ILKOM Jurnal Ilmiah Vol. 13, No. 2, August 2021, pp. 168-174
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Performance comparison of support vector machine (SVM) with linear kernel and polynomial kernel for multiclass sentiment analysis on twitter

Rifqatul Mukarramah [a, 1, *]; **Dedy Atmajaya** [a, 2]; **Lutfi Budi Ilmawan** [a, 3]
[a] *Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM.5, Makassar dan 90231, Indonesia*
[1] *rifqatulmukarramah@gmail.com; 2 dedy.atmajaya@umi.ac.id; 3 lutfibudi.ilmawan@umi.ac.id*
** Corresponding author*

Article history: Received April 30, 2021; Revised July 04, 2021; Accepted July 04, 2021; Available online August 07, 2021

Abstract

Sentiment analysis is a technique to extract information of one's perception, called sentiment, on an issue or event. This study employs sentiment analysis to classify society's response on covid-19 virus posted at twitter into 4 polars, namely happy, sad, angry, and scared. Classification technique used is support vector machine (SVM) method which compares the classification performance figure of 2 linear kernel functions, linear and polynomial. There were 400 tweet data used where each sentiment class consists of 100 data. Using the testing method of k-fold cross validation, the result shows the accuracy value of linear kernel function is 0.28 for unigram feature and 0.36 for trigram feature. These figures are lower compared to accuracy value of kernel polynomial with 0.34 and 0.48 for unigram and trigram feature respectively. On the other hand, testing method of confusion matrix suggests the highest performance is obtained by using kernel polynomial with accuracy value of 0.51, precision of 0.43, recall of 0.45, and f-measure of 0.51.

Keywords: SVM Algorithm; Covid-19; Kernel Linear; Kernel Polynomial; Sentiment Analysis

Introduction

Global society was shocked by the emergence of corona virus, named Corona Virus Disease 2019 (Covid-19), that infects human respiratory system and can easily be transmitted between people [1][2][3]. The first case was found at Wuhan, China at the end of December 2019 and started spreading to other countries including Indonesia. The first case in Indonesia was announced on 2nd March, 2020 [4][5]. The rapid spread and intensive increase in the number of infected patients led people to feel restless, anxious, and scared of this outbreak. This triggers people to express all feelings and opinions on social media, especially Twitter. Twitter is a micro-blogging system technology that allows users to easily upload activities, feelings, and experiences, called tweets, to the internet, anywhere, anytime, and in real time [6].

Based on the problems above, it is necessary to analyze public opinion on Twitter about Covid-19 with the classification of happy, sad, angry, and scared. Sentiment analysis or also known as opinion mining is a process of understanding, extracting and processing textual data automatically to obtain sentiment information contained in a sentence whether it has a positive or negative opinion [7]. Several previous studies on data classification that compared several methods [8][9][10][11][12] showed that the Support Vector Machine method offered better performance than other classification methods in their respective case studies. Therefore, this study applied the Support Vector Machine (SVM) method with Linear Kernel and Polynomial Kernel in the classification process. Support Vector Machine (SVM) is a method that classifies linear and non-linear data sets by converting the original training data to a higher dimension so that these dimensions look for the decision limits needed to separate classes and their records from others [13].

The results of this study will provide an overview of people's perceptions of Covid-19, whether they tend to be happy, sad, angry, or scared. This study also compares the performance results of the SVM method with the Linear Kernel and Polynomial Kernel.

Method

The initial stage of this research is to carry out a data crawling process to get tweets that discuss Covid-19 on Twitter. The data that has been obtained will go through a labeling process where each tweet will be labeled according to the polarity of the sentiment being experienced. The next stage is pre-processing, feature extraction and

http://dx.doi.org/10.33096/ilkom.v13i2.851.168-174

tf-idf. The data will be used as a dataset for the sentiment classification process. The output of this process is a prediction of the sentiment experienced by Twitter users about the Covid-19 outbreak. In addition, this process produces performance values for each kernel used by calculating the values of accuracy, precision, recall and f-measure.

This research uses the Support Vector Machine (SVM) method with 2 kernel functions, namely linear and polynomial. Extraction of the unigram and trigram features was also carried out to produce tweet predictions related to the Covid-19 outbreak with 4 sentiment classes, namely happy, sad, angry, and afraid. This process also measures and compares the best performance of the classification method used. The program flow of this research can be seen in the following **Figure 2a and 2b**.

