Optimising the Particle Swam Optimazion Usage for Predicting Indonesia Presidential Election Result Period 2019-2024

Dinar Ajeng Kristiyanti[1]
STMIK Nusa Mandiri: Teknik Informatika
Jakarta, Indonesia
dinar@nusamandiri.ac.id

Normah[2]
STMIK Nusa Mandiri: Teknik Informatika
Jakarta, Indonesia
normah.nor@nusamandiri.ac.id

Abstract—Indonesia is a Democrat nation. A general election known as the PEMILU has become a tradition of the nation that is synonymous with political issues and leadership turnover. Social media is one place in expressing the opinions and aspirations of people including politics, Twitter is one of the social media used as a place for politicians including two couples of presidential candidate and vice president of INDONESIA in Campaign to win a vote in the elections of 17 April 2019. This research analyzes public opinion i.e. comments on Twitter accounts consisting of 4000 Indonesian language comments, consisting of 500 positive opinions and 500 negative opinions from each account @jokowi, @KyaiMarufAmin, @prabowo, @sandiuno into two categories of positive and negative opinions by comparing the text classifier model Naïve Bayes and SVM, and the implementation of The PSO algorithm to obtain optimal accuracy results. The results of the study show Prabowo Sandi won the prediction of presidential candidate with the best accuracy result of 77.00% acquired model Naïve Bayes + PSO, and 86.20% acquired model SVM + PSO, with an increase in accuracy 7.5% on model SVM, And 2.1% on the model Naïve Bayes when compared before done optimization with PSO algorithm.

Keywords—Naïve Bayes; Opnion Mining; Presidential Election; SVM, PSO.

I. INTRODUCTIVON

Indonesia is a Democrat nation. A general election known as the PEMILU has become a tradition of the nation that is synonymous with political issues and leadership turnover, such as president, legislative, governor, until the head of village. Where in a democracy country, elections are the main pillars of a process of accumulating the will of society. Elections as well as a democratic procedure for selecting leaders, (Dani, 2016). Presidential and vice president elections are held every five years.

Since 20 September 2019, two pairs of presidential candidates and vice-Presidents for the general election of 2019 by KPU RI are 01 for the candidate Joko Widodo and Ma'ruf Amin, and 02 for the candidate Prabowo Subianto, and Sandiaga Uno, with a campaign of 20 September 2018 to 19 April 2019, (Saubani, 2019).

Social media is one place in expressing the opinions and aspirations of people including politics. Social media widely used include Twitter, Facebook, Instagram, and the online mass media. It is also utilized by the support team of both presidential and vice presidents candidates in conducting a campaign to succeed and win the elections. The social media

activity during the period of the 2019 election campaign led to increased political tensions like the war on social media by forming public opinion, (Lazuardi, 2019).

Predicting the victory of presidential and vice presidents candidates is not easy, and should be done objectively and independently. Everyone has different views and same rights to express opinions or give opinions toward particular event. Public opinion is view of someone for certain problem comes out due to prior conversation with another person who may have an effect on the opinion given, (Kristiyanti, Umam, Wahyudi, Amin, & Marlinda, 2019).

Opinions expressed online by the general public through the media as Twitter can be followed up and used as a source of data in predicting the victory of presidential and vice presidential candidates through an analysis of positive and negative opinion sentiments.

Naïve Bayes and SVM are models that are widely used in sentiment analysis, but each model has a deficiency. Naive Bayes is a method that can be used for classification. In the process of text classification, but the lack of features sample in training data will provide an absolute zero probability value, causing errors in the classification process. Smoothing methods

are used to minimize the possibility of calculation errors on Naive Bayes if the features of test data did not occur oᵢn the training data, (Ramadhani, Indriani, & Nugrahadi, 2017).

SVM shows a better level of generalization than multilayer perceptron neural networks that are trained with backpropagation. The learning process in SVM is generally much faster than a neural network, but not in the testing process (recognition), because the speed of the testing process depends on the scale of the training set subset chosen as a support vector. Another open problem in SVM is its use in multiclass cases, because SVM requires a combination of various modules which results in high computational requirements in learning and testing, (Nugroho, 2008).

The PSO is initialize with a population of random solutions and searches for optima by updating generations, (Parimala & Nallaswamy, 2012). Particle Swarm Optimization (PSO) was motivated from the simulation of simplified social behavior of bird flocking, firstly developed by Kennedy and Eberhart. It is easy to implement with few parameters, and it is widely used to solve the optimization problems, as well as feature selection problem, (Yan & Jiao, 2016).

The purpose of this research is to predict who the President and vice president are elected based on the number of positive and negative opinions obtained through sentiment analysis, as well as testing how optimized the PSO when used to SVM and NB classifications in the case study of sentiment analysis contained in Twitter user comments in predicting President of the Republic of Indonesia 2019-2024.

Used Textblob for pre-processing, polarity, the polarity confidence calculation, and they validated the obtained results by SVM and Naïve Bayes using Weka; they reported the highest accuracy of Naïve Bayes with a 65.2% rate, which was 5.1% more than the SVM accuracy rate, Saha, Rajav and Ranjan at (Hasan, Moin, Karim, & Shamshirband, 2018)

On the research about Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization, This research concerns on binary classification which is classified into two classes. Those classes are positive and negative. The Positive class shows good message opinion; Otherwise the negative class shows the bad message opinion of certain movies. This justification is based on the accuracy level of SVM with the validation process uses 10-Fold cross validation and confusion matrix. The hybrid Partical Swarm Optimization (PSO) is used to improve the election of best parameter in order to solve the dual Optimization problem. The result shows the improvement of accuracy level from

71.87% to 77%, (Basari, Hussin, Ananta, & Zeniarja, 2013).

II. LITERATURE REVIEW

A. Opinion Mining

Opinion mining refers to the broad area of natural language processing, text mining, computational linguistics, which involves the computational study of sentiments, opinions and emotions expressed in text. Although, view or attitude based on emotion instead of reason is often colloquially referred to as a sentiment. Hence, lending to an equivalent for opinion mining or sentiment analysis. Stated that opinion mining has many application domains including accounting, law, research, entertainment, education, technology, politics, and marketing. In earlier days many social media have given web users avenue for opening up to express and share their thoughts and opinions, (Sarlan, Nadam, & Basri, 2015).

Sentiment analysis or opinion mining is a computational study of a person's opinions, behavior and emotions towards an entity. These entities can describe individuals, events or topics. The topic is most likely to be a review, (Kristiyanti, 2015a)

B. Naïve Bayes algorithm

The Naïve Bayes classification method is one of the supervised learning to provide a classification of text documents. Imagine that a document is described from a document class number which can be modeled as a set of words where an opportunity (independent) Word i of a document appears on a document from class C can be written as (Hadi, Bagus, & Hasan, 2017): $p(w_i/C)$

Then the probability that given document D contains all the words wi from class C is

$$P(D/C) \, \Pi_i \, p(w_i/C)$$

In classifying text we mark (tokenize) The document to be able to classify in its appropriate class. Using decision making rules with "Max a Posterior Probability" then obtained classifications as follows, (Hadi et al., 2017):

$$c_{MAP} = \arg\max_{c \in C}(P(c|d)) = \arg\max_{c \in C}\left(P(c)\prod_{1 \le i \le n}P(w_i|c)\right)$$

with:
d	= documents
w_i	= word ᵢ in document
c	= The set of class use Classification
$P(c/d)$	= conditional probability of Class C provided by the document

POLITEKNIK
GANESHA
Medan

P(c)   = Probabilitas prior from class c
$p(w_i/C)$ = conditional probability from the word w_i
          given class c

C. SVM

SVM has the advantage of being able to identify a separate hyperplane that maximizes margins between two different classes. But the Support Vector Machine has a lack of the problem of selecting the appropriate parameters or features. The selection of features and parameter tuning in SVM significantly affects the classification accuracy results, (Kristiyanti, 2015b).

SVM has become a popular method of classification and regression for linear and non linear problems. This method tries to find the optimal linear separator between the data with a maximum margin that allows positive values above the margin and negative values below it. This problem is described as a "quadratic programming optimization problem". Let $\{(x_{11},y_1),(x_{12},y_2),\ldots,(x_{mn},y_m)\}$ denote the set of training data, where x_{ij} denotes the occurrences of the events j in time $_i$, and y_i $\square$ {-1,1}. A support vector machine algorithm is solving the following quadratic problem, (Haddi, Liu, & Shi, 2013):

$$\min_{w,b} \frac{1}{2}w^2 + C * \sum_{i=1}^{n} \varepsilon_i \ st \ \forall \ i: y_i(\langle w, x_{ij}\rangle + b) \geq 1 - \varepsilon_i \quad \varepsilon \geq 0$$

Where $\square_i$ are the slack variables in which there are non-separable case and $C>0$ is the soft margin which controls the differences between margin b and the sum of errors. In other words, it performs a penalty for the data in the incorrect side of classification (misclassified), this penalty rises as the distance to the margin rises. w is the slope of the hyperplane which separates the data. The speciality of SVM comes from the ability to apply a linear separation on the high dimension non linear input data, and this is gained by using an appropriate kernel function. SVM effectiveness is often affected by the types of kernel function that are chosen and tuned based on the characteristics of the data

D. PSO

Feature Selection is optimization process to reduce a large set of the original great features in order to feature subset that relatively small and significantly improve the accuracy of classification for fast and effective, (Kristiyanti & Wahyudi, 2017).

The Particle Swarm Optimization algorithm (shortened as PSO) is a novel population-based stochastic search algorithm. It is an alternative solution to the complex non-linear optimization problem. PSO is an evolutionary computation technique developed by Dr. Ebernet and Dr. Kennedy in 1995 inspired by social behaviour of bird flocking or fish schooling. PSO is particularly attractive for feature selection in that particle swarms will discover the best feature combinations as they fly within the problem space. PSO has strong search ability in the problem space and can discover optimal solutions quickly. The PSO is initialize with a population of random solutions and searches for optima by updating generations, In PSO, the potential solutions, called particles, are "flown" through the problem space by following the current optimum particles. The members of entire population upheld through the search procedure so information socially shared among individuals to direct the search towards the best solution in the search space. In last few years, the PSO became frequently applied in wrappers, (Parimala & Nallaswamy, 2012).

III. RESEARCH METHODOLOGY

This research analyses the sentiment word contained in a Twitter comment on both the presidential candidate and the vice president. The comparison of the results between two models of text classification used are Naïve Bayes and SVM. Then resume with PSO implementation to optimize accuracy results.
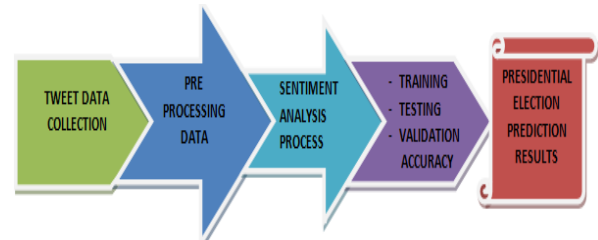

Figure 1. Stages of Research

Where after the comments on Twitter have been collected, the preprocessing stage is carried out by conducting Tokenization & Generate N-grams. In this study, the N-gram used is bigram, N = 3. Then the sentiment analysis process is carried out on each word contained in the comments. followed by training, as well as testing and validation on the accuracy results obtained. Then conclude the results of predictions of who the elected presidential and vice presidential candidates are 2019-2024.

IV. RESULTS AND DISCUSSION

The method which used is the method of experimental research with research phases as follows: Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Data Collection Phase

The data used in this study are in the form of public opinion regarding the two candidates for President and Vice President for the 2019-2024 period obtained from www.twitter.com from each Twitter account determined as candidates for President and Vice President for the 2019-2024 period, amounting to 4000 Indonesian language comments, consisting of 500 positive opinions and 500 negative opinions from each account @jokowi, @KyaiMarufAmin, @prabowo, and @sandiuno, this data was collected from the campaign period to the day of the Presidential Election on 17 April 2019.

B. Data Processing Phase

Preprocessing is done by *Tokenization & Generate N-gram*. Tokenization done to break down the tweet into some word or set of words that stand alone, (Buntoro, 2017). All the words in each document reviews are collected and punctuation marks, and symbols were removed, (Normah, 2019). The N-gram based character language model is a new model in natural language processing (Carpenter, 2005). It is derived from the N-gram language models. Instead of taking words as the basic unit, this model takes characters (letters, space, or symbols) as the basic unit in the algorithm. In this research, the N-gram used is bigram, N=3.

C. Sentiment Text Classification Phase Using Naïve Bayes Algorithm and SVM

After doing the preprocessing stage against the 4000 comments that have been obtained, the next step is to perform the analysis sentiment with the Rapidminner application, namely by classifying where comments are positive and which are negative Based on the number of words found in the comment. To get the maximum accuracy from classifying the data, a comparison of two models are Naïve Bayes and SVM..

K-fold cross-validation is a validation technique with initial data randomly split into k sections mutually exclusive or "fold", (Wahyudi & Kristiyanti, 2016).

The tests were conducted with 10 Cross Validation, and the evaluation was done by comparing the accuracy results measured by the Confusion matrix, and visualizing in the ROC curve.

ROC curve (Receiver Operating Characteristic) is a two-dimensional graph with false positives as horizontal lines and true positives to measure the performance difference method is used. ROC curves are used to measure the AUC (Area Under the Curve). AUC was calculated to measure the difference performances method was used. ROC curve divides the positive results in the y-axis and the negative results in the x-axis, so the larger the area under the curve, the better the prediction results, Witten & Frank dalam (Normah, 2019)

The following is a comparative evaluation of the classification results conducted by both models Naïve Bayes and SVM on public opinion about both the presidential and vice President candidate of the Republic of Indonesia:

TABLE I.   CONFUSION MATRIX MODEL SVM (JOKOWI MA'RUF)

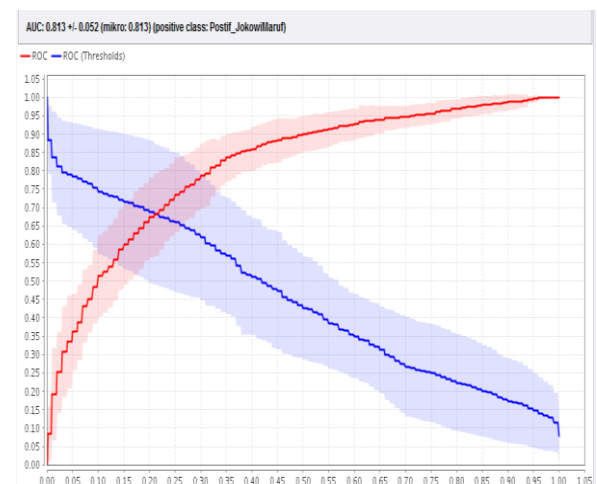| Naïve Bayes Accuracy: 73.30% +/- 4.97% (micro: 73.30%)			
	True Positive	True Negative	Class precision
Positive Prediction Jokowi Ma'ruf	626	150	79.54 %
Negative Prediction Jokowi Ma'ruf	374	840	69.19 %
Class Recall	62.60%	84.00%	

Figure 2. ROC Curve of SVM Jokowi Ma'ruf

TABLE II.   CONFUSION MATRIX MODEL SVM (PRABOWO SANDI)

Naïve Bayes Accuracy: 78.70% +/- 3.09% (micro: 78.70%)			
	True Positive	True Negative	Class precision
Positive Prediction Jokowi Ma'ruf	872	298	74.53 %

POLITEKNIK
GANESHA
Medan

Naïve Bayes Accuracy: 78.70% +/- 3.09% (micro: 78.70%)			
Negative Prediction Jokowi Ma'ruf	128	702	84.58 %
Class Recall	87.20 %	70.20 %	
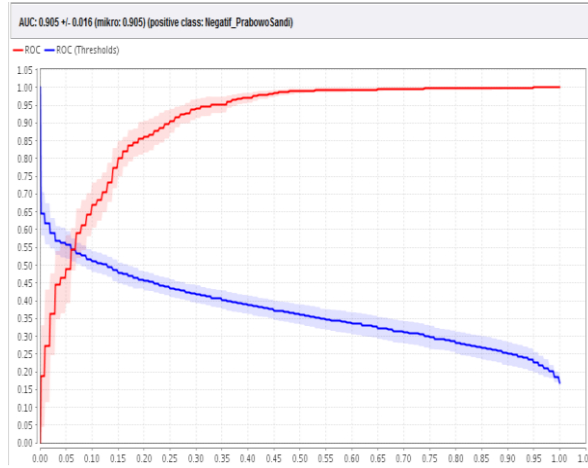
Figure 3. ROC Curve of SVM Prabowo Sandi

TABLE III.   COMPARISON OF ACCURACY RESULTS NAÏVE BAYES WITH SVM

	Accuracy		AUC	
	NB	**SVM**	**NB**	**SVM**
JM	66,25%	73,30%	0,681	0,813
PS	74,90%	78,70%	0,659	0,905

Table III. Shows that the SVM model generates a public opinion classification with a better level of accuracy than the Naïve Bayes model with the elected candidate, Prabowo Sandi. The best accuracy results generated by SVM is 78.70% with a AUC value of 0905 and is classified as excellent classification. While Naïve Bayes generates accuracy of 74.90% with a AUC value of 0659, which is classified as poor classification.

The study continued by adding the PSO algorithm, to find out how optimizing the accuracy results obtained by both models if the PSO algorithm was added.

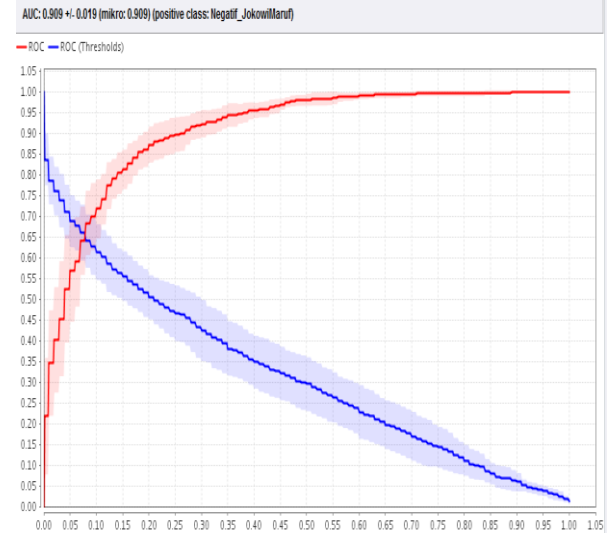| Naïve Bayes Accuracy: 82.85% +/- 2.77% (micro: 82.85%)			
	True Positive	True Negative	Class precision
Positive Prediction Jokowi Ma'ruf	790	133	85.59 %
Negative Prediction Jokowi Ma'ruf	210	867	80.50 %
Class Recall	79.00%	86.70%	

Figure 4. ROC Curve of SVM + PSO Jokowi Ma'ruf

TABLE V.   CONFUSION MATRIX MODEL SVM + PSO (PRABOWO SANDI)

Naïve Bayes Accuracy: 86.20% +/- 2.23% (micro: 86.20%)			
	True Positive	True Negative	Class precision
Positive Prediction Jokowi Ma'ruf	830	106	88.68 %
Negative Prediction Jokowi Ma'ruf	170	894	84.02 %
Class Recall	83.00 %	89.40 %	

TABLE IV.   CONFUSION MATRIX MODEL SVM + PSO (JOKOWI MA'RUF)

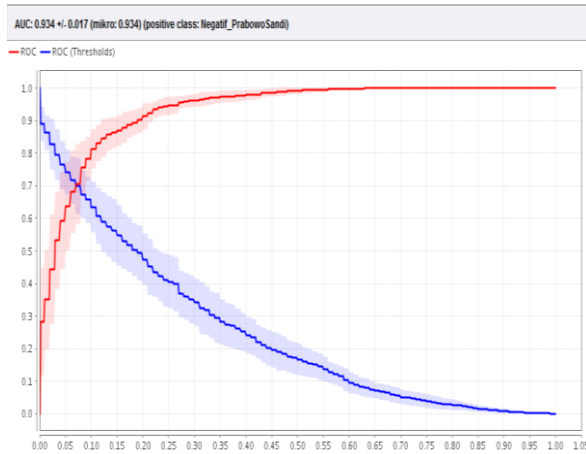**POLITEKNIK GANESHA** *Medan*

Figure 5. ROC Curve of SVM + PSO Prabowo Sandi

Table V. Pointing out that after the addition of the PSO algorithm in both models, it obtained the accuracy results that improved in both the accuracy results using Naïve Bayes and SVM. Model SVM resulted in public opinion classification with better level of accuracy compared with model Naïve Bayes and increased its accuracy after being added with PSO algorithm of 7.5% to 86.20% with the AUC value of 0.934 is With the elected candidate **Prabowo Sandi**.

The Naïve Bayes model also increased accuracy by 2.1% to 77.00% with auc value of 0.693 with the selected candidate Prabowo Sandi.

TABLE VI.   CONFUSION MATRIX MODEL NB + PSO (JOKOWI MA'RUF)

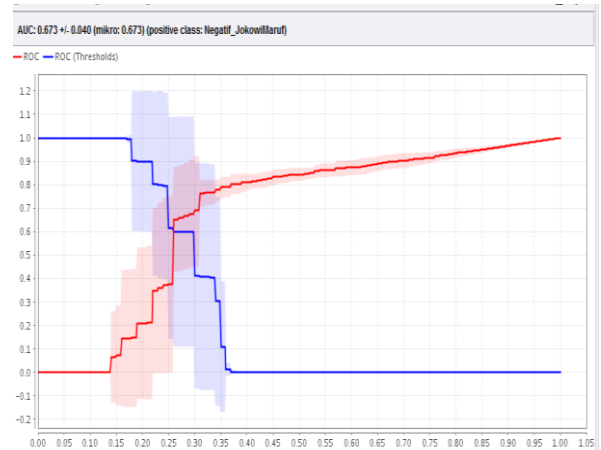| Naïve Bayes Accuracy: 73.60% +/- 3.14% (micro: 73.60%)			
	True Positive	True Negative	Class precision
Positive Prediction Jokowi Ma'ruf	714	242	74.69 %
Negative Prediction Jokowi Ma'ruf	285	758	72.61 %
Class Recall	71.40 %	75.80%	

Figure 6. ROC Curve of NB + PSO Jokowi Ma'ruf

TABLE VII.   CONFUSION MATRIX MODEL NB + PSO (PRABOWO SANDI)

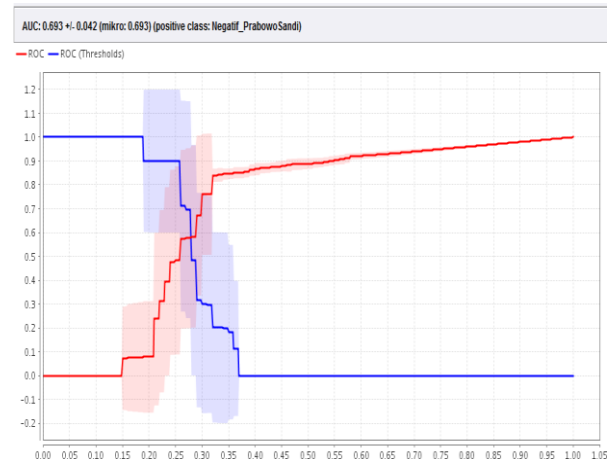| Naïve Bayes Accuracy: 77.00% +/- 1.66% (micro: 77.00%)			
	True Positive	True Negative	Class precision
Positive Prediction Jokowi Ma'ruf	716	176	80.27 %
Negative Prediction Jokowi Ma'ruf	284	824	74.37 %
Class Recall	71.60 %	82.40 %	

Figure 7. ROC Curve of NB + PSO Prabowo Sandi

TABLE VIII.   COMPARISON OF ACCURACY RESULTS NAÏVE BAYES WITH SVM

	Accuracy		AUC	
	NB+PSO	SVM+PSO	NB+PSO	SVM+PSO
JM	73,60%	82.85%	0,673	0,909

37

POLITEKNIK
GANESHA
Medan

PS	77,00%	86,20%	0,693	0,934

V. CONCLUSION

1. The classification of public opinion on both the presidential candidate and the vice President of the Republic of Indonesia 2019-2024 using two comparisons of the classification model namely Naïve Bayes and SVM is gaining the best accuracy by 74.90% Acquired model Naïve Bayes to Prabowo Sandi, and 78.70% acquired model SVM for Prabowo Sandi.

2. To improve the accuracy of the implementation of PSO algorithms applied on both models, and obtained the best accuracy result of 77.00% obtained by model Naïve Bayes + PSO for Prabowo Sandi, and 86.20% acquired SVM + PSO model for the Prabowo Sandi.

3. The final result is that the SVM model has a higher accuracy value compared to the Naïve Bayes model.

4. The PSO algorithm is able to optimize the accuracy generated by both models, namely an increase of 7.5% with SVM model, and 2.1% improvement in accuracy on the model Naïve Bayes.

REFERENCES

Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, *53*, 453–462. https://doi.org/10.1016/j.proeng.2013.02.059

Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *Integer Journal Maret*, *1*(1), 32–41. Retrieved from https://www.researchgate.net/profile/Ghulam_Buntoro/publication/316617194_Analisis_Sentimen_Calon_Gubernur_DKI_Jakarta_2017_Di_Twitter/links/5907eee44585152d2e9ff992/Analisis-Sentimen-Calon-Gubernur-DKI-Jakarta-2017-Di-Twitter.pdf

Dani, W. (2016). PENGERTIAN, FUNGSI DAN SISTEM PEMILIHAN UMUM. Retrieved August 10, 2019, from Komisi Pemilihan Umum Daerah Istimewa Yogyakarta website: https://diy.kpu.go.id/web/2016/12/19/pengertian-fungsi-dan-sistem-pemilihan-umum/

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, *17*, 26–32. https://doi.org/10.1016/j.procs.2013.05.005

Hadi, A. F., Bagus, D., & Hasan, M. (2017). Text Mining Pada Media Sosial Twitter Studi Kasus : Masa Tenang Pilkada Dki 2017 Putaran 2. *Seminar Nasional Matematika Dan Aplikasinya, 21 Oktober 2017 Surabaya, Universitas Airlangga*. Retrieved from http://math.fst.unair.ac.id/wp-content/uploads/2017/10/50-Dimas-Bagus__Sistem-Informasi_.pdf

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Mathematical and Computational Applications*, *23*(1), 11. https://doi.org/10.3390/mca23010011

Kristiyanti, D. A. (2015a). Analisis Sentimen Review Produk Kosmetik Melalui Komparasi Feature Selection. *Konferensi Nasional Ilmu Pengetahuan Dan Teknologi (KNIT)*, *2*(2), 74–81.

Kristiyanti, D. A. (2015b). Analisis Sentimen Review Produk Kosmetik Menggunakan Algoritma Support Vector Machine Dan Particle Swarm Optimization Sebagai. *Seminar Nasional Inovasi & Tren (SNIT) 2015 "Peluang Dan Tantangan Indonesia Dalam Menyikapi Afta 2015,"* 134–141. Retrieved from http://lppm.bsi.ac.id/SNIT2015/BidangA/A22-134-141_2015-SNIT-Dinar Ajeng Kristiyanti_ ALGORITMA SUPPORT VECTOR.pdf

Kristiyanti, D. A., Umam, A. H., Wahyudi, M., Amin, R., & Marlinda, L. (2019). Comparison of SVM Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter. *2018 6th International Conference on Cyber and IT Service Management, CITSM 2018*, (Citsm), 1–6. https://doi.org/10.1109/CITSM.2018.8674352

Kristiyanti, D. A., & Wahyudi, M. (2017). Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review. *2017 5th International Conference on Cyber and IT Service Management, CITSM 2017*. https://doi.org/10.1109/CITSM.2017.8089278

Lazuardi, G. (2019). Aktivitas Medsos Sebabkan Tensi Panas di Pemilu 2019. Retrieved July 20, 2019, from Tribunnews.com website: https://www.tribunnews.com/nasional/2019/05/21/pengamat-aktivitas- medsos-sebabkan-tensi-panas-di-pