and a smaller testing set (20%). The training set is used to teach the model how to classify data, while the testing set is used to evaluate how well the trained model performs on unseen data.

Support Vector Machines (SVM) are adept at handling intricate datasets by leveraging kernel functions, which enable them to operate effectively in higher-dimensional spaces. The fundamental premise of SVM revolves around finding the hyperplane that best separates different classes within the data. This hyperplane is not merely any boundary but is strategically positioned to maximize the distance, or margin, between the closest data points of different classes. By doing so, SVM aims to achieve robust classification performance, ensuring that the model can generalize well to unseen data (Moch Arifqi Ramadhan, 2022). This approach is particularly advantageous in scenarios where the data points are not linearly separable in their original feature space, as SVM can map them to a higher-dimensional space where separability becomes feasible. Thus, SVM represents a powerful tool in machine learning for addressing complex classification tasks by effectively delineating boundaries between different categories based on the inherent structure of the data.

3. Results and Discussion

3.1. Data Reprentation

Once trained with labeled data, the SVM model will be tested on unseen data to assess its accuracy in classifying sentiment towards Rohingya refugees on Twitter. This evaluation will involve metrics like accuracy, precision, recall, and F1-score. By analyzing these results, we hope to gain deeper insights into public opinion on this sensitive issue based on Twitter conversations. Additionally, the evaluation will assess the SVM model's effectiveness in sentiment classification for this specific context. The following is a comparative mapping of the amount of positive, negative and neutral sentiment data obtained. The total amount of sentiment data collected for this analysis was 1347 sentiments, divided into 962 sentiments in the negative category, 101 sentiments in the neutra , and 284 sentiments in the positive.