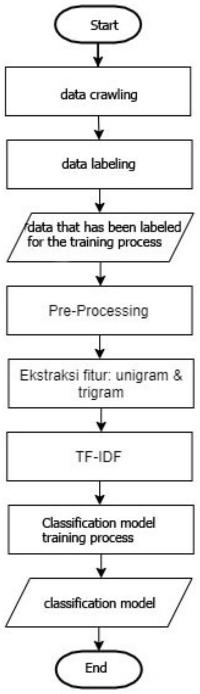

Start
data crawling
data labeling
data that has been labeled for the training process
Pre-Processing
Ekstraksi fitur: unigram & trigram
TF-IDF
Classification model training process
classification model
End

Figure 2.a Flowchart of classification model determination

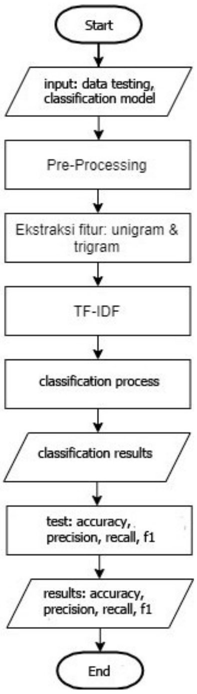

Start
input: data testing, classification model
Pre-Processing
Ekstraksi fitur: unigram & trigram
TF-IDF
classification process
classification results
test: accuracy, precision, recall, f1
results: accuracy, precision, recall, f1
End

Figure 2.b Flowchart of classification process

The data set is obtained from the process of crawling data on Twitter using the Twitter API and then labeled into 4 sentiment classes, namely happy, sad, angry, and scared. After going through the labeling stage, the next stage is pre-processing which is selecting data that is likely to cause problems in the results of data processing due to inappropriate data selected to make it easier to process by the system created. Pre-processing consists of several stages, namely:

a) Cleansing is used to remove URLs, mentions, usernames, RTs, hashtags, numbers, punctuation marks, and emoticons.
b) Case folding is used to uniform all letters into lowercase letters.
c) Tokenizing is used to cut each word in a sentence
d) Stemming is used to change the word in the sentence to the basic word form
e) Stopword is used to remove words that have no meaning and are not needed in the classification process.

After pre-processing, feature extraction using unigram and trigram is carried out. While the former is the word extract in the review sentence with n=1 or single term, the latter is extract of n-words in the review sentence with n=3 [14]. The next stage is TF-IDF, a statistical-based weighting technique, which is applied in various information mining issues [15]. This technique is used to compute a weight to each word which signifies the importance of the word in the document.

After weighting each word using TF-IDF, the data will be stored as training data which will then be used as a classification model to determine sentiment analysis from Twitter using the Support Vector Machine SVM method). The Support Vector Machine (SVM) classification process uses the python programming language with the Scikit-Learn library.

Results and Discussion

This section discusses the classification and testing results of the tweet classification model that has been built. After the process of crawling Twitter data and labeling them according to the sentiment experienced, the next stage is pre-processing.

A. *Pre-Processing*

The initial stage step is cleansing which is to remove URLs, mentions, usernames, RTs, hashtags, numbers, punctuation marks, and emoticons as shown in **Table 1**.

Table 1. Tweet after Cleansing

Tweet	Label
Mengikuti arus dunia yang menyedihkan Corona anjing	Angry
Orang orang sedang phobia virus Corona flu dan demam efek kehujanan kemarin aja di jauhin terus suruh periksa kedokter	Scared

The next step is case folding which is to uniform all letters into lowercase as given in **Table 2**.

Table 2. Tweet after Case Folding

Tweet	Label
mengikuti arus dunia yang menyedihkan corona anjing	Angry
orang orang sedang phobia virus corona flu dan demam efek kehujanan kemarin aja di jauhin terus suruh periksa kedokter	Scared

Next, *tokenizing* cut each word in a sentence as shown in **Table 3**.

