

The performance of Naïve Bayes, support vector machine, and logistic regression on Indonesia immigration sentiment analysis

Priati Assiroj¹, Asep Kurnia², Sirojul Alam³

¹Department of Technology Management, Polytechnic of Immigration, Ministry of Law and Human Rights, Depok, Indonesia

²The Expert for Strengthening Bureaucratic Reform, Ministry of Law and Human Rights, Jakarta, Indonesia

³Data Governance Officer, Department of Digital Governance and Data Protection, SBU Digital, Perum Peruri, Jakarta, Indonesia

Article Info

Article history:

Received Jan 4, 2023

Revised May 27, 2023

Accepted Jun 5, 2023

Keywords:

Immigration

Logistic regression

Naïve Bayes

Sentiment analysis

Support vector machine

ABSTRACT

In recent years various attempts have been made to automatically mine opinions and sentiments from natural language in online networking messages, news, and product review businesses. Sentiment analysis is needed as an effort to improve service performance in the organization. In this paper, we have explored the polarization of positive and negative sentiments using Twitter user reviews. Sentiment analysis is carried out using the Naïve Bayes (NB), support vector machine (SVM), and logistic regression (LR) model then compares the results of these three models. The results of the experiment showed that the accuracy of LR was better than SVM and NB, namely 77%, 76%, and 70%.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Priati Assiroj

Department of Technology Management, Polytechnic of Immigration, Ministry of Law and Human Rights
Gandul Raya Street, No. 4, Depok, Jawa Barat, Indonesia

Email: priati.assiroj@poltekim.ac.id

1. INTRODUCTION

Communication or interaction is a basic need of human life as a social being. Human communication needs to aim to observe the environment so that humans can survive and adapt to the environment [1]. The communication process requires an instrument as a connector for information. Today, mass media communication has become a vital element in leading and changing human opinion. In the past, media were described as newspapers, magazines, radio, films, and television. However, with current technological advances, mass media is associated with the use of the internet in interactive mass media. Today's communication technology plays a role in sustaining the shift from conventional media to modern media [2].

In recent years, opinion and sentiment mining has been automated using online messages, such as Twitter threads, news, and product reviews. This research utilizes Indonesian language threads or tweets that talk about immigration services in Indonesia. Sentiment analysis is a method whose implementation utilizes data in the form of text by evaluating and identifying feelings and opinions, both positive and negative [3]. Twitter users can provide objective opinions on various topics or issues [4]. One of the earliest studies on Twitter sentiment analysis was conducted by [5], who considers problems as two classes of classification and characterizes tweets as positive or negative. Researchers conducted sentiment analysis on reviews using the Naïve Bayes (NB), support vector machine (SVM), and logistic regression (LR) model. SVM has been widely used for classification and regression. Theoretically and practically, this algorithm has proven its achievements in various domains [6].

A study has proven the effectiveness of SVM for processing Arabic tweet data [7] with satisfactory results. Another study was conducted by [8] using a dataset for sentiment analysis with NB and SVM. Alves *et al.* [8] also presents the method used to classify the polarity of tweet sentiment by considering spatial and temporal information. Kharde and Sonawane [9] prove the effectiveness of SVM and NB in sentiment analysis data through results and tables collected using datasets from Twitter. Troussas *et al.* [10] also used NB to classify Facebook status and the results were compared with the rocchio classifier and perceptron classifier. The results showed that NB has a better precision level of 77%, compared to the last two classifiers. The NB method can also be combined with the feature selection method of the genetic algorithm. Muthia [11] have been done by using hotel review data. As a result, the original NB method obtained an accuracy rate of 78.5% and after being combined with feature selection from the genetic algorithm the accuracy rate became 83%. Martiti and Juliane [12] also found that the accuracy of NB used in the sentiment analysis application they made was 86.6%.

The Directorate General of Immigration (Ditjenim) is a government agency that provides public services in the field of immigration, is also uses Twitter as a communication instrument. Not only that, but the technical service units (UPT) spread across Indonesia also have accounts Twitter in order to realize good governance. Twitter as a medium of relations must be able to accommodate providing two-way communication facilities between government administrators and the community. Departing from the number of internet and Twitter users in Indonesia, of course, this is a big potential for the use of big data by the Ditjenim and other UPT immigration. Big data is a system that unites the real world, humans, and the virtual world (social media) [13]. Sourced from the data record of user conversations on Twitter, if the processing is carried out, of course, it will produce a certain pattern or characteristic of information. This can be used in formulating strategies, research, and market (community) responses to an immigration service or product.

Ditjenim has great potential in processing and utilizing big data, the article is that there are 126 UPT spread throughout Indonesia that provide immigration services in the form of issuing passports, visas, and residence permits, of course, it requires professional data handling related to public complaints. Big data contained in social media is an unstructured form of data [14] and has no pattern or schema [15]. Text mining is used to process unstructured and patterned text data [16].

