International Applied Science
UMJember Proceeding Series (2022) Vol. 1, No. 1 : 57-66

ICRD 2020

Assessing Technique For Mapping Public Response To DKI Jakarta Governor Policy In Handling COVID-19 Pandemic Using SVM BASED Sentiment Analysis

Bagus Setya Rintyarna[1*], Wahyu Nurkholis Hadi Syahputra[2], Triawan Adi Cahyanto[1], Riska Nur Maulida[1]
[1]Universitas Muhammadiyah Jember; Indonesia
[2]Chiang Mai University; Thailand

DOI: https://doi.org/10.32528/ias.v1i1.50
*Correspondence: Bagus Setya Rintyarna
Email: bagus.setya@unmuhjember.ac.id

Published: Januari, 2022

Abstrak: Since the coronavirus outbreak or known as COVID-19 spread throughout the world, especially in Indonesia. The Governor of DKI Jakarta issued several policies to deal with the spread of COVID-19. However, this policy has become a conversation on social media such as Youtube. Through audience interaction in the comments column, giving lots of positive and negative sentiment comments, the audience response is classified using the sentiment analysis technique of comments to find out which sentiments are positive, negative, and neutral for each comment. In this study, the data were taken from news video comments. The method used is the Support Vector Machine and the selection feature uses the Term Frequency-Inverse Document Frequency (TF-IDF). The data used amounted to 945 Indonesian language comments. Accurate results obtained by using the addition of a stoplist at the preprocessing stage a.

Keywords: *Sentiment Analysis, TF-IDF, Support Vector Machine, Youtube, News.*

INTRODUCTION

Youtube is the most popular social media. The presence of Youtube is the social media that is most in demand by the public because various videos provide a lot of information such as entertainment, tips, and tricks, to the latest news. Not only that, but viewers can also comment on videos in the comments column. Youtube is increasingly developing with various features according to the social conditions of the community [1].

Youtube users in Indonesia experienced an 88% increase in accessing Youtube in 2020 [2]. The increase has occurred since the outbreak of the coronavirus or known as COVID-19 began to spread around the world. This is because many people are looking for information about the latest developments related to COVID-19, both national and international information.

With the emergence of the COVID-19 outbreak in Indonesia, the Governor of DKI Jakarta issued a regional quarantine-oriented policy such as the Large-Scale Social Restriction policy. The policy aims to prevent the spread of COVID-19. However, this policy is often rejected by the central government and certain parties. This has become the hottest news on Youtube so that many viewers have expressed their opinions in the news video commentary column. Expressing opinions or opinions that are positive or negative can be a reference to find out the public sentiment in the form of criticism or suggestions [3].

The classification of sentiments in comments is one of the functions of text mining. The classification process uses the Support Vector Machine (SVM) method. SVM is machine learning method based on the structural theory of statistical learning with the aim of finding the boundaries that separate each class [4].

Sentiment analysis is carried out to see a person's opinion on a problem or a certain figure, where an opinion can be categorized as a positive or negative opinion [5]. Sentiment analysis is carried out on news video comments, namely to see positive, neutral, or negative opinions addressed to the Governor of DKI Jakarta. From the sentiment analysis, the resulting sentiment can be used as a parameter of the success of a person's perfomance.

Sentiment Analysis is the process of extracting and computational processing of opinion sentences so that the sentiment information they contain is generated [6]. In extracting opinion sentences, the Lexicon-based method extracts all sentiment terms for a given text and assign the sentiment value using the sentiment lexicon while the machine learning-based method uses various machine learning algorithms that are often used including the Decision Tree Classifier, Neural Network, Naïve Bayes, Maximum Entropy, and Support Vector Machine (SVM) [7].

Sentiment classification is one of the functions of Text Mining [8]. The process of classifying comment data is carried out using the Support Vector Machine method. Support Vector Machine is a machine learning method based on the structural theory of statistical learning which aims to obtain a separator function or hyperplane as a separator for each class [9]. SVM was developed to solve classification problems because it has a better technical ability to generalize data compared to other techniques [10].

Same of the studies that have been caried out related to classification are research conducted by Faradhillah who has conducted a sentiment analysis on tweet data related to public services in the city of Surabaya to assess the perfomance of public service. Classifier uses 2 algorithms, namely Naïve Bayes and Support Vector Machine. The best classification model is obtained using the SVM algorithm with an accuracy than Naïve Bayes method [11].

Subsequent research conducted research on sentiment analysis on Twitter by Lavanya and Deisy. The data used from four different topics will be classified into three classes of sentiment, namely positive, neutral, and negative. Classifier using the multiclass SVM method and the results obtained an accuracy rate between 68%-70% [12].

The third research was conducted by Somantri, in a research journal discussing determining the choice to visit a place to eat or culinary with Sentiment Analysis technique. The research compared the perfomance of the feature extraction algorithm Information Gain (IG) and Chi Square on the best model results produced vy Support Vector Machine algorithm. The results showed that the best level of accuracy was produced by the SVM-IG [13].

Another research conducted by Risnantoyo, used 833 tweets related to the coronavirus. The tweet sentiment is classified into three classes including positive, negative, and neutral using three classification algorithms, namely Naïve Bayes, Support Vector Machine, and KNN to build a classification model. The results of the comparison of three classification algorithms show that the SVM algorithm provides better results than the other 2 algorithms with an accuracy value of 76.21% [14].

The last research by Nuansa, this study compare two algorithm methods, Naïve Bayes Classifier and Support Vector Machine in classfying positive and negative sentiment. The data uses taken from Twitter in form of tweet data totaling 1000. Using TF-IDF as weighting feature. The comparison between the two methods shows that the SVM with RBF kernel results with an accuracy of 87.80% better than the Naïve Bayes Classifier, namely an accuracy of 85.77% where both use tweet data with 10-fold cross validation [3].

Based on previous research, this research tries to classify the news video commentary data on Youtube into three classes, namely positive, neutral, and negative. Classification using the Support Vector Machine method and parameter optimization will be carried out.

.

METHOD

The stage of the sentiment analysis research methodology using the SVM method begin with data input data in the form of training data and testing data then processed in the preprocessing stages, feature weighting, classification to the testing and validation process. The general stages are described in Figure 1.

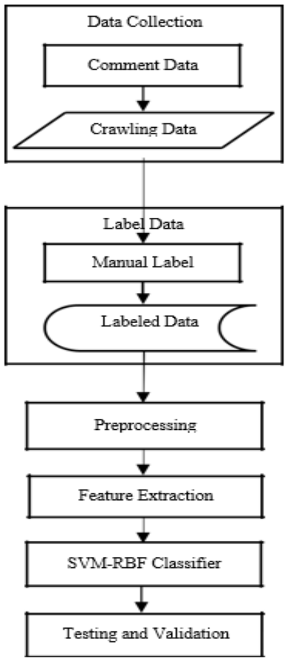

Figure 1: Proposed System for Sentiment Analysis

Data Collection

In the research, data collection was carried out by crawling comment data with API Keys obtained through the Google Developers website using the R language which was then saved in .csv format. The data used is the comment data from the news video on Youtube, which is obtained as much as 945 data. The data were then labeled by linguists into positive, neutral,and negative categories. The data will be processed to the preprocessing stage.

Preprocessing

The Preprocessing stages are presented in Figure 2. In this study, all labeled comment data will be processed to clean unstructured data into structured data or eliminate noise in comment data.