Table 3. Tweet after Tokenizing

Tweet	Label
‘mengikuti’, ‘arus’, ‘dunia’, ‘yang’, ‘menyedihkan’, ‘corona’, ‘anjing’	Angry
‘orang’, ‘orang’, ‘sedang’, ‘phobia’, ‘virus’, ‘corona’, ‘flu’, ‘dan’, ‘demam’, ‘efek’, ‘kehujanan’, ‘kemarin’, ‘aja’, ‘di’, ‘jauhin’, ‘terus’, ‘suruh’, ‘periksa’, ‘kedokter’	Scared

Stemming is the next step which is to change word in the sentence to the basic form as shown in **Table 4**.

Table 4. Tweet after Stemming

Tweet	Label
‘ikut’, ‘arus’, ‘dunia’, ‘yang’, ‘sedih’, ‘corona’, ‘anjing’	Marah
‘orang’, ‘orang’, ‘sedang’, ‘phobia’, ‘virus’, ’corona’, ‘flu’, ‘dan’, ‘demam’, ‘efek’, ‘hujan’, ‘kemarin’, ‘aja’, ‘di’, ‘jauh’, ‘terus’, ‘suruh’, ‘periksa’, ‘dokter’	Takut

The final step is *stopword* which is to remove words that have no meaning as shown in **Table 5**.

Table 5. Tweet after Stopword

Tweet	Label
arus dunia sedih corona anjing	Marah
orang orang phobia virus corona flu demam efek hujan kemarin jauhin suruh periksa dokter	Takut

B. *Feature Extraction*

This study applies n-gram with word extraction on tweet data with n=1, called unigram, and n=3, called trigram. The feature extraction results can be seen in **Table 6**.

Table 6. *Feature Extraction Using Unigram and Trigram*

Unigram	Trigram	Label
- arus - dunia - sedih - corona - anjing	- arus dunia sedih - dunia sedih corona sedih corona anjing	Angry
- orang - phobia	- orang phobia virus - phobia virus corona	Scared

Unigram	Trigram	Label
- virus - corona - flu - demam - efek - hujan - kemarin - jauh - suruh - periksa - dokter	- virus corona flu - corona flu deman - flu demam efek - demam efek hujan - efek hujan kemarin - hujan kemarin	

C. *TF-IDF*

The extracted data is still in the form of qualitative data so that it needs to be reprocessed into quantitative data using the TF-IDF calculation. Below are three documents to show the results of the TF-IDF calculation. The results of the calculation of tweet data using TF-IDF can be seen in **Table 7**.

Table 7. TF-IDF Calculation

TF	IDF	TF-IDF	Term
0.142857	1.393043	0.199006	Korona
0.142857	5.892852	0.841836	Rasa
0.142857	6.298317	0.899760	Rebah
0.142857	6.991465	0.998781	Selasa
0.142857	4.912023	0.701718	Semenjak
0.142857	6.991465	0.9987871	Senin

D. *Classification Process*

After passing through the labelling, pre-processing, feature extraction, and TF-IDF, the next stage is to carry out classification process using the SVM model. In order to determine the prediction results of the classification model used suit the label or not, tweet data obtained from the testing data set is given. The data can be seen in **Table 8**.

Table 8. *Data Testing*

Tweet	Label
alhamdulillah, negative corona positif cantik..	Happy
pensi sekolahku gagal tahun ini garagara corona	Sad
kenapa sih masih banyak orang yang gasuka pakai masker, gak takut kena corona yah	Angry
orang-orang takut ibunya kena korona jadi pada nggak berani bantuin	Afraid

The prediction results can be seen in **Table 9**.

Table 9. Prediction Results of Unigram Feature

Tweet	*Kernel*	*Prediksi*
alhamdulillah, negative corona positif cantik..	Linear	Afraid
	Polynomial	Afraid
pensi sekolahku gagal tahun ini garagara corona	Linear	Sad
	Polynomial	Happy
kenapa sih masih banyak orang yang gasuka pakai masker, gak takut kena corona yah	Linear	Afraid
	Polynomial	Sad
orang-orang takut ibunya kena korona jadi pada nggak berani bantuin	Linear	Afraid
	Polynomial	Sad

Table 9 shows the prediction results from the classification model using the unigram feature providing some results which are not suitable. Using a linear kernel, it was found that only the second tweet is predicted to be correct. Meanwhile, by using a polynomial kernel, all tweets are predicted to be wrong.