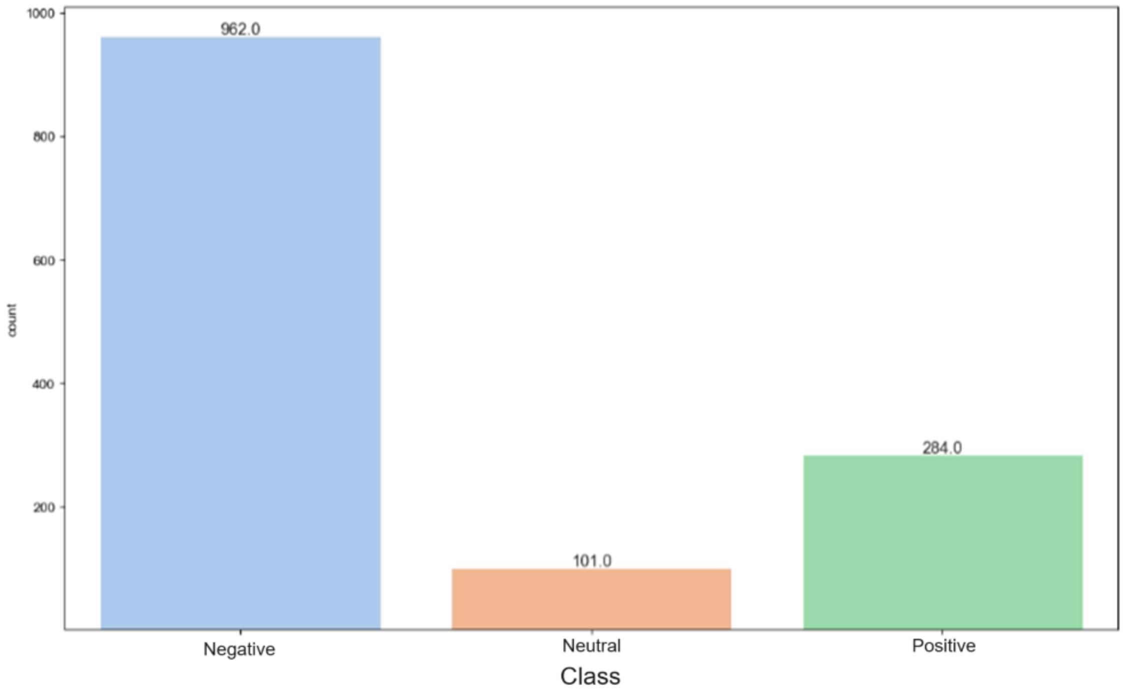

Figure 3. Distribution of Sentiment Data

3.2. TF-IDF Vectorization

The vectorization method allows for assigning numerical values to words in a document or tweet, facilitating pattern processing and search during the classification phase. By analyzing a sample of tweet data and performing calculations based on formula (1), this method employs the word weighting technique known as TF-IDF.

Example of vectorization results:

Table 1. Sentiment Sample

Class	Sentiment for training
Neutral	['usir', 'penduduk', 'aceh']
Negative	['periksa', 'ajak', 'rohingnya', 'ganggu', 'asing']
Negative	['pukul', 'deru']
Negative	['dapat', 'jajah', 'tempat', 'pengusi', 'punya', 'tanda']
Positive	['langsung', 'bebas', 'nyaman', 'penduduk', 'warga', 'prioritas']

After TF-IDF weighting, the results of the weighting carried out can be seen in Table 2.

3.3. Support Vector Machine Process

To develop a sentiment analysis model for categorizing Twitter content related to Rohingya refugees, our initial step involves partitioning our dataset of tweets. This division separates the data into training and testing sets, a common practice where approximately 80% of the dataset is allocated for training the model and the remaining 20% for testing its performance. During the training phase, the model learns from the training data to discern sentiment patterns. Subsequently, the testing data is utilized to assess how accurately the model can classify sentiments in new, unseen tweets. This approach ensures that the sentiment analysis model is both trained effectively and evaluated rigorously before deployment.

Table 2. TF-IDF Result

Term	**TF Vectorization**				
	D1	**D2**	**D3**	**D4**	**D5**
usir	0.206	0	0	0	0
penduduk	0.166	0	0	0	0.166
aceh	0.206	0	0	0	0
periksa	0	0.206	0	0	0
ajak	0	0.206	0	0	0
rohingnya	0	0.166	0.166	0	0
ganggu	0	0.206	0	0	0
asing	0	0.206	0	0	0
kena	0	0	0.206	0	0
dapat	0	0	0	0.206	0
jajah	0	0	0	0.206	0
tempat	0	0	0	0.206	0
pengungsi	0	0	0	0.206	0
punya	0	0	0	0.206	0
tanda	0	0	0	0.206	0
langsung	0	0	0	0	0.206
bebas	0	0	0	0	0.206
nyaman	0	0	0	0	0.206
warga	0	0	0	0	0.412
prioritas	0	0	0	0	0.206

Five examples of tweets with variable X as the input data term and Y as the label are shown in Table 3, where -1 indicates a negative label, 0 indicates a neutral label and 1 indicates a positive label. The SVM algorithm is used to calculate the training data. The system learns from the results of training data which will be used as a reference in the future. Test data findings can be used to find out whether a tweet is included in the positive, neutral or negative category. The following are the stages of the SVM process on sample data.

3.3.1. RBF Kernels

In the classification carried out, the type of kernel used is the RBF kernel because the data entered is non-linear data. The following is the equation used to calculate the RBF kernel value.

$$K\left(x_i, x_j\right) = \exp(-\gamma * |x_i - x_j|) \tag{2}$$

The results of the RBF kernel calculation with a value of gamma = 1 from the sample data are as presented in the table 3.

3.3.2. Hessian Matrix

After establishing the kernel value, the next critical task is to compute the Hessian Matrix. Before proceeding with this calculation, several key parameters must be set, such as αi, C (for regularization), γ (related to the kernel), λ (if applicable), and the maximum number of iterations for the algorithm. These parameters play pivotal roles in shaping the behavior and performance of the model during its training phase (Table 4).

The Hessian matrix calculation step begins by initializing the value α = 0 then carrying out calculations using the following equation

$$D_{ij} = y_i y_j (K\left(x_i, x_j\right) + \lambda^2 \tag{3}$$

The results of the Hessian Matrix calculation for the sample data shown on Table 5.

Table 3. RBF Kernel Result

No.	**1**	**2**	**3**	**..**	**18**	**19**	**20**
1	1.000	0.971	1.000	..	0.919	0.809	0.919
2	0.971	1.000	0.971	..	0.971	0.916	0.971
3	1.000	0.971	1.000	..	0.919	0.809	0.919
4	0.919	0.907	0.919	.	0.919	0.809	0.919
5	0.919	0.907	0.919	..	0.919	0.809	0.919
6	0.907	0.896	0.907	..	0.907	0.799	0.907
7	0.919	0.907	0.919	..	0.919	0.809	0.919
8	0.919	0.907	0.919	..	0.919	0.809	0.919
9	0.919	0.907	0.919	..	0.919	0.809	0.919
10	0.919	0.907	0.919	..	0.919	0.809	0.919
11	0.919	0.907	0.919	..	0.919	0.809	0.919
12	0.919	0.907	0.919	..	0.919	0.809	0.919
13	0.919	0.907	0.919	..	0.919	0.809	0.919
14	0.919	0.907	0.919	..	0.919	0.809	0.919
15	0.919	0.907	0.919	..	0.919	0.809	0.919
16	0.919	0.971	0.919	..	1.000	0.959	1.000
17	0.919	0.971	0.919	..	1.000	0.959	1.000
18	0.919	0.971	0.919	..	1.000	0.959	1.000
19	0.809	0.916	0.809	..	0.959	1.000	0.959
20	0.919	0.971	0.919	..	1.000	0.959	1.000

Table 4. Parameter Value

Parameter	**Value**
Ai	0
C	1
Γ	0.1
λ	0.5
iteration	2

Table 5. Hessian Matrix Result

No.	**1**	**2**	**3**	**..**	**18**	**19**	**20**
1	1.250	1.221	1.250	..	1.169	1.059	1.169
2	1.221	1.250	1.221	..	1.221	1.166	1.221
3	1.250	1.221	1.250	..	1.169	1.059	1.169
4	1.169	1.157	1.169	..	1.169	1.059	1.169
5	1.169	1.157	1.169	..	1.169	1.059	1.169
6	-1.157	-1.146	-1.157	..	-1.157	-1.049	-1.157
7	-1.169	-1.157	-1.169	..	-1.169	-1.059	-1.169
8	1.169	1.157	1.169	..	1.169	1.059	1.169
9	1.169	1.157	1.169	..	1.169	1.059	1.169
10	1.169	1.157	1.169	..	1.169	1.059	1.169
11	-1.169	-1.157	-1.169	..	-1.169	-1.059	-1.169
12	1.169	1.157	1.169	..	1.169	1.059	1.169
13	-1.169	-1.157	-1.169	..	-1.169	-1.059	-1.169
14	1.169	1.157	1.169	..	1.169	1.059	1.169
15	1.169	1.157	1.169	..	1.169	1.059	1.169
16	1.169	1.221	1.169	..	1.250	1.209	1.250
17	1.169	1.221	1.169	..	1.250	1.209	1.250
18	1.169	1.221	1.169	..	1.250	1.209	1.250
19	1.059	1.166	1.059	..	1.209	1.250	1.209
20	1.169	1.221	1.169	..	1.250	1.209	1.250

3.3.3. *Sequential Training SVM*

Sequential training SVM is the process of updating alpha values. This process includes calculating error values, calculating delta alpha values and calculating new alpha. The sequential training process will repeat until the specified target error limit is reached. In this sample, the results of sequential training with 2 iterations shown on Table 6.