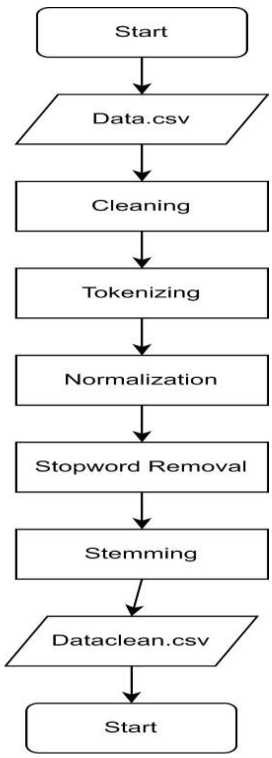

Figure 2: Preprocessing Process

TF-IDF

The feature weighting process uses the Term Frequency – Inverse Document Frequency (TF-IDF) method. Feature weighting combines two concepts, namely the frequency of occurrence of words and the frequency inverse that contains words in the document. The frequency with which the word appears in the comment data indicates the important information of the word in the document. The TF-IDF formula in equation 1 and equation 2 [15].

$$\boldsymbol{w_{dt}} = \boldsymbol{tf_{dt}} * \boldsymbol{idf_{ft}} \qquad (1)$$

$$\boldsymbol{w_{dt}} = \boldsymbol{tf_{ij}}\ \boldsymbol{x} \log \frac{\boldsymbol{D}}{\boldsymbol{df_i}} \qquad (2)$$

Where:

: The weight of term in the document
: Number of occurrences of term in the document
D : Number of all documents in database
: Number of documents containing term

The feature weighting process is presented in figure

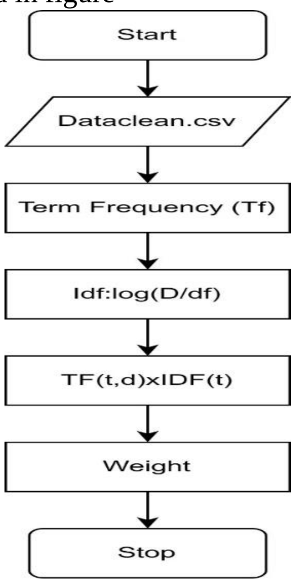

Figure 3: TF-IDF process

Classification

The classification process is presented with a flow of stages as in figure 4, namely, data has gone through the text preprocessing and feature weighting stages will then be classified

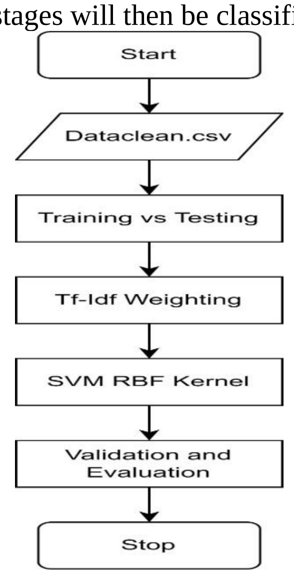

Figure 4: Classification Process

The classification process starts from the data that has gone through the preprocessing stage, then the data will be divide into 3 types, Training data, Validation data, and Testing data. Training data is data used to train or build models. Validation data is carried out to evaluate the model during model during the model construction process so that the model is able to recognize patterns generically [16]. After the model is built, then the model is tested using testing data which is to measure the extent to which the model succeeds in classifying correctly [17].

Classification using the Support Vector Machine method of the RBF kernel function. Selection of the right function is very important because it will determine the feature space in which the classifier function will be searched [18].

As long as the kernel functions are appropriate, the SVM method will operate completely without knowing what kind of map the SVM uses [4]. The RBF kernel is the most popular kernel among all the kernels in SVM. The RBF kernel function has the advantage that it automatically determines the value, the location of the center and can cover an infinite range of values [19]. where the RBF kernel is expressed in equation 3.

$$(\qquad) = \exp(- \| \qquad \|^2) , \quad > 0 \quad (3)$$

The RBF kernel effectively avoids overfitting by choosing the right value for the parameters C (cost) and γ (gamma) which can see how far the influence of training data [20]. Then the C (cost)

(gamma) parameters were optimized to improve the perfomance results of the SVM classification using the RBF kernel function. The combination of optimization parameter values can be seen in table 1.