Table 10. Prediction Results of Trigram Feature

Tweet	***Kernel***	***Prediction***
alhamdulillah, negative corona positif cantik..	Linear	Afraid
	Polynomial	Happy
pensi sekolahku gagal tahun ini garagara corona	Linear	Afraid
	Polynomial	Sad
kenapa sih masih banyak orang yang gasuka pakai masker, gak takut kena corona yah	Linear	Happy
	Polynomial	Sad
orang-orang takut ibunya kena korona jadi pada nggak berani bantuin	Linear	Happy
	Polynomial	Sad

Table 10 shows the prediction results from the classification model using the trigram feature, some results were found not suitable. There were no tweets predicted to be correct in the result of linear kernel. Meanwhile, using a kernel polynomial, the first and second tweets are predicted to be correct, the third and fourth tweets are predicted incorrect.

E. Testing

The tweet classification test was carried out by measuring the accuracy, precision, recall, and f-measure values from the Support Vector Machine calculation with 2 kernel functions, namely linear and polynomial, and using unigram and trigram feature extraction. The data used for the classification model testing process was obtained from Indonesian-language tweets raising the topic of Covid-19 with a total of 400 tweets divided into 4 classes, namely 100 tweets with happy labels, 100 tweets with sad labels, 100 tweets with angry labels, and 100 tweets with scared labels. The testing of the performance of the classification model employs the k-fold cross validation and confusion matrix methods.

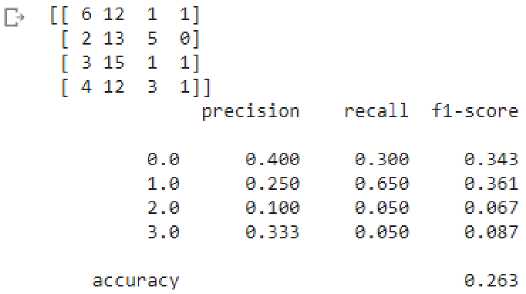
The cross-validation test uses a value of k=4. In each iteration, one of the folds will be selected as testing data and the remaining will become training data. Each data can only be used as testing data once. The process of calculating the accuracy of the data testing will continue to repeat until all iterations are complete.

Table 11. Accuracy Test Results with 4-Fold Cross Validation

Iteration	**Linear**		**Polynomial**	
	Unigram	**Trigram**	**Unigram**	**Trigram**
1	0.337	0.437	0.425	0.512
2	0.275	0.300	0.337	0.437
3	0.275	0.287	0.312	0.450
4	0.237	0.425	0.287	0.512
Avg	0.280	0.360	0.340	0.480

Table 11 shows the results of the evaluation process of the classification model using 4-fold cross validation. In the first iteration, the highest accuracy is obtained using a polynomial kernel with trigram feature extraction. Similarly, the highest accuracy at the second, third and fourth iterations were obtained by using a polynomial kernel and trigram feature extraction, with 0.437, 0.450, and 0.512 respectively.

The confusion matrix test is used to visualize the performance of the classification algorithm which is used by comparing the actual value with the predicted value. The parameters that will be measured in this test are the values of accuracy, precision, recall, and f-measure.

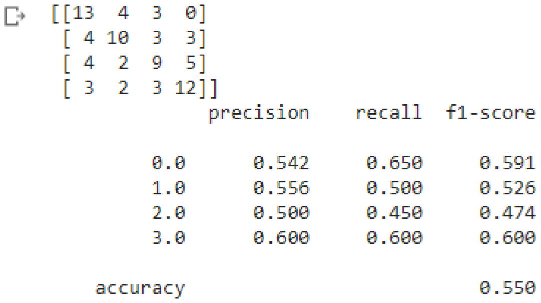


```
[[ 6 12  1  1]
 [ 2 13  5  0]
 [ 3 15  1  1]
 [ 4 12  3  1]]
              precision    recall  f1-score

         0.0      0.400     0.300     0.343
         1.0      0.250     0.650     0.361
         2.0      0.100     0.050     0.067
         3.0      0.333     0.050     0.087

    accuracy                          0.263
```

