Sentiment Analysis of Technology Utilization by Pekanbaru City Government Based on Community Interaction in Social Media

Bunga Nanti Pikir [1], M. Khairul Anam [2], Hadi Asnal [3], Rahmaddeni [4], Triyani Arita Fitri [5], Hamdani

[1,2,3,4,5]STMIK Amik Riau, Jl. Purwodadi Indah KM 10 Panam, Pekanbaru, Indonesia

Abstract

Government services for the public are currently utilizing technology, especially in the city of Pekanbaru. The government has currently centralized all services for the public, both online and offline, in public service malls. The type of service that uses technology, especially for online services, has received criticism from the netizen in online media such as Twitter. In an effort to improve services for the public, the government must listen to the wishes of the people. To resolve that issue, this study will use sentiment analysis to see positive, negative, and neutral comments. The method used is to see the accuracy generated using the Naïve Bayes Classifier (NBC) method. Bayes Classifier is a statistical classifier, where the classifier can predict the probability of class membership of a data tuple that will fall into a certain class, according to the probability calculation. Accuracy results are obtained by dividing training data and testing data with a comparison of 70%:30% with an accuracy value of 55.56%, Precision 64%, recall 80%, f-score 71.2%.

1. Introduction

Pekanbaru City is currently experiencing fairly rapid development, from the infrastructural development to the technological developments [1][2]. The development of information technology requires the Pekanbaru City Government to provide services that are faster, more effective and efficient. For this reason, a technology-based government system such as e-money has been developed at Sultan Syarif Kasim II airport [3]. Problems that occur to utilize the technology by the Pekanbaru City Government are still receiving complaints from the public such as the command center

public service [4] which only some people know about this public service and the utilization of CCTV which is applied in traffic light [5] which is not working properly. We can see the utilization of technology by the Pekanbaru City Government from several web portal sites and other official web portals. Meanwhile, to see the various comments of netizens seen from social media Twitter. Netizens' comments on the government are a form of community participationfor regional progress [6].

This research start from the web scraping stage, sentiment analysis, and analyzes

Bunga Nanti Pikir
Email: bunganantipikir@sar.ac.id.

sentiment results using the Naïve Bayes classifier method. Web Scraping is the process of retrieving documents or data from web pages that are needed for analysis [7]. The web scraping process can be done manually and automatically, and then the data that has been taken from web pages is stored in a Spreadsheet or Microsoft excel. At the web scraping stage in this study, data were collected from several web portals which were the top sites as well as from official portals discussing the Pekanbaru City Government. The results of web scraping from several web portals are used as reference material to see how the utilization of technology by the Pekanbaru City Government. After we get the data from the web scraping, we can see other netizen comments on social media.

Social media that is used as an object to see netizen comments is Twitter, where netizen comments obtained on Twitter are classified into 3 types of comments, namely positive, negative and neutral comments. The types of comments are obtained after crawling the data by utilizing the Twitter API, namely by using sentiment analysis in the comment classification process. 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 [8]. In this study, sentiment analysis was carried out to see the opinions or tendencies of netizens' opinions towards the Pekanbaru City Government which contained positive, negative and neutral sentiments regarding the utilization of technology, then the sentiment results were analyzed using the Naïve Bayes Classifier method.

Naïve Bayes is a data mining algorithm with a classification method. Naïve Bayes is a classification with probability and statistical methods [9]. The Naïve Bayes Classifier research aims to carry out the process of classifying the results of netizen comments regarding the application of

technology that has gone through a sentiment analysis process.

There are several methods used to apply sentiment analysis, including: Naïve Bayes Classfier, Support Vector Machine (SVM), Neural Network, K-Nearest Neighbor, K-Means, and so forth [10][11]. Based on previous research [12][13][14][15], revealed that the Naïve Bayes Classifier method has high speed and accuracy when applied to large databases and diverse data and the advantages of the Naïve Bayes Classifier method are the simplicity, and the accuracy is quite high, by 90%. Because of that reason, this study uses the Naïve Bayes classifier method. The results of this study will be compared with previous studies using 3 methods with fairly high accuracy [16].

2. Research Methods

The stages of the research process carried out in this study are described in the flow of the research methodology as follows:
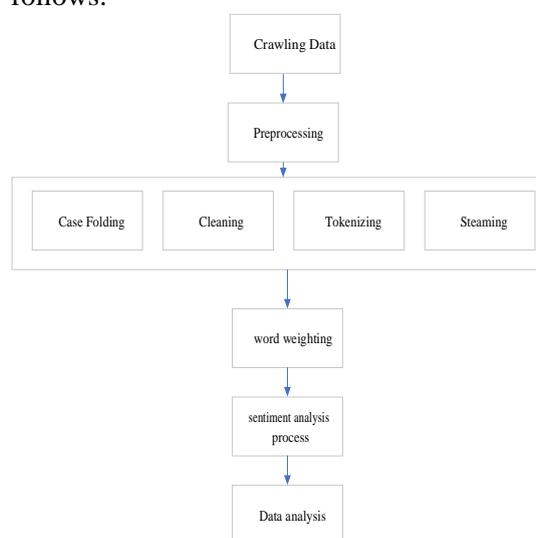

Figure 1. Methodology flow

The following is the explanation of Figure 1.

2.1. Crawling Data

The following is the data crawling process using the Twitter API with
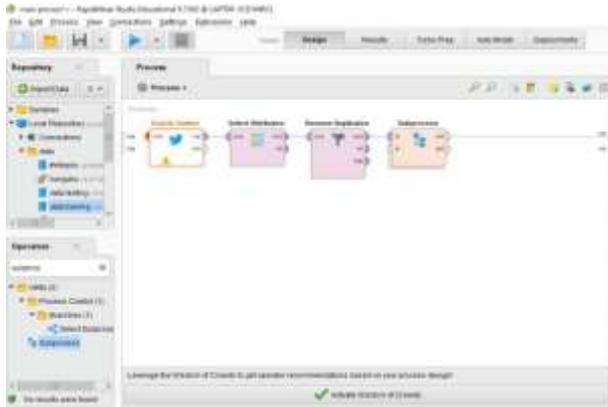
RapidMiner.

Figure 2. Crawling data

After the Twitter data crawling process is carried out as shown in Figure 2, the next step is to pre-process the data obtained.

2.2. Preprocessing

The following are the tools used to perform preprocessing in this study:

A. The first stage is case folding, this is the process to convert all document text into lowercase letters [17].

B. The second stage is cleaning, which is the stage where characters other than letters are removed and are considered as delimiters, and delete URLs, Mentions and Hashtags [18].

C. The third stage is tokenizing which is the process of separating text data into several tokens. Tokenizing broadly breaks down a set of characters in a text into a word, distinguishing certain characters that can be treated as word separators or not. [19].

D. The last stage is Stemming, where this stage occurs the removal of affixes on each word so that it becomes a basic word, and at this stage it also aims to clean up a word from improper spelling [20].

2.3. Word Weighting

Data that has gone through the preprocessing stage must be in numeric form. To convert the data into numeric, using the TF-IDF weighting method. Metode Term Frequency Invers Document Frequency (TF-IDF) is a method used to determine how far the terms are connected to the document by giving each word a weight [21][22]. This TF-IDF method combines two concepts, namely the frequency of occurrence of a word in a document and the inverse frequency of the document containing the word [23].

In calculating the weights using TF-IDF, the TF value per word is calculated first with the weight of each word valued 1. While the IDF value is formulated in Equation (1).

$$IDF(word) = \log \frac{td}{df} \qquad (1)$$

IDF (word) is the IDF value of each word to be searched, td is the total number of existing documents, df is the number of words that appear in all documents.

2.4. Classification Process with Naïve Bayes

Based on the analysis of data requirements in the previous text mining process, this process will explain the processes that will be carried out in data classification.

A. Training Data

At this stage, the value of the data that weight is known will be used as training data for being a reference in making a classification model. Then we will look for the category probability value and the probability of each word in each term for each class from the training data.

1. Prior Determination Process

First, calculate the probability of each category (prior), in this study, there are three categories, namely positive, negative and neutral categories.

$$P(Pos/Neg/Net) = \frac{d(Pos/Neg/Net)}{|c|}$$

$$P(Pos) = \frac{d(Pos)}{|c|} = \frac{1}{3}$$

$$P(Neg) = \frac{d(Neg)}{|c|} = \frac{1}{3}$$

$$P(Net) = \frac{d(Net)}{|c|} = \frac{1}{3}$$

2. Probability

Then the probability is calculated on each term of all documents. The total number of terms used in this calculation is 35 (table 3.7), 24 positive class terms, 27 negative class terms, and 23 neutral class terms. The number of terms depends on the results of preprocessing the data. The following is a probability calculation for each term:

$$P\left(W_k \middle| \begin{array}{c} \text{Pos} \\ \text{Neg} \\ \text{Net} \end{array}\right) = \frac{(n_k,\text{Pos/Neg/Net})+1}{(n,\text{Pos/Neg/Net})+|\text{vocabulary}|}$$

Known:
|vocabulary| = 74
Positive Term = 24
Negative Term = 27
Neutral Term = 23

For example, the probability of the word "ada"

$$P(\text{"ada"}|\text{"Pos"}) = \frac{(\text{"ada"}|\text{"Pos"})+1}{(\text{"Pos"})+|\text{vocabulary}|} = \frac{1+1}{24+74} = 0.0102$$

$$P(\text{"ada"}|\text{"Neg"}) = \frac{(\text{"ada"}|\text{"Neg"})+1}{(\text{"Neg"})+|\text{vocabulary}|} = \frac{0+1}{27+74} = 0.0198$$

$$P(\text{"ada"}|\text{"Net"}) = \frac{(\text{"ada"}|\text{"Net"})+1}{(\text{"Net"})+|\text{vocabulary}|} = \frac{1+1}{23+74} = 0.0103$$

Then calculated all the probability values. After getting all the probability values for each word, then a test data search is carried out.

B.   Test Data
In this process, the testing process will be carried out on the test data using the test data in the previous training process. Here is a tweet for the test data.

Table 1. Example Comments

Comment	Class
@sunguikkusuik Agar saat diterapkan nanti dapat berjalan dengan baik #PkuSmartCity	

In the test data above, the probability value is calculated based on the probability value in each term. Calculating the probability value using the Naïve Bayes method in this test data is by multiplying the probability value of all categories by the

probability in each term taken from all the training data.
Before going through the calculation process, the test data first performs the text mining process. Table 2 is the result of test data that has gone through the preprocessing stage.

Table 2. Example Comments after preprocessing

Comment	Class
agar saat diterapkan nanti dapat berjalan dengan baik pkusmartcity	

Known:
|vocabulary| = 74
Positive Term = 24
Negative Term = 27
Neutral Term = 23

Misalnya Probabilitas kata "agar"

$$P(\text{"agar"}|\text{"Pos"}) = \frac{(\text{"agar"}|\text{"Pos"})+1}{(\text{"Pos"})+|\text{vocabulary}|} = \frac{0+1}{24+74} = 0,0102$$

$$P(\text{"agar"}|\text{"Neg"}) = \frac{(\text{"agar"}|\text{"Neg"})+1}{(\text{"Neg"})+|\text{vocabulary}|} = \frac{0+1}{27+74} = 0,0099$$

$$P(\text{"agar"}|\text{"Net"}) = \frac{(\text{"agar"}|\text{"Net"})+1}{(\text{"Net"})+|\text{vocabulary}|} = \frac{0+1}{23+74} = 0,01031$$

The following are the results of the probability calculation process from the test data.

Table 3. Hasil Probablitas Kata Uji

No	Term (t)	Probability		
		Neg	Pos	Net
1	Agar	0,0102	0,0099	0,01031
2	Saat	0,0102	0,0099	0,01031
3	terap	0,01031	0,0101	0,01042
4	nanti	0,0102	0,0099	0,01031
5	dapat	0,01031	0,0101	0,01042
6	jalan	0,01031	0,01	0,01052
7	dengan	0,0102	0,0099	0,01031
8	baik	0,01031	0,01	0,01063
9	pkusmartcity	0,01041	0,0101	0,01052

The next step is to calculate the category probability from the test data.

$$P(Pos/Neg/Net|d) = P(Pos/Neg/Net) * \Pi_i P(a_i|Pos/Neg/Net)$$

P(uji|Pos) = P(Pos) x P(agar | Pos) x P(saat | Pos) x P(terap | Pos) x P(nanti | Pos) x P(dapat | Pos) x P(jalan | Pos) x P(dengan | Pos) x P(baik | Pos) x P(pkusmartcity | Pos)

P(uji|Pos) = 3,27161E-19

P(uji|Neg) = P(Neg) x P(agar | Neg) x P(saat | Neg) x P(terap | Neg) x P(nanti | Neg) x P(dapat | Neg) x P(jalan | Neg) x P(dengan | Neg) x P(baik | Neg) x P(pkusmartcity | Neg)

P(uji|Neg) = 4,19866E-19

P(uji|Net) = P(Neg) x P(agar | Net) x P(saat | Net) x P(terap | Net) x P(nanti | Net) x P(dapat | Net) x P(jalan | Net) x P(dengan | Net) x P(baik | Net) x P(pkusmartcity | Net)

P(uji|Net) = 4,75809E-19

The results of the calculation of the probability value above obtained the highest value on P(uji|Net) which is 3.27161E-19, so that the results of the comments on the test data are classified in the **POSITIF** category.

2.5. Data Analysis

After doing manual calculations, the next step is to analyze data using RapidMiner to see accuracy, recall, and precision.

3. Results and Discussion

The process carried out in this study is almost the same as previous research [16]. However, this study did not use Cross-Validation. Previous studies got 100% results using the Naïve Bayes method. Figure 3 is the process flow design that is carried out.
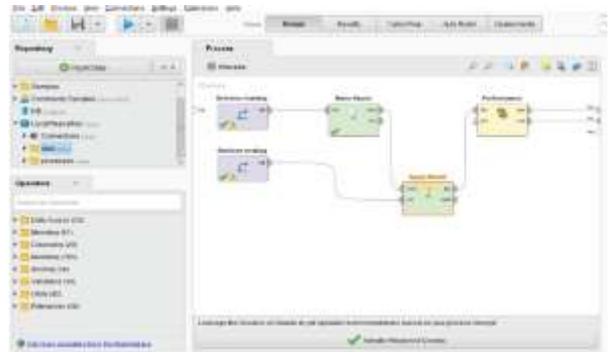

Figure 3. Process flow design

Here is the explanation of figure 3:
a. The apply model is used to accommodate the resulting data from training data and testing data.
b. Naïve Bayes is used for calculating the methods used in RapidMiner.
c. Performance is used to see the results of data accuracy in the form of percentages and in the form of confidence.

3.1. RapidMiner Analysis Results

The display of accuracy using RapidMiner is shown in the following figure.
a. Sentiment Analysis

Figure 4. Display of analysis result

RapidMiner can also see the performance results from the Naïve Bayes calculation, so the performance results obtained from the Naïve Bayes analysis are 55.56%.
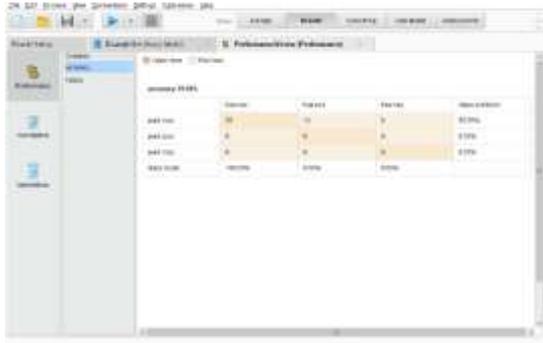
Figure 5. Display of sentiment accuracy results

RapidMiner can also see the results on the confidence graph from the Naïve Bayes calculation, and the graph results show NEUTRAL data, the graph tends to be NEUTRAL as well. So it can be concluded that the sentiment towards the tweet data is NEUTRAL.
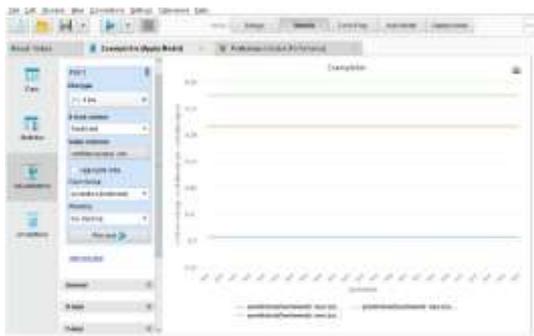

Figure 6. Display of analysis graph results

3.2. Testing

The testing stage is carried out by taking the results from the RapidMiner which is the result of the test data process as many as 45 test data from a sample of 70 data. The following is the result of the rapid miner. In this test, 70 comments of training data and 45 comments of test data are used. The results are as in the table below.

Figure 7. Testing Dataset 70% : 30%

Manual Classification	RapidMiner Classification		
	Pos	Neg	Net
Pos	0	0	14
Neg	0	0	6
Net	0	0	25

Based on the table above, it shows that the 70: 30 dataset test has an accuracy value of:

$$accuracy = \frac{0+0+25}{45} x100\%$$

$$accuracy = \frac{25}{45} x100\%$$
$$accuracy = 55,55\%$$

The results of manual testing and RapidMiner testing get the same results with an accuracy value of: 55.55%.

Precision and Recall Calculation

Precision is the ratio of true positive predictions to the overall positive predicted results. And recall is the ratio of true positive predictions compared to the overall data that are true positive.

True Positive (TP): cases where the comments are predicted to be Positive, and it is True as predicted.

True Negative (TN): cases where the comment is predicted Negative and in fact the sentiment is True Negative.

False Positive (FP): cases where comments that were predicted to be positive, turn out to be non-positive. The prediction is False.

False Negatif (FN): cases where the predicted Negative comment that is False predicted.

Positive Class

Precision	= (TP) / (TP+FP)
	= 0 / (0+14) = 0
Recall	= (TP) / (TP + FN)
	= 0 / (0+6) = 0
F Score	= 2 * ((precision*recall)/(precision+recall))
	= 2 * 0 = 0

Negative Class

Precision	= (TP) / (TP+FP)
	= 0 / (0+14) = 0
Recall	= (TP) / (TP + FN)
	= 0 / (0+6) = 0
F Score	= 2 * ((precision*recall)/(precision+recall))
	= 2 * 0 = 0

Neutral Class

Precision	= (TP) / (TP+FP)
	= 25 / (25+14) = 0,64/64%
Recall	= (TP) / (TP + FN)

37

$= 25 / (25+6) = 0,80/$ **80%**

F Score $= 2 *$
$((precision*recall)/(precision+recall))$
$= 2 * ((0,64*0,8)/(0,64+0,8))$
$= 2 * 0,35$
$= 0,712/71,2\%$

4. Conclusion

After completing a series of steps towards analyzing sentiment data on Twitter with the hashtag #pemerintahanpku, #teknologipku, #smartcitypekanbaru, #pekanbaru, #infopku regarding the technology utilization by the Pekanbaru City Government, several conclusions can be drawn including the graphic data showing that the application of technology by the Pekanbaru City Government has a neutral sentiment with the largest value with a percentage of 55.56%.

Then the graph shows that netizens or the public have not fully cared about the utilization of technology applied by the Pekanbaru City Government, seen from the results of the sentiments that have been analysed which are neutral. The results of comparison with previous studies show a significant decrease. This is because this research does not use cross-validation in the classification process

5. Reference

[1] F. Firmansyah and A. B. Raharja, "Quantification of Land Cover Changes in Sub-urban Areas of Pekanbaru City," in *IOP Conference Series: Earth and Environmental Science*, 2021, vol. 887, no. 1, doi: 10.1088/1755-1315/887/1/012020.

[2] M. K. Anam, Purwanto, T. A. Fitri, and A. N. Ulfah, "SMART method utilization for meetinghouse elections in Pekanbaru City," *JAIA – J. Artif. Intell. Appl.*, vol. 1, no. 1, pp. 11–18, 2020.

[3] F. F. Adelia, M. K. Anam, T. A. Fitri, and F. Zoromi, "ANALISIS PERSPEKTIF PADA PENERAPAN E-MONEY MENGGUNAKAN DELONE AND MCLEAN IS SUCCESS MODEL DI BANDARA SULTAN SYARIF KASIM II PEKANBARU," *J. Inform. Rekayasa Elektron.*, vol. 3, no. 2, pp. 100–110, 2020, doi: https://doi.org/10.36595/jire.v3i2.256.

[4] G. Meiwanda, "Challenges of Smart City: Local Government in Pekanbaru City and Community," in *Annual Conference of Indonesian Association for Public Administration*, 2020, vol. 122, pp. 40–53, doi: 10.2991/aebmr.k.200301.003.

[5] H. Yodiansyah, N. Yuzalmi, and I. Zain, "Access To Communication Planning and Transportation Management Technology in Assessment of Vehicle Operations in Pekanbaru City'S, Riau, Indonesia," *Int. J. Res. -GRANTHAALAYAH*, vol. 6, no. 9, pp. 1–16, 2018, doi: 10.29121/granthaalayah.v6.i9.2018.1 200.

[6] M. K. Anam, "Analisis Respons Netizen Terhadap Berita Politik Di Media Online," *J. Ilm. Ilmu Komput.*, vol. 3, no. 1, pp. 14–21, 2017, doi: 10.35329/jiik.v3i1.62.

[7] E. R. Arumi and P. Sukmasetya, "Exploiting Web Scraping for Education News Analysis Using Depth-First Search Algorithm," *JOIN (Jurnal Online Inform.)*, vol. 5, no. 1, pp. 19–26, 2020, doi: 10.15575/join.v5i1.548.

[8] B. Saberi and S. Saad, "Sentiment analysis or opinion mining: A review," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 5, pp. 1660–1666, 2017, doi: 10.18517/ijaseit.7.5.2137.

[9] I. B. A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, "Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm," *Int. J. Eng. Emerg. Technol.*, vol. 2, no. 1, p. 53, 2017, doi:

10.24843/ijeet.2017.v02.i01.p11.

[10] Z. H. Moe, T. San, M. M. Khin, and H. M. Tin, "Comparison of Naive Bayes and Support Vector Machine Classifiers on Document Classification," in *2018 IEEE 7th Global Conference on Consumer Electronics, GCCE 2018*, 2018, pp. 285–286, doi: 10.1109/GCCE.2018.8574785.

[11] M. K. Anam, B. Nanti, P. Gulo, M. B. Firdaus, and S. Erlinda, "Penerapan Naïve Bayes Classifier , K-Nearest Neighbor dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah Applications of Naïve Bayes Classifier , K-Nearest Neighbor and Decision Tree to Analyze Sentiment on Netizen and Gove," *Matrik J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 1, pp. 139–150, 2021, doi: 10.30812/matrik.v21i1.1092.

[12] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive Bayes classification algorithm for traffic risk management," *EURASIP J. Adv. Signal Process.*, vol. 30, 2021, doi: 10.1186/s13634-021-00742-6.

[13] A. W. Syaputri, E. Irwandi, and M. Mustakim, "Naïve Bayes Algorithm for Classification of Student Major's Specialization," *J. Intell. Comput. Heal. Informatics*, vol. 1, no. 1, p. 17, 2020, doi: 10.26714/jichi.v1i1.5570.

[14] S. Khomsah, "Naive Bayes Classifier Optimization on Sentiment Analysis of Hotel Reviews," *J. Penelit. Pos dan Inform.*, vol. 10, no. 2, p. 157, 2020, doi: 10.17933/jppi.2020.100206.

[15] S. Sendari, I. A. E. Zaeni, D. C. Lestari, and H. P. Hariyadi, "Opinion Analysis for Emotional Classification on Emoji Tweets using the Naïve Bayes Algorithm," *Knowl. Eng. Data Sci.*, vol. 3, no. 1, pp. 50–59, 2020, doi: 10.17977/um018v3i12020p50-

59.

[16] M. K. Anam, B. N. Pikir, M. B. Firdaus, S. Erlinda, and Agustin, "Penerapan Naïve Bayes Classifier , K-Nearest Neighbor dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah Applications of Naïve Bayes Classifier , K-Nearest Neighbor and Decision Tree to Analyze Sentiment on Netizen and Gove," *Matrik J. Manajemen, Tek. Inform. dan Rekayasa Komput. ?141*, vol. 21, no. 1, pp. 139–150, 2021, doi: 10.30812/matrik.v21i1.1092.

[17] Y. Yunitasari, A. Musdholifah, and A. K. Sari, "Sarcasm Detection For Sentiment Analysis in Indonesian Tweets," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 1, pp. 53–62, 2019, doi: doi.org/10.22146/ijccs.41136.

[18] E. Sutoyo and A. Almaarif, "Twitter sentiment analysis of the relocation of Indonesia's capital city," *Bull. Electr. Eng. Informatics*, vol. 9, no. 4, pp. 1620–1630, 2020, doi: 10.11591/eei.v9i4.2352.

[19] E. S. Romaito, M. K. Anam, Rahmaddeni, and A. N. Ulfah, "Perbandingan Algoritma SVM Dan NBC Dalam Analisa Sentimen Pilkada Pada Twitter," *CSRID J.*, vol. 13, no. 3, pp. 169–179, 2021, doi: 10.22303/csrid.13.3.2021.169-179.

[20] A. N. Ulfah and M. K. Anam, "Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM)," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 1, pp. 1–10, 2020, doi: 10.35957/jatisi.v7i1.196.

[21] S. E. Pratama, W. Darmalaksana, D. Sa'adillah Maylawati, H. Sugilar, T. Mantoro, and M. A. Ramdhani, "Weighted inverse document frequency and vector space model for hadith search engine," *Indones. J.*

Electr. Eng. Comput. Sci., vol. 18, no. 2, pp. 1004–1014, 2020, doi: 10.11591/ijeecs.v18.i2.pp1004-1014.

[22] S. H. Mahanani, T. Mauritsius, and N. Legowo, "Recommender System using Content Based Filtering for News Portal in Indonesia," *Int. J. Recent Technol. Eng.*, vol. 8, no. 6, pp. 173–178, 2020, doi: 10.35940/ijrte.f7231.038620.

[23] W. B. Trihanto, R. Arifudin, and M. A. Muslim, "Information Retrieval System for Determining The Title of Journal Trends in Indonesian Language Using TF-IDF and Na?ve Bayes Classifier," *Sci. J. Informatics*, vol. 4, no. 2, pp. 179–190, 2017, doi: 10.15294/sji.v4i2.11876.