Table 1: Parameter Value

Parameter	Nilai Optimasi					
	1	2	3		4	5
Cost (C)	0.1	0.5	1		10	100
Gamma (γ)	0.001	0.01	0.1		0.2	0.5

3.6 Test and Evaluation

Testing the classification model using the K-fold Cross Validation method. The method is used to classify new data that does not yet exist in the dataset[21]. In this study, the K values were 5-fold and 10-fold.

Evaluation of the classification model using the Confusion Matrix method is a matrix that displays a visualization of the perfomance of the classification algorithm by comparing the prediction classification againts the actual classification in the form of False Posittive (FP), True Positive (TP), False Negative (FN), and True Negative (TN).

Table 2: Confusion Matrix

Actual	Prediction	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

From the calculation of the confusion matrix, the perfomance value of the classification model will be found from the calculation of accuracy, precision, recall, and F-Measure.

RESULTS AND DISCUSSION

Dataset

The comment data are classes into three categories namely positive, neutral, and negatives. The process of labeling comment data is labeled by Indonesia linguists. The dataset used is 945 data with 302 positive, 379 negative, and 264 neutral.

Table 3. Dataset Labeled

Komentar	**Label**	**Inisial**
Untung jkrta dipimpin oleh seorang anisbaswedan.	Positif	2
Alhamdulillah libur 2 minggu :)	Netral	1
anis km gak ada otak nya kalau di suruh lockdown kasih makan rakyatmu anis.	Negatif	0

Cleansing

The initial process, namely the cleansing stage, will carry out several processes for changing uppercase letters to lowercase letters, removing punctuation and numbers, and removing symbols, and deleting URLs.

Table 4: Cleansing Process

Komentar	Output
Untung jkrta dipimpin olehseorang anis baswedan.	untung jkrta dipimpin olehseorang anis baswedan
Alhamdulillah libur 2 minggu :)	alhamdulillah libur minggu
anis km gak ada otak nya kalau di suruh lockdown kasih makan rakyatmu anis.	anis km gak ada otak nya kalau di suruh lockdown kasih makan rakyatmu anis

Tokenizing

Tokenizing aims to pronounce sentences into tokens or words.

Table 5: Tokenizing Process

Input	Output
untung jkrta dipimpin olehseorang anis baswedan	['untung', 'jkrta', 'dipimpin', 'oleh', 'seorang', 'anis', 'baswedan']
alhamdulillah libur minggu	['alhamdulillah', 'libur', 'minggu']
anis km gak ada otak nya kalau di suruh lockdown kasih makan rakyatmu anis	['anis', 'km', 'gak', 'ada', 'otak', 'nya', 'kalau', 'di', 'suruh', 'lockdown', 'kasih', 'makan', 'rakyatmu', 'anis']

Normalization

The stage of normalization is changing non-standart words, informal words, abbreviations into standard words that are in accordance with Kamus Besar Bahasa Indonesia (KBBI) such as the word "kpd" to "kepada". Language normalization is processed with a normalization dictionary which is compiled manually and uses standard word translations according to KBBI in .csv format.

Table 6: Normalization Process

Input	Output
['untung', 'jkrta', 'dipimpin', 'oleh', 'seorang', 'anis', 'baswedan']	['untung', 'jakarta', 'dipimpin', 'oleh', 'seorang', 'anis', 'baswedan']
['alhamdulillah', 'libur', 'minggu']	['alhamdulillah', 'libur', 'minggu']
['anis', 'km', 'gak', 'ada', 'otak', 'nya', 'kalau', 'di', 'suruh', 'lockdown', 'kasih', 'makan', 'rakyatmu', 'anis']	['anis', 'kamu', 'tidak', 'ada', 'otak', 'nya', 'kalau', 'di', 'suruh', 'lockdown', 'kasih', 'makan', rakyatmu', 'anis']

Stopword Removal

The stage of removing meaningless words and words that often appear but have no meaning such as conjuctions, substitute words for people and so on. This process uses the python stopword Indonesia library which was adapted from the F. Tala stopword.