Table 6. Alpha Value

No.	α
1	0.257893
2	0.255541
3	0.257893
4	0.256946
5	0.256946
6	0.256688
7	0.256946
8	0.256946
9	0.258367
10	0.343225
11	0.259225
12	0.259225
13	0.259225
14	0.343225
15	0.343225
16	0.25599
17	0.25599
18	0.25599
19	0.252738
20	0.25599

3.3.4. *Bias Calculation*

The bias value in Support Vector Machine (SVM) is a term that refers to certain parameters in the SVM model which function to shift the hyperplane – the space separating classes – to the optimal position. The equation used to find the bias value is as follows.

$$b = \frac{1}{2}[\sum_{i=1} \alpha_i y_i K(x_i x^+) + \sum_{i=1} \alpha_i y_i K(x_i x^-)] \tag{4}$$

By using the equation 4, a bias value of -0.895 is obtained. This value will be used in the SVM testing process.

3.3.5. *Testing Proccess*

For the testing process, the TF-IDF weight calculation step is carried out on the test data. After obtaining the TF-IDF vector value from the test data, the next step is to calculate the Euclidean square distance between the test vector and each support vector. The f sample of the test data used shown on Table 7.

Table 7. Testing sentiment

Sentiment for testing
['datang', 'masalah', 'periksa', 'kena', 'rohingnya', 'ajak', 'kena', 'sebar', 'gelap', 'periksa']

By calculating the vector value in the test used, the value of the function f(x) will be obtained. Where the function (x) is the determinant of the test data class. The following is the equation used for test data.

$$f(x) = sign(\sum \mathrm{K}(V_{uji}, V_i) + b) \tag{5}$$

After carrying out the testing process on test data 1, it is known that the classification function gets a value of -0.76 so that test data 1 is classified as class -1 category where class -1 is the negative class.

3.4. Evaluation

After completing the testing process on the Support Vector Machine algorithm, the results will be obtained. Where the results in question are labels from test data obtained from the model during the training process. The results of the classification of test data in the form of sentiment classes obtained from the program will be compared with the actual class data so that the accuracy, precision, recall and f1-score values of the model used on the dataset will be known. During the evaluation phase, the effectiveness of the model applied to the training dataset is assessed. Utilizing a classification report helps determine metrics like precision, recall, F1-score, and accuracy, providing insights into the model's performance across distinct datasets.

Support Vector Machines (SVM) make predictions by examining and leveraging the patterns ingrained within the training data. This data typically encompasses classes representing positive and negative sentiments, enabling SVM to discern unique characteristics associated with words in each sentiment class. Additionally, visual representations like word clouds offer intuitive insights into the prevalent terms associated with positive, neutral, and negative sentiments within the dataset.

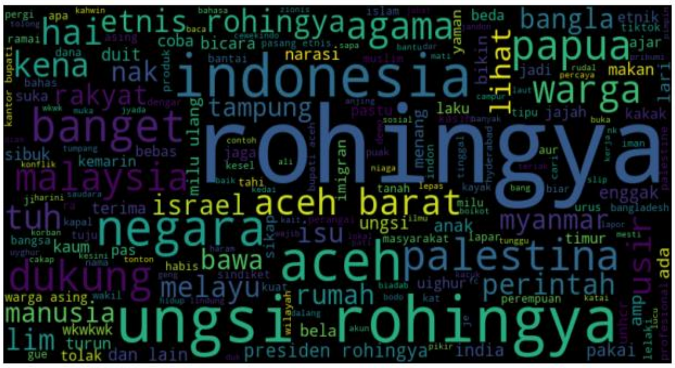

Figure 4. Wordcloud for “Positive”

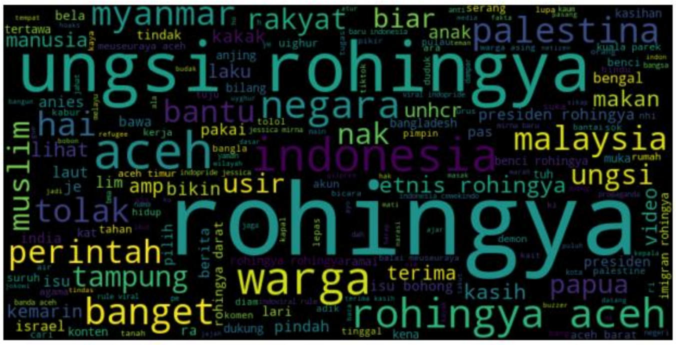

Figure 5. Wordcloud for “Negative”

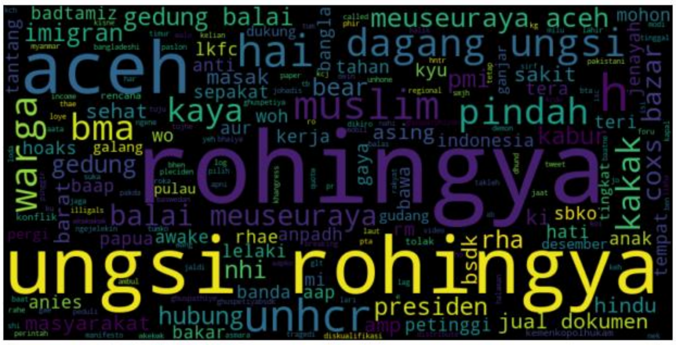

Figure 6. *Wordcloud* for “Neutral”

Upon completion of the testing phase using the Support Vector Machine (SVM) algorithm, the results are obtained by predicting labels *for* the test data, which were unseen during training. These predictions reflect the sentiment classes assigned by the model. To assess the model's performance, these predicted sentiment labels are compared against the actual labels in the test dataset. This comparison allows for the calculation of metrics such as accuracy, precision, recall, and F1-score, providing insights into how well the SVM model generalizes to new data and accurately classifies sentiments based on the patterns learned from the training dataset.

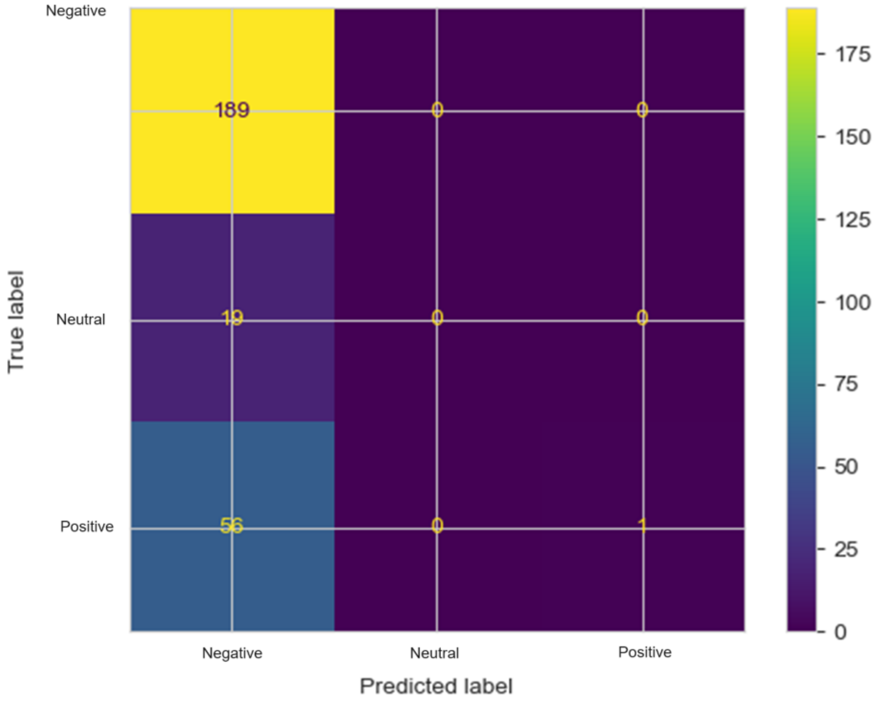

Figure 7 Confussion Matrix

From Figure 7, the values for accuracy, precision, recall and f1-score can be calculated using the equation:

$$Accuracy = \frac{189+0+1}{189+0+0+19+0+0+56+0+1} \text{ x } 100\% = 72\%$$

$$Precision = \frac{1}{1+0} \text{ x } 100\% = 100\%$$

$$Recall = \frac{1}{1+156} \text{ x } 100\% = 2\%$$

$$F1\text{-}Score = \frac{2 \, x \, 100 \, x \, 2}{100+2} \text{ x } 100\% = 3\%$$

4. Conclusion

It is known that of the 1347 data obtained regarding sentiment, 21.08% of the sentiment was positive, 71.41% of the sentiment was negative and 7.49% of the sentiment was neutral. 2. In the analysis carried out, the ratio between training data and test data is 8:2. 1056 of them are training data while 265 are test data. The results of classifying sentiment towards Rohingya immigrants using the support vector machine algorithm which corresponds to the actual data amounts to 190 data from a total of 265 test data. The level of accuracy resulting from classifying sentiment towards Rohingya immigrants using the Support Vector Machine (SVM) algorithm is accuracy of 72%, precision of 100%, recall of 2%, and f1-score of 3%

References

Ariansyah, A., & Kusmira, M. (2021). Analisis Sentimen Pengaruh Pembelajaran Daring Terhadap Motivasi Belajar Di Masa Pandemi Menggunakan Naive Bayes Dan Svm. *Faktor Exacta*, *14*(3), 100.

https://doi.org/10.30998/faktorexacta.v14i3.10325

Dhea Nada, T. gunawan. (2021). Peran Pemerintah Dalam Menangani Pengungsi Rohingya Di Kota Lhokseumawe (Studi pada Kantor Imigrasi Kelas II TPI kota Lhokseumawe) Dhea. *Kajian Administrasi Negara: Riset Dan Pengabdian*, *01*(01), 45–50.

Giovani, A. P., Haryanti, T., & Kurniawati, L. (2020). Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi. 14(2), 116–124.

Hermawan, L., & Bellaniar Ismiati, M. (2020). Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval. *Jurnal Transformatika*, *17*(2), 188. https://doi.org/10.26623/transformatika.v17i2.1705

Ismail, A. R., Bagus, R., & Hakim, F. (2023). Implementasi Lexicon Based Untuk Analisis Sentimen Dalam Mengetahui Trend Wisata Pantai Di DI Yogyakarta Berdasarkan Data Twitter. 1(1), 37–46.

Moch Arifqi Ramadhan, R. A. (2022). *Klasifikasi Text Spam Menggunakan Metode Support Vector Machine Dan Naive Bayes* (R. M. Awangga (ed.)). Penerbit Buku Pedia.

Que, V. K. S., Iriani, A., & Purnomo, H. D. (2020). Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *9*(2), 162–170. https://doi.org/10.22146/jnteti.v9i2.102

Rosyida, T., Putro, H. P., Wahyono, H., Teknik, F., & Krisnadwipayana, U. (2024). *Opini Dari Twitter Menggunakan Naïve Bayes Dan Svm*. *26*(1).

Sriani, Suhardi, I. F. G. (2023). Analisis Sentimen Kebijakan Pemberian Subsidi Motor Listrik Menggunakan Metode Support Vector Machine. 13(3), 511–517.

Yulian Azhari, W. (2022). Pencegahan Potensi Konflik Antara Pengungsi Rohingya Dan Masyarakat Lokal Indonesia. *Pengabdian Mandiri*, *20*(1), 105–123.

Zulkarnain, I. K. (2020). Bersama Untuk Kemanusiaan: Penanganan Lintas Sektor Terhadap Masalah Pengungsi Rohingya Di Aceh 2015. *Jurnal HAM*, *11*(April), 475–478.