Complaints in the form of tweets when compiled and analyzed can provide important information that has characteristics or patterns to certain trends or issues related to immigration services. The data can be used as a means of conducting sentiment analysis on various immigration policies, so that it can provide feedback for institutions to improve in the future. In addition to SVM, this study also uses a highly probabilistic NB classifier model. This method is simple but powerful because it has a high value of accuracy and performance in classification [17]. The NB method can be used to polarize sentiment into positive and negative categories [10]. Likewise, the LR model which is a supervised learning algorithm and can be used to classify text data also applied to this research so that the performance of the three algorithms can be compared. The comparison is done by looking at the confusion matrix and the area under curve (AUC) value on the receiver operating characteristic (ROC) curve of each algorithm. Classification quality is seen from the AUC value which is divided into several groups [18], 0.90-1.00 for very good classification, 0.80-0.90 for good classification, 0.70-0.80 for adequate classification, 0.60-0.70 for poor classification, and 0.50-0.60 for the wrong classification.

2. METHOD

Singh and Dubey [19] have conducted a literature study on sentiment analysis and opinion research on social issues. The selected newspapers have extracted data from the website. They argued that different types of classification techniques when combined can produce better results. Akbani *et al.* [20] classified the sentiment of tweets in Arabic using NB, decision trees, and SVM. In this study, the framework for classifying Arabic tweets consists of several subtasks such as term frequency inverse document frequency (TF-IDF) and Arabic mood.

In addition, three information-seeking metrics were used for performance evaluation: precision, recall, and F-score. Shoukry and Rafea [21] focuses on the effect of preprocessing features in the sentiment classification process. Ahmad and Aftab [22] analyzed the performance of the SVM for polarity detection from textual data. Davidov *et al.* [23] using an SVM classifier prepared with eleven features for transient stability assessment (TSA). With the ability to survey public opinion (sentiment) on a subject, data can be collected and analyzed from social media such as Twitter in real-time [24]. Social media sentiment analysis has been widely used in the topics studied [25]–[28]. Community or customer understanding of the perception of a product is very useful for business marketing strategies [29], [30].

The workflow of this paper is shown in Figure 1. After we collect data from Twitter related to Indonesian immigration and its passport services, we conducted a data understanding of the process. In this

step, we try to understand the data we have and identify potential problems that exist in the dataset. The next stage is data preparation. At this stage, we perform several steps. We perform cleaning, case-folding, tokenizing, filtering, and stemming until we get data which we call preprocessed data before sentiment polarization is carried out. Then after knowing the polarization of sentiment, we did modeling using the three models we mentioned earlier, NB, SVM, and LR, we evaluated the results with a confusion matrix and ROC curve to find out which model has the best performance.

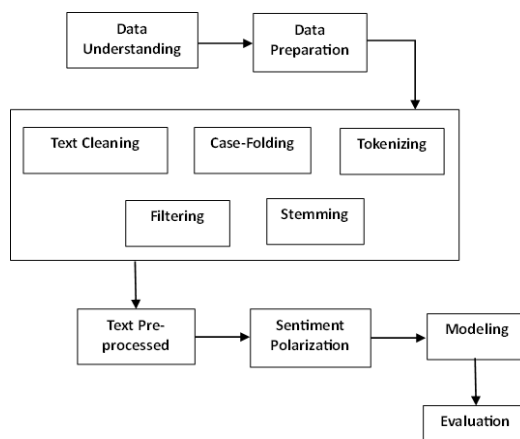


Figure 1. Research workflow

2.1. Data understanding

Data understanding is conducted to understand the dataset. We first import it into our jupyter notebook workspace and read it as shown in Figure 2. Then we use the pandas library to see the details of the dataset. This is the original data that we got and we call it raw data.

	Text
0	#Repost @ditjen_imigrasi\n...\nSahabat Mido, p...
1	Reposted from @ditjen_imigrasi Sahabat Mido, p...
2	@ditjen_imigrasi Saya sudah mengisi data pada ...
3	@sashaasays @ditjen_imigrasi Bukannya emg yg b...
4	@tempodotco "Resmi" berarti masuk melalui imig...
5	Imigrasi Ngurah Rai Bali Tangkap WNA Nigeria y...
6	@Kemenkumham_RI @Kumham_Sulsel Kemenkumham Sul...
7	Imigrasi Blitar ungkap peningkatan permohonan ...
8	@Kemenkumham_RI @Kumham_Sulsel Kemenkumham Sul...
9	RT @elrinyuliana: @saidiman sama pilih beberap...

Figure 2. Raw data

Data understanding is the first step in data analysis. The data is checked so that it will be known what problems exist in the data. In addition, a summary and identification of potential problems can also be made. This stage must be done carefully because it will determine the results in the next stage. The summary is used as a reference to ensure the data distribution is appropriate, or it can also be used to find out the deviations that must be handled in data preparation. Problems like null, outliers, and bad data density can be fixed in data preparation [31]. After understanding the raw data, then we check for duplicated tweets or if there are any duplicates. The duplication of the tweets may happen among the huge of data. The result is shown in Figure 3.