Table 7: Stopword Removal Process

Input	Output
['untung', 'jakarta', 'dipimpin', 'oleh', 'seorang', 'anis', 'baswedan']	['untung', 'jakarta', 'dipimpin', 'anis', 'baswedan']
['alhamdulillah', 'libur', 'minggu']	['alhamdulillah', 'libur', 'minggu']
['anis', 'kamu', 'tidak', 'ada', 'otak', 'nya', 'kalau', 'di', 'suruh', 'lockdown', 'kasih', 'makan', 'rakyatmu', 'anis']	['anis', 'otak', 'lockdown', 'makan', 'rakyatmu', 'anis']

Stemming

The last stage where all words are processed to remove the affixes to the word so that it becomes the root word. The stemming process is carried out using the Python library, the Sastrawi library.

Table 8: Stemming Process

Input	Output
['untung', 'jakarta', 'dipimpin', 'anis', 'baswedan']	['untung', 'jakarta', 'pimpin', 'anis', 'baswedan']
['alhamdulillah', 'libur', 'minggu']	['alhamdulillah', 'libur', 'minggu']
['anis', 'otak', 'lockdown', 'makan', 'rakyatmu', 'anis']	['anis', 'otak', 'lockdown', 'makan', 'rakyat', 'anis']

TF-IDF

The implementation of the TF-IDF process uses the python library, namely Tfidf Vectorizer.

Table 9: TF-IDF Process

Term	Tf			idf	TF*IDF		
	D1	D2	D3		D1	D2	D3
terima kasih	1	0	0	0.602	0.602	0	0
anies	1	0	0	0.602	0.602	0	0
baswedan	1	0	0	0.602	0.602	0	0
anis	0	0	2	0.602	0	0	1.204
libur	0	0	1	0.602	0	0	0.602
otak	0	1	0	0.602	0	0.602	0
calon	1	0	0	0.602	0.602	0	0
publik	1	0	0	0.602	0.602	0	0

Classification Experiments

Classification experiments were carried out using four types of data models with different text preprocessing stages. Experiments were carried out to determine how much influence the preprocessing stage of the text by using the addition of a stoplist and the language normalization stage on the classification results. The four types of data models are as follows:

Testing using data divide by the ratio 80:20, 756 training data and 189 testing data. The test was carried out using 5-fold and 10 fold. the following are the results of the SVM classification test perfomance without parameter optimization for the four types of models.

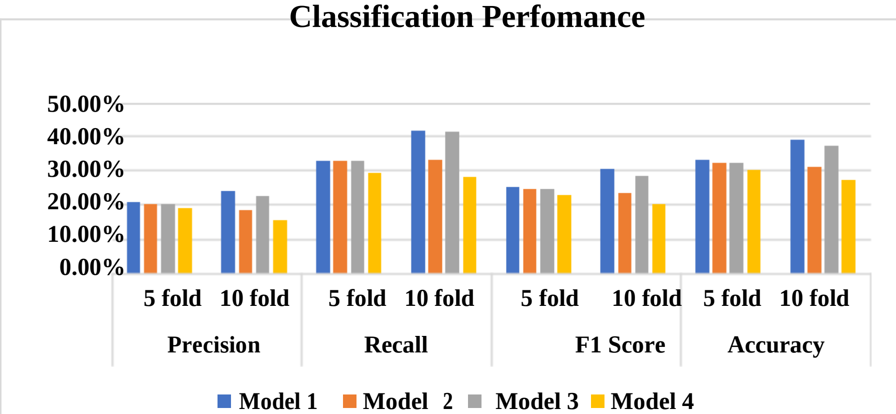

Figure 1: Classification Perfomance without Parameter Optimization

Figure 1 shows the perfomance value of each of the four data models, the results are not good. The average accuracy value is 31.75% on the 5-fold validation test and the accuracy value is 33.50% on the 10-fold validation test. The results are not good because the data has classification deviations or the classification is not in accordance with the class. From the poor perfomance results, parameter optimization will be carried out in the RBF kernel, namely the value of C and gamma (γ) in the four data models. Following are the results of the SVM classification perfomance testing, parameter optimization is carried out on four types of models.

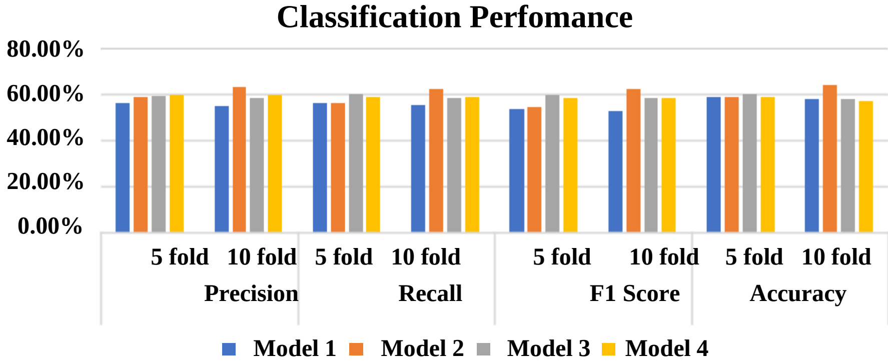

Figure 2: Classification Perfomance with Parameter Optimization

Figure 2 shows that after optimizing the parameter values of C and gamma (γ) in the RBF kernel, the results are quite good. The best classification perfomance in model 2 obtained an accuracy value of 64%, 63%

precision, 62.33% recall, and F1 Score of 62.33% with the 10-fold cross validation test. The results of the perfomance are quite good because the data are classified according to their class.

Analysis of Results

Based on the results of the evaluation carried out, the results of the comparison of the SVM classification values in the four experimental models without optimization of the parameters shown in Figure 4.

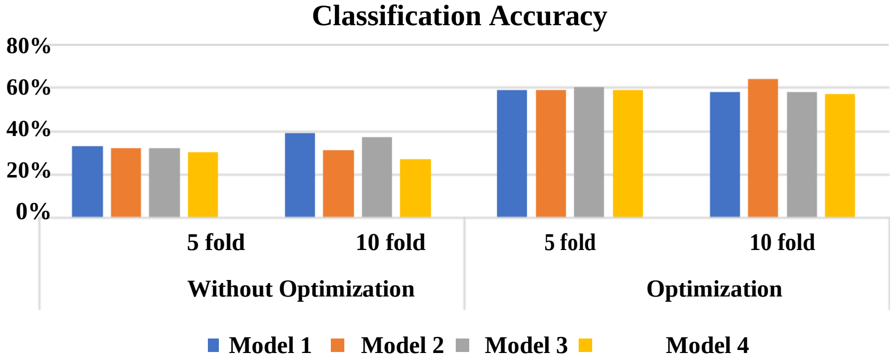

Figure 3: Comparison of Classification Accuracy

Figure 3 shows that the comparison of results of SVM Classification accuracy with the 5-fold validation test on four data models without parameter optimization has an average accuracy increase of 27% after optimization of the RBF kernel parameters. Whereas in the 10-fold validation test without parameter optimization, the accuracy increased by an average of 26%. In model 1 and model 2, the preprocessing stages both does not use normalization, but in model 2 involves adding a stoplist, the best accuracy value is 64%. This experiment proved that the preprocessing stage did not use a stoplist resulting in words that did not contain meaning or words that often appeared would have an effect on the classification process. In model 3, using the preprocessing stage of normalization and not using a stoplist, the accuracy value is quite good at 60% in the 5-fold validation test. The lowest accuracy value is obtained in model 4 by 57% on the 10-fold validation test.

This show thas the preprocessing stage which does not involve an additional stoplist, results in words that do not contain meaning such as names of people, place names, country names, and so on will be counted in the feature weighting process so that can affect the results of the classification process. In contrast to the results of the accuracy of model 4 which is lower than the acquisition of accuracy of model 2 where the preprocessing stages both involves an additional stoplist, but in model 4 the normalization stage is added, this is due to ambiguous words such as “kaya” if normalized to “seperti” but the word “kaya” is a formal word so that the word is omitted by the stoplist, this cause meaningful words that contain countless sentiments in the feature weighting process and affect the classification process.

Measuring the level of system effectiveness is not only seen from the value obtained, but also from the value of precision, recall, and F-Measure. The precision and recall values are used to see if any data deviations are causing a low value. A system is said to be optimal when the value is balanced between precision and withdrawal as seen from the F-Measure value. Figure 8 shows the comparison of the classification perfomance ratings of precision, recall, and F-Measure.

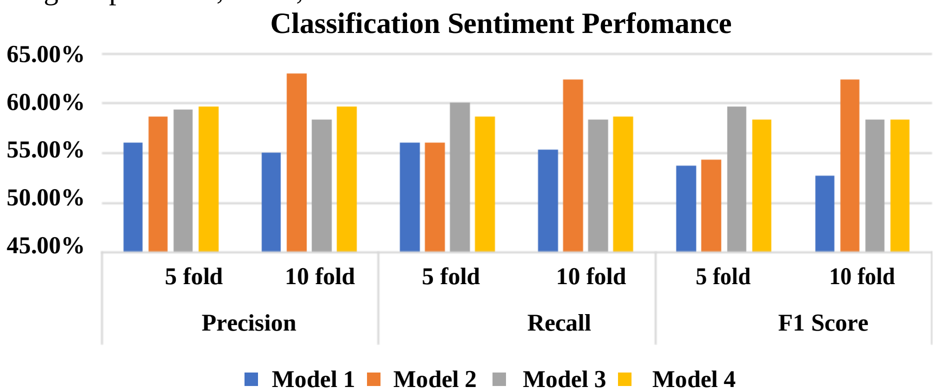

Figure 4: *Comparison of Classification Perfomance*

In Figure 8, results of the comparison of classification perfomance on 4 experimental models in the preprocessing stage are obtained. Model 1 perfomance obtained the lowest score, namely 56% precision, 56% recall, and 53.66% F-Measure on the 5-fold validation test as well as the 10-fold validation test with 55% precision, 55.33% recall, and 52.66% F-Measure compared to other models. This indicates that the rank can be used by the class in the relevant data. In Model 2, the best perfomance values are obtained among other models, 63% of precision, 62.33% recall, and 62.33% F-Measure in 10-fold validation test using a preprocessing stage experiment involving additional stoplist. Model 3 obtained a fairly good perfomance, 59.33% of precision, 60% recall, and 59.66% F-Measure using the preprocessing stage of normalization.

Whereas in Model 4 which involves the preprocessing stages of normalization and additional stoplists, the precision value is 59.66%, 58.66% recall, and F-Measure is 58.3% The cause of the low classification is due to several factors such as, the comment data is still unstructured, there are many satirical sentences that can make sentences ambiguous, sentimentally, or have multiple meanings resulting in deviation of sentiment calssification. There are still words that have not been normalized due to the lack of normalization dictionaries.

CONCLUSION

From the results fo the analysis sentiment classification regarding the policy of the governor of DKI Jakarta using the Support Vector Machine method after optimizing the RBF kernel parameter, the classification accuracy is quite good at 64%, 63% precision, 62.33% recall, and 62.33% F-Measure with the preprocessing stage using cleansing, tokenizing, stopword and stoplist, and stemming. Using the Support Vector Machine method with the RBF kernel function, parameter optimization is performed to produce a fairly good perfomance compared to not optimizing the RBF kernel parameters.

Based on the research that has been done, there are still some deficiencies that need to be fixed or developed. The suggestion given for the next research is the need to add a normalization dictionary due to the lack of translation such as colloquial words and it is advisable to try using other feature extraction methods such as the Word2Vec or Bag of Words methods to produce better classification perfomance.

REFERENCES

[1] Setiadi, “Pemanfaatan media sosial untuk efektifitas komunikasi,” vol. 3, no. 1, 2012.

[2] S. Khomsah, “Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia,” vol. 1, no. 10, pp. 648–654, 2021.

[3] E. Putra Nuansa, “Analisis Sentimen Pengguna Twitter Terhadap Pemilihan Gubernur Dki Jakarta Dengan Metode Naïve Bayesian Classification Dan Support Vector Machine,” 2017.

[4] B. Santosa, “Ide Dasar Support Vector Machine,” no. x, 2007.

[5] S. D. Harijiatno, “Analisis Sentimen Pada Twitter Menggunakan Multinominal Naive Bayes,” 2019.

[6] O. Ahlgren, “Research On Sentiment Analysis : The First Decade,” 2016 IEEE 16th Int. Conf. Data Min. Work. Res., 2016, doi: 10.1109/ICDMW.2016.94.

[7] B. S. Rintyarna, R. Sarno, and C. Fatichah, “Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks,” J. Big Data, 2019, doi: 10.1186/s40537-019-0246-8.

[8] G. A. Buntoro, “Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter,” vol. 2, no. 1, pp. 32–41, 2017.

[9] A. C. Najib, A. Irsyad, G. A. Qandi, and N. Aini, “Perbandingan Metode Lexicon-based dan SVM untuk Analisis Sentimen Berbasis Ontologi pada Kampanye Pilpres Indonesia Tahun 2019 di Twitter Abstrak,” vol. 4, no. 2, 2019.

[10] A. S. Nugroho, A. B. Witarto, and D. Handoko, “Support Vector Machine,” 2003.

[11] N. Y. A. Faradhillah, R. P. Kusumawardani, I. Hafidz, J. S. Informasi, and F. T. Informasi, "Eksperimen Sistem Klasifikasi Analisa Sentimen Twitter Pada Akun Resmi Pemerintah Kota Surabaya Berbasis Pembelajaran," Semin. Nas. Sist. Inf. Indones., pp. 16–23, 2016.

[12] K. Lavanya and C. Deisy, "Twitter Sentiment Analysis Using Multi-Class SVM," 2017 Int. Conf. Intell. Comput. Control Twitter, 2017.

[13] O. Somantri, D. Apriliani, J. T. Informatika, P. Harapan, and B. Tegal, "Support Vector Machine Berbasis Feature Selection untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal," J. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 5, pp. 537–548, 2018, doi: 10.25126/jtiik20185867.

[14] K. Risnantoyo, Ricky; Nugroho, Arifin; Mandara, "Sentiment Analysis on Corona Virus Pandemic Using Machine Learning Algorithm," JITE (Journal Informatics Telecommun. Eng. Available, vol. 4, no. 1, pp. 86–96, 2020.

[15] S. Mohammad, "A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification," 2nd IEEE Int. Conf. Eng. Technol., no. March, pp. 16–20, 2016.

[16] J. Wira and G. Putra, Pengenalan Konsep Pembelajaran Mesin dan Deep Learning, vol. 4. 2020.

[17] M. Lailiyah, "Sentiment Analysis Menggunakan Rule Based Method Pada Data Pengaduan Publik Berbasis Lexical Resources," Institut Teknologi Sepuluh Nopember, 2017.

[18] V. Apostolidis-afentoulis, "SVM CLASSIFICATION WITH LINEAR AND RBF KERNELS," no. July, 2015, doi: 10.13140/RG.2.1.3351.4083.

[19] K. S. Gopi, Arepalli Peda; Jyothi, R. Naga Sravana; Narayana, Lakshman; Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," 2019.

[20] S. Prangga, "Optimasi Parameter Pada Support Vector Machine Menggunakan Pendekatan Metode Taguchi untuk Data High-Dimensional," Institut Teknologi Sepuluh Nopember, 2017.

[21] I. D. Onantya and P. P. Adikara, "Analisis Sentimen Pada Ulasan Aplikasi BCA Mobile Menggunakan," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 3, no. 3, pp. 2575–2580, 2019.