Figure 2.a Output of SVM evaluation using *linear kernel* with unigram feature

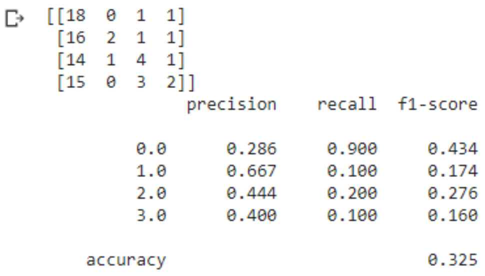


```
[[13  4  3  0]
 [ 4 10  3  3]
 [ 4  2  9  5]
 [ 3  2  3 12]]
              precision    recall  f1-score

         0.0      0.542     0.650     0.591
         1.0      0.556     0.500     0.526
         2.0      0.500     0.450     0.474
         3.0      0.600     0.600     0.600

    accuracy                          0.550
```

Figure 2.b Output of SVM evaluation using linear kernel with trigram feature

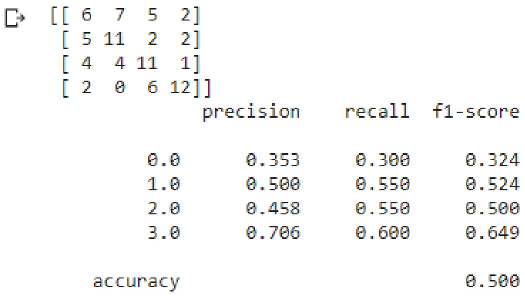


```
[[18  0  1  1]
 [16  2  1  1]
 [14  1  4  1]
 [15  0  3  2]]
              precision    recall  f1-score

         0.0      0.286     0.900     0.434
         1.0      0.667     0.100     0.174
         2.0      0.444     0.200     0.276
         3.0      0.400     0.100     0.160

    accuracy                          0.325
```

Figure 2.c Output of SVM evaluation using *polynomial kernel* with unigram feature



```
[[ 6  7  5  2]
 [ 5 11  2  2]
 [ 4  4 11  1]
 [ 2  0  6 12]]
              precision    recall  f1-score

         0.0      0.353     0.300     0.324
         1.0      0.500     0.550     0.524
         2.0      0.458     0.550     0.500
         3.0      0.706     0.600     0.649

    accuracy                          0.500
```

Figure 2.d Output of SVM evaluation using *polynomial kernel* with trigram feature

Figures 2.a, 2.b, 2.c and 2.d show the output of the testing process using linear and polynomial kernel functions with unigram and trigram feature extraction. The best performance in both linear and polynomial kernel were obtained by using the trigram feature with an accuracy value of 0.55 and 0.50 respectively.

Table 12 shows a comparison of the results of the Support Vector Machine performance using linear and polynomial kernels obtained by calculating the average value of each performance obtained for each kernel and feature used. Classification calculations using polynomial kernels and trigram features produce better test scores overall than using linear kernels with unigram or trigram features.

Table 12. Comparison of overall classification testing results

Performance	**Linear**		**Polynomial**	
	Unigram	**Trigram**	**Unigram**	**Trigram**
Accuracy	0.337	0.437	0.425	0.512
Precision	0.275	0.300	0.337	0.437
Recall	0.275	0.287	0.312	0.450
F-Measure	0.237	0.425	0.287	0.512

The comparison of performance results of support vector machine using linear and polynomial kernel can also be seen in the following. Chart.

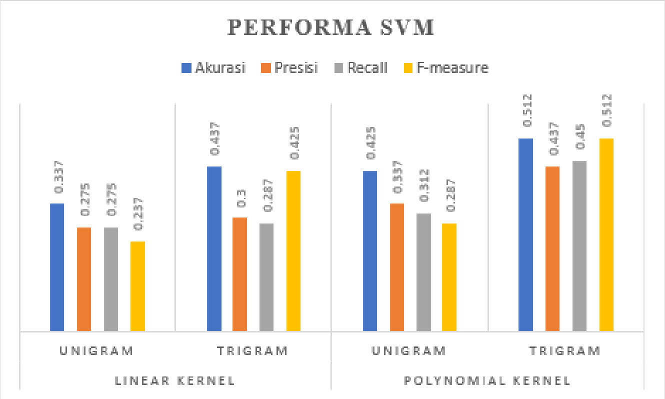


Figure 3. Chart of accuracy results in each kernel and feature used

Figure 3 portraits that polynomial kernel with trigram features resulted in better overall performance values with accuracy of 0.512, precision of 0.437, recall of 0.45, and f-measurement of 0,512, compared to the use of linear kernel with unigram features that provided accuracy of 0.337, precision of 0.275, recall of 0.275, and f-measurement of 0.237 and the trigram features with accuracy, precision, recall, and f-measurement of 0.437, 0.3, 0.287, and 0.425 respectively.

Conclusion

Based on the results and discussion, it is concluded that the best performance of the support ventor machine method is obtained by using the polynomial kernel function with an accuracy value of 0.512, precision of 0.437, recall of 0.45, and f-measurement of 0.512. Therefore, the most appropriate kernel function and feature extraction to be applied in this study is a polynomial kernel with trigram features.

References

[1] W. A. F. Dewi, “Dampak COVID-19 terhadap Implementasi Pembelajaran Daring di Sekolah Dasar,” Edukatif J. Ilmu Pendidik., vol. 2, no. 1, pp. 55–61, 2020.

[2] D. Telaumbanua, “Urgensi Pembentukan Aturan Terkait Pencegahan Covid-19 di Indonesia,” QALAMUNA J. Pendidikan, Sos. dan Agama, vol. 12, no. 01, pp. 59–70, 2020.

[3] M. Siahaan, “Dampak Pandemi Covid-19 Terhadap Dunia Pendidikan,” J. Kaji. Ilm., vol. 1, no. 1, pp. 73–80, 2020.

[4] A. Syahadati, N. C. Lengkong, O. Safitri, S. Machsus, Y. R. Putra, and R. Nooraeni, “ANALISIS SENTIMEN PENERAPAN PSBB DI DKI JAKARTA DAN DAMPAKNYA TERHADAP PERGERAKAN IHSG,” vol. 15, no. 1, pp. 20–25, 2021.

[5] R. Nuraini, “No Title,” indonesia.go.id. 2020.

[6] R. Habibi, D. B. Setyohadi, and E. Wati, “Analisis Sentimen Pada Twitter Mahasiswa Menggunakan Metode Backpropagation,” J. Inform., vol. 12, no. 1, 2016.

[7] A. V. Sudiantoro and E. Zuliarso, “Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma NAÏVE BAYES CLASSIFIER,” Pros. SINTAK 2018, pp. 398–401, 2018.

[8] L. Budi and A. Mude, “Perbandingan Metode Klasifikasi Support Vector Machine dan Naïve Bayes untuk Analisis Sentimen pada Ulasan Tekstual di Google Play Store,” vol. 12, no. 2, pp. 154–161, 2020.

[9] N. Hafidz, S. Anggraeni, and W. Gata, “Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan Naïve Bayes,” 2019.

[10] F. Sodik and I. Kharisudin, “Analisis Sentimen dengan SVM , NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter,” vol. 4, pp. 628–634, 2021.

[11] H. Azis et al., “Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah,” vol. 19, no. 3, pp. 286–294, 2020.

[12] M. S. Hadna, P. I. Santosa, and W. W. Winarno, “Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter,” Semin. Nas. Teknol. Inf. dan Komun., vol. 2016, no. Sentika, pp. 57–64, 2016.

[13] S. Sharma, S. Srivastava, A. Kumar, and A. Dangi, “Multi-Class Sentiment Analysis Comparison Using Support Vector Machine (SVM) and BAGGING Technique — An Ensemble Method,” 2018 Int. Conf. Smart Comput. Electron. Enterp., no. February 2019, pp. 1–6, 2018.

[14] N. B. N-gram, I. Pujadayanti, M. A. Fauzi, and Y. A. Sari, “Prediksi Rating Otomatis pada Ulasan Produk Kecantikan dengan Metode Prediksi Rating Otomatis pada Ulasan Produk Kecantikan dengan Metode Naïve Bayes dan N-gram,” no. November, 2018.

[15] D. H. Wahid and A. SN, “Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity,” IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 10, no. 2, p. 207, 2016.