According to Figure 3, we can see that there are 4,809 duplicated tweets. We use the duplicate function from the pandas library. After we know the duplicate, we delete it all in the data preparation stage. Then at this data understanding stage, we also check how much data we have and check if there are empty data or null. Null data is a noise that can be a problem in the analysis. For the best result, we should make our data clean. We found that in our dataset, there are 10,000 tweet rows and there are no blank rows or data. We

did not find the null data, so we can use this data for analysis. If we found the null data, we must delete it and make sure again that there are no null data.

	Text
991	RT @SeruniPuspaAlam: Negara asing akan pikir-p...
994	RT @SeruniPuspaAlam: Negara asing akan pikir-p...
999	sore ini ya.
1000	@kompascom @tempodotco @repu...
1002	sore ini ya.
...	...
9978	1. Kakanwil Kemenkumham Sumsel Harun Sulianto ...
9980	Masih inget kejadiannya ya @Dennysiregar7?InYa...
9981	Bokep Indo Doyan Sperma'nInBokep Indo Viral Sk...
9992	1. Kakanwil Kemenkumham Sumsel Harun Sulianto ...
9998	Bokep Indo Doyan Sperma'nInBokep Indo Viral Sk...

4809 rows × 1 columns

Figure 3. Duplicate tweets

2.2. Data preparation

This stage is done to fix the problems that existed in the previous stage. This stage is also a determinant of the suitability of the data to the algorithm to be used because ideally this stage is reviewed repeatedly when problems occur in the modeling until an appropriate one is found. Activities include data selection, transformation, and data cleansing so that the data is finally ready for modeling [32]. One of the actions we take is to delete data that has duplicates. We found that we have 4,809 rows, as shown in Figure 3, have been deleted, then leaving 5,191 data to be used for this research. The deleted data is data that is duplicated as we said in the previous stage. We now have data that do not have null and duplicate.

2.2.1. Text cleansing

At this stage, we perform data cleansing after the duplicate data is deleted. We use the re library which is already available in the python programming language. We also use the natural language toolkit (NLTK) library to tokenize and remove stopwords, then we use the Sastrawi library for text processing in Indonesian. The text cleaning function cleans text from unnecessary characters such as excessive spaces, symbols, numbers, links, hashtags, and mentions. Then we continue by creating the functions needed for the next process, namely case-folding, tokenizing, filtering, and stemming. These functions are needed to do text analysis.

2.2.2. Case-folding

Case-folding refers to the process of converting text to a standard lowercase form and removing any distinctions between uppercase and lowercase letters. It means the process will convert all text to lowercase. This process uses the lower function in the string library. The result is lowered text data. Then after we gain the case-folded data, we do the tokenization process.

2.2.3. Tokenizing

Tokenizing is the process to encode words. The words in the text column will be grouped word by word into an array of strings. This process uses the word tokenize function in the NLTK library. The sentences will be separated into words. The result is text data that consist of words from separated sentences.

2.2.4. Filtering

In this process, we deliberately filter out only Indonesian tweets that will be used for research. The filtering process is conducted to get tweets that are only in Indonesian. This process utilizes the corpus and stopwords functions provided by the NLTK library. The filtered data is then selected for the next process, which is word stemming.

2.2.5. Stemming

The final step in text processing is to convert all the words in the existing text to their basic form. For example, the word “reading” will be changed to “read”. This process takes quite a long time depending on the number of datasets. After all the text processing functions are created, then we apply these functions one by one so that we get the data as shown in Figure 4. The data that we have cleaned is then saved with the name 'cleaned_data.csv', and in the next process, we use this dataset. Figure 4 is the data that is ready for the next process. But as seen in the column 'text_preprocessed' there are still some special characters such as quotes, which we should remove. The next process is to remove some special characters that still exist as shown in Figure 5.

	text_clean	text_preprocessed
0	imigrasisahabat mido pengen ubah nama panggilan...	['imigrasisahabat', 'mido', 'ken', 'ubah', 'na...
1	reposted from imigrasi sahabat mido pengen uba...	['reposted', 'from', 'imigrasi', 'sahabat', 'm...
2	imigrasi saya sudah mengisi data pada app m pa...	['imigrasi', 'isi', 'data', 'app', 'm', 'paspo...
3	imigrasi bukannya emg yg buat bali udh bisa sa...	['imigrasi', 'emg', 'yg', 'bal', 'udh', 'sa', '...
4	resmi berarti masuk melalui imigrasi mana nih	['resmi', 'masuk', 'imigrasi', 'nih']
...
5186	optimalkan nilai ikpa kantor imigrasi kelas i ...	['optimal', 'nilai', 'ikpa', 'kantor', 'imigra...
5187	halo sahabat mido untuk menjawab kebutuhan pas...	['halo', 'sahabat', 'mido', 'butuh', 'paspor', '...
5188	kayanya hari ini akan lembur sampe sahur lagi ...	['kaya', 'lembur', 'sampe', 'sahur', 'gapapa', '...
5189	jakarta—kantor imigrasi kelas i khusus non tpi...	['jakarta kantor', 'imigrasi', 'kelas', 'i', '...
5190	halo sahabat midoterimakasih banyak atas apres...	['halo', 'sahabat', 'midoterimakasih', 'apres...

5191 rows x 2 columns

Figure 4. Results application functions that have been made

	text_clean	text_preprocessed
0	imigrasisahabat mido pengen ubah nama panggilan...	[imigrasisahabat, mido, ken, ubah, nama, pangg...
1	reposted from imigrasi sahabat mido pengen uba...	[reposted, from, imigrasi, sahabat, mido, ken, ...
2	imigrasi saya sudah mengisi data pada app m pa...	[imigrasi, isi, data, app, m, paspor, mentok, ...
3	imigrasi bukannya emg yg buat bali udh bisa sa...	[imigrasi, emg, yg, bal, udh, sa, kalo, cgk, e...
4	resmi berarti masuk melalui imigrasi mana nih	[resmi, masuk, imigrasi, nih]
...
5186	optimalkan nilai ikpa kantor imigrasi kelas i ...	[optimal, nilai, ikpa, kantor, imigrasi, kelas...
5187	halo sahabat mido untuk menjawab kebutuhan pas...	[halo, sahabat, mido, butuh, paspor, sahabat, ...
5188	kayanya hari ini akan lembur sampe sahur lagi ...	[kaya, lembur, sampe, sahur, gapapa, ri, imigr...
5189	jakarta—kantor imigrasi kelas i khusus non tpi...	[jakarta, kantor, imigrasi, kelas, i, khusus, ...
5190	halo sahabat midoterimakasih banyak atas apres...	[halo, sahabat, midoterimakasih, apresiasi, pe...

5191 rows x 2 columns

Figure 5. Pre-processed text

2.3. Text pre-processed

Text preprocessing refers to a set of techniques and steps applied to raw textual data before it is used for further analysis or natural language processing tasks. It involves transforming the text into a format that is more suitable and efficient for subsequent processing. Text preprocessing typically includes tasks such as removing punctuation, converting to lowercase, tokenization (splitting text into individual words or tokens), removing stop words (commonly used words that do not carry significant meaning), stemming or lemmatization (reducing words to their base or root form), and handling special characters or encoding issues. The goal of text preprocessing is to clean and standardize the text data, making it easier to analyze and derive meaningful insights. Figure 5 shows that the data in the text_preprocessed column is ready for sentiment analysis. The process is done by making sentiment polarity based on the Indonesian language lexicon dictionary. This lexicon dictionary is generally available on the internet and can be downloaded by anyone who needs it. We use two lexicon dictionaries, namely positive and negative.

2.4. Sentiment polarization

Sentiment polarity is an expression that defines the sentimental aspect of an opinion. In text data, the results of sentiment analysis can be determined for each entity in a sentence, document, or sentence. Mood polarity can be defined as positive, negative, or neutral [33]. The fundamental task of sentiment analysis is to classify whether the opinion expressed in a document, sentence, or entity attribute/aspect is positive, negative, or neutral [34]. The polarity of sentiment for an item determines the orientation of the expressed sentiment; determines whether the text expresses the user's positive, negative, or neutral feelings toward the entity in question [35].

Figure 6 shows the defined function in python to align data with the Indonesian lexicon dictionary. Then we define a function for sentiment analysis in Indonesian whose data polarity is based on the lexicon dictionary and we group it into positive if the value in the lexicon dictionary is more than 0, negative if the value in the lexicon dictionary is less than 0, and neutral if the value in the lexicon dictionary is the same with 0 and the result is shown in Figure 7. The result shows that negative sentiment is 3,655, positive sentiment is 973, and neutral is 563. We also provide a sample of the polarized data that shows a sample of these three types of sentiments. Figure 7 is an illustration of the polarization obtained from the data regarding the existing lexicon dictionary.

```

1 def sentiment_analysis_lexicon_indonesia(text):
2     score = 0
3     for word in text:
4         if(word in lexicon_positive):
5             score = score + lexicon_positive[word]
6         for word in text:
7             if(word in lexicon_negative):
8                 score = score + lexicon_negative[word]
9
10    polarity = ''
11    if(score > 0):
12        polarity = 'positive'
13    elif (score < 0):
14        polarity = 'negative'
15    else:
16        polarity = 'neutral'
17    return score, polarity
18

```

```

1 #hasil polarisasi sentimen
2
3 results = tweets['text_preprocessed'].apply(sentiment_analysis_lexicon_indonesia)
4 results = list(zip(*results))
5
6 tweets['polarity_score'] = results[0]
7 tweets['polarity'] = results[1]
8 print(tweets['polarity'].value_counts())
9
10 tweets

```

negative	3655
positive	973
neutral	563

Name: polarity, dtype: int64

Figure 6. Sentiment polarity

Figure 7. Polarization result

We have known the data with its polarity then we visualized the data to see it deeply as shown in Figure 8. Figure 8 shows that as many as 70.4% of Twitter users have negative sentiments about Indonesian immigration, 18.7% have positive sentiments, and 10.8% are neutral. From the Figure 8, it can be seen that the sentiment data for the modeling process has not been balanced. There are around 2,682 data that need to be manipulated so that the sentiment dataset is balanced and can be used for modeling. By using the NumPy library we do the dataset balancing process before modeling. Data balancing is conducted by including positive and negative polarizations. Then we visualize the data that we have balanced. Visualization is conducted to see in detail the percentage of each data. The visualization results can be seen in Figure 9. Figure 9 shows that the existing sentiment data is balanced. The positive sentiment is 50% and the negative sentiment is also 50%. So that it is ready for the next step, which is modeling.

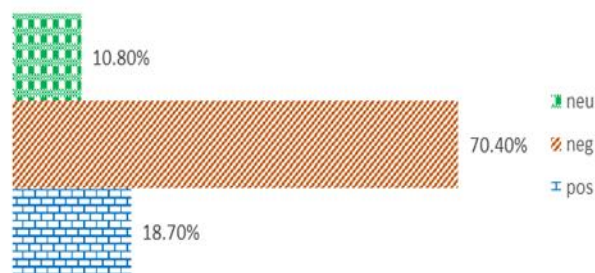


Figure 8. Polarity

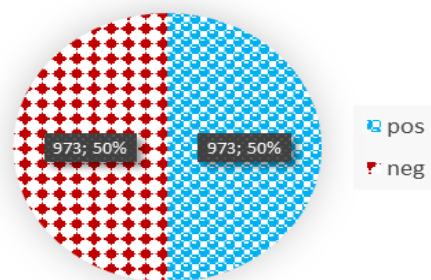


Figure 9. Balanced visualization

3. RESULTS AND DISCUSSION

3.1. Modeling

As previously explained, this study uses a NB algorithm, SVM, and LR for sentiment analysis with 70% data division as training data and 30% as testing data. The results of each model are shown in Figure 10

for NB, Figure 11 for the SVM, and Figure 12 for the LR model. Figure 10 shows that the accuracy of the NB model is 70%, while the precision values are 85% and 64%, respectively. Figure 11 shows that the accuracy of the SVM model is 76% higher than the NB model, while the precision values are 82% and 72%, respectively. Figure 12 shows that the accuracy of the LR model is 77% higher than the NB model and SVM, while the precision values are 81% and 74%, respectively.

```
#Bayesian Model
BNBmodel = BernoulliNB()
BNBmodel.fit(X_train, y_train)
model_Evaluate(BNBmodel)
y_pred1 = BNBmodel.predict(X_test)
```

	precision	recall	f1-score	support
0	0.85	0.47	0.60	62
1	0.64	0.92	0.76	64
accuracy			0.70	126
macro avg	0.75	0.69	0.68	126
weighted avg	0.75	0.70	0.68	126

Figure 10. Result of NB

```
#SVM Model
from sklearn.svm import LinearSVC
SVCmodel = LinearSVC()
SVCmodel.fit(X_train, y_train)
model_Evaluate(SVCmodel)
y_pred2 = SVCmodel.predict(X_test)
```

	precision	recall	f1-score	support
0	0.82	0.66	0.73	62
1	0.72	0.86	0.79	64
accuracy			0.76	126
macro avg	0.77	0.76	0.76	126
weighted avg	0.77	0.76	0.76	126

Figure 11. Result of SVM

```
#Logistic Regression Model
LRmodel = LogisticRegression(C = 2, max_iter = 1000, n_jobs=-1)
LRmodel.fit(X_train, y_train)
model_Evaluate(LRmodel)
y_pred3 = LRmodel.predict(X_test)
```

	precision	recall	f1-score	support
0	0.81	0.69	0.75	62
1	0.74	0.84	0.79	64
accuracy			0.77	126
macro avg	0.78	0.77	0.77	126
weighted avg	0.77	0.77	0.77	126

Figure 12. Result of LR

3.2. Evaluation

To evaluate the used model, we use the confusion matrix that we get for each model along with its ROC curve. This is the confusion matrix of each model. Figure 13 is a confusion matrix from NB, Figure 14 is a SVM, and Figure 15 is a confusion matrix from LR. From the NB confusion matrix in Figure 13, we can see that the prediction result for the true positive is 46.83%, the true negative is 23.02%, the false positive is 26.19%, and the false negative is 3.97%. In Figure 14, SVM confusion matrix, the prediction result for true positive is 43.65%, for true negative is 32.54%, for false positive is 16.67%, and for false negative is 7.14%. In Figure 15, LR confusion matrix, the prediction result for true positive is 42.86%, for true negative is 34.13%, for false positive is 15.08%, and for false negative is 7.94%.

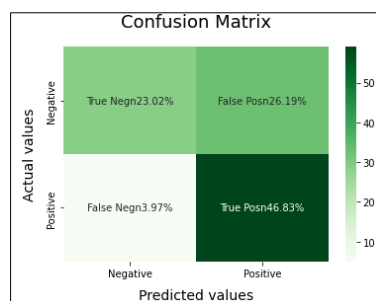


Figure 13. NB

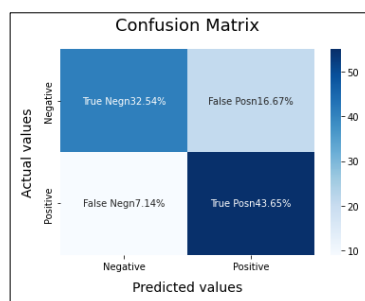


Figure 14. SVM

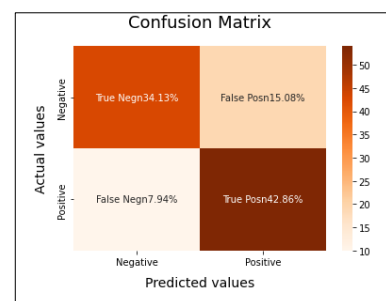


Figure 15. LR

Hereafter we also obtain a visualization of the ROC curve from the NB model. The ROC curve from the NB model can be seen in Figure 16, for SVM model can be seen in Figure 17, and for LR model in

Figure 18. The ROC curve shows the performance of classification models at all classification thresholds. From this curve, we know the AUC graphs. This graph is located between the true positive rate with sensitivity on y-axis and the false positive rate with specificity on x-axis [36]. It seems a bid among this both sensitivity and specificity. From the curve, in Figure 16 we can see that the AUC value of the NB model is 0.69. AUC provides an aggregate measure of performance across all possible classification thresholds. Figure 17 shows that the AUC value of the SVM model is 0.76, which is higher than the NB model. Figure 18 provides information that the AUC value of the LR model is 0.77, higher than the NB model and SVM.

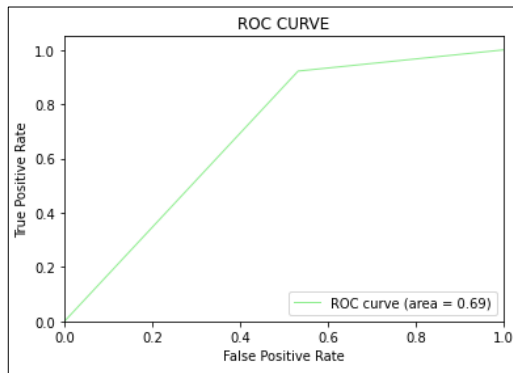


Figure 16. ROC of NB

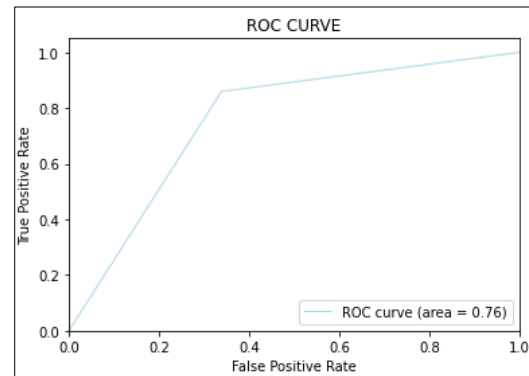


Figure 17. ROC of SVM

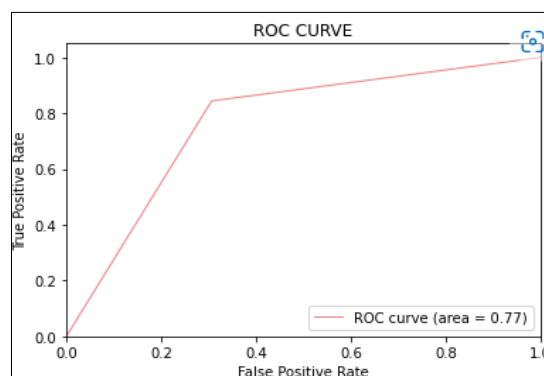


Figure 18. ROC of LR

4. CONCLUSION

From testing using three models, NB, SVM, and LR, we can conclude that accuracy: in terms of model accuracy, LR outperforms SVM which in turn outperforms Bernoulli NB. LR has 77%, SVM 76%, and Bayesian 70%. F1-score: the F1-scores are: i) for class 0: Bernoulli NB (accuracy=0.60)<SVM (accuracy=0.73)<LR (accuracy=0.75) and ii) for class 1: Bernoulli NB (accuracy=0.76)<SVM and LR (accuracy=0.79). AUC score: NB model has 0.69 AUC score, SVM model has 0.76 AUC score, and LR model has 0.77 AUC score. We conclude that the best model for the given dataset is LR. In our problem, LR follows Occam's Razor principle, which defines that for a given problem statement, if the data has no assumptions, the simplest model will work best. Since our data set has no assumptions and LR is a simple model, the concept applies to the above dataset.

ACKNOWLEDGEMENTS

We would like to express our sincere appreciation to Politeknik Imigrasi for their generous support and funding that made this research project possible. Their financial assistance has played a crucial role in facilitating the successful completion of this study.





REFERENCES

- [1] H. D. Lasswell, "The Structure and Function of Communication in Society," *İletişim kuram ve araştırma dergisi*, vol. 24, pp. 215-228, 2007.
- [2] J. Straubhaar, R. LaRose, and L. Davenport, *Media now: Understanding media, culture, and technology*. Boston, MA: Cengage Learning, 2015.
- [3] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," *Computational Linguistics*, vol. 35, no. 3, pp. 399-433, 2009, doi: 10.1162/coli.08-012-R1-06-90.
- [4] L. F. S. Coletta, N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, "Combining Classification and Clustering for Tweet Sentiment Analysis," in *2014 Brazilian Conference on Intelligent Systems*, IEEE, 2014, pp. 210-215, doi: 10.1109/BRACIS.2014.46.
- [5] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," US, 2009.
- [6] R. Batuwita and V. Palade, "Class Imbalance Learning Methods for Support Vector Machines," in *Imbalanced Learning*, Hoboken, USA: John Wiley & Sons, 2013, pp. 83-99, doi: 10.1002/9781118646106.ch5.
- [7] M. M. Altaawier and S. Tiun, "Comparison of machine learning approaches on Arabic twitter sentiment analysis," *Int J Adv Sci Eng Inf Technol*, vol. 6, no. 6, pp. 1067-1073, 2016, doi: 10.18517/ijaseit.6.6.1456.
- [8] A. L. F. Alves, C. de S. Baptista, A. A. Firmino, M. G. de Oliveira, and A. C. de Paiva, "A Spatial and Temporal Sentiment Analysis Approach Applied to Twitter Microtexts," *Journal of Information and Data Management*, vol. 6, no. 2, pp. 118-129, 2015.
- [9] V. A. Kharde and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," *Int J Comput Appl*, vol. 139, no. 11, pp. 5-15, 2016, doi: 10.5120/ijca2016908625.
- [10] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning," in *IISA 2013*, IEEE, 2013, pp. 1-6, doi: 10.1109/IISA.2013.6623713.
- [11] D. A. Muthia, "Sentiment Analysis of Hotel Review Using Naïve Bayes Algorithm and Integration of Information Gain and Genetic Algorithm As Feature Selection Methods," in *International Seminar on Scientific Issues and Trends (ISSIT)*, 2014, pp. 25-30, doi: 10.1109/ICIC47613.2019.8985946.
- [12] Martiti and C. Juliane, "Implementation of Naive Bayes Algorithm on Sentiment Analysis Application," in *Proceedings of the 2nd International Seminar of Science and Applied Technology (ISSAT 2021)*, 2021, pp. 193-200, doi: 10.2991/aer.k.211106.030.
- [13] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: a survey on big data," *Inf Sci (N Y)*, vol. 275, pp. 314-347, 2014.
- [14] J. E. Wieringa, "Unstructured data: Can its power be unleashed?," 2016.
- [15] P. O'Sullivan, G. Thompson, and A. Clifford, "Applying data models to big data architectures," *IBM J Res Dev*, vol. 58, no. 5/6, pp. 1-11, 2014, doi: 10.1147/JRD.2014.2352474.
- [16] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," *AI Soc*, vol. 30, no. 1, pp. 89-116, 2015, doi: 10.1007/s00146-014-0549-4.
- [17] P. Routray, C. K. Swain, and S. P. Mishra, "A survey on sentiment analysis," *Int J Comput Appl*, vol. 76, no. 10, pp. 1-8, 2013.
- [18] F. Gorunescu, *Data Mining Concepts, Models and Technique*. Berlin, Heidelberg: Springer, 2013, doi: 10.1007/978-3-642-19721-5.
- [19] V. Singh and S. K. Dubey, "Opinion mining and analysis: A literature review," in *2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence)*, IEEE, 2014, pp. 232-239, doi: 10.1109/CONFLUENCE.2014.6949318.
- [20] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: European Conference on Machine Learning*, 2004, pp. 39-50, doi: 10.1007/978-3-540-30115-8_7.
- [21] A. Shoukry and A. Rafea, "Preprocessing Egyptian Dialect Tweets for Sentiment Mining," in *Fourth Workshop on Computational Approaches to Arabic-Script-based Languages*, 2012, pp. 47-56.
- [22] M. Ahmad and S. Aftab, "Analyzing the Performance of SVM for Polarity Detection with Different Datasets," *International Journal of Modern Education and Computer Science*, vol. 9, no. 10, pp. 29-36, 2017, doi: 10.5815/ijmecs.2017.10.04.
- [23] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, 2010, pp. 241-249.
- [24] G. Eysenbach, "Infodemiology and infoveillance: Tracking online health information and cyberbehavior for public health," *Am J Prev Med*, vol. 40, no. 5, pp. S154-S158, 2011, doi: 10.1016/j.amepre.2011.02.006.
- [25] T. Nasukawa and J. Yi, "Sentiment Analysis: Capturing Favorability Using Natural Language Processing," in *Proceedings of the 2nd international conference on Knowledge capture*, New York, NY, USA: ACM, 2003, pp. 70-77, doi: 10.1145/945645.945658.
- [26] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-90, 2008.
- [27] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, Boston, MA: Springer, 2012, pp. 415-463, doi: 10.1007/978-1-4614-3223-4_13.
- [28] C. Chew and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," *PLoS One*, vol. 5, no. 11, pp. 1-13, 2010, doi: 10.1371/journal.pone.0014118.
- [29] K. Danno and T. Horio, "Sunburn cell: factors involved in its formation," *Photochem Photobiol*, vol. 45, no. 5, pp. 683-690, 1987.
- [30] E. E. Kent et al., "'Obesity is the New Major Cause of Cancer': Connections Between Obesity and Cancer on Facebook and Twitter," *Journal of Cancer Education*, vol. 31, no. 3, pp. 453-459, 2016, doi: 10.1007/s13187-015-0824-1.
- [31] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *Knowledge Engineering Review*, vol. 25, no. 2, pp. 137-166, 2010, doi: 10.1017/S0269888910000032.
- [32] P. Chapman, "The CRISP-DM User Guide," 1999.
- [33] I. A. Özen and I. İlhan, "Opinion Mining in Tourism: A Study on 'Cappadocia Home Cooking' Restaurant," in *Smart Technology Applications in the Tourism Industry*, Pennsylvania, USA: IGI Global, 2020, pp. 43-64, doi: 10.4018/978-1-7998-1989-9.ch003.
- [34] B. Güçlü, M. Garza, and C. Kennett, "What Are Basketball Fans Saying on Twitter?: Evidence From Euroleague Basketball's Final Four Event," in *Big Data and Knowledge Sharing in Virtual Organizations*, Pennsylvania, USA: IGI Global, 2019, pp. 176-197, doi: 10.4018/978-1-5225-7519-1.ch008.
- [35] A. Kumar and D. Gupta, "Sentiment Analysis as a Restricted NLP Problem," in *Natural Language Processing for Global and Local Business*, Pennsylvania, USA: IGI Global, 2021, pp. 65-96, doi: 10.4018/978-1-7998-4240-8.ch004.





- [36] P. Assiroj, H. L. H. S. Warnars, and A. Fauzi, "Comparing CART and C5.0 Algorithm Performance of Human Development Index," in *2018 Third International Conference on Informatics and Computing (ICIC)*, IEEE, 2018, pp. 1–5, doi: 10.1109/IAC.2018.8780439.

BIOGRAPHIES OF AUTHORS







Priati Assiroj     was born in Cirebon, Jawa Barat, Indonesia. From 2014 to 2016, she was a lecturer at Universitas Singaperbangsa Karawang, Indonesia, and from 2016 to recent she is a lecturer at Universitas Buana Perjuangan Karawang in Department of Information System. Since January 2019 she is a lecturer in Politeknik Imigrasi, Ministry of Law and Human Rights, Republic of Indonesia. Since March 2018, she has been a scholar of the Bina Nusantara Graduate Program, Doctor of Computer Science, Bina Nusantara University Jakarta, Indonesia and in this doctoral program she received an international research grant from Bina Nusantara International research endowment and then she completed her Ph.D. in computer science in May 2022. Her research fields are data mining, high-performance computing and evolutionary algorithm. She can be contacted at email: priati.assiroj@poltekim.ac.id and tie.assiroj@gmail.com.



Asep Kurnia     is a doctor of educational science from Jakarta State University, Indonesia. He received his Master of Economics in Management from Krisnadwipayana University Jakarta and received his Bachelor of Law from the Islamic University of Djakarta. He actives in Indonesian immigration with training and education activities such as training in the identification of fraudulent travel documents at Australia in 2004, intellectual property academy at Korea in 2005, and many more related to the Ministry of Law and Human Rights Indonesia. Recently he is the expert for strengthening bureaucratic reform, Ministry of Law and Human Rights Indonesia. His research interest is management, human resources, policy, and data-driven. He can be contacted at email: asepk234@gmail.com.



Sirojul Alam     received a Bachelor of Computer in Information Technology from Buana Perjuangan University, Indonesia. Currently, he is a Data Governance Officer of Strategic Business Unit Digital Business of the Indonesian Government Minting and Security Printing Corp a.k.a Perum Peruri Indonesia. He is also certified as an SQL database administrator from BNSP Indonesia and holds international certification from Cisco AppDynamics application performance monitoring. His research interests are data mining, machine learning, and sentiment analysis in python programming. He can be contacted at email: sirojmu@gmail.com and sirojul.alam@peruri.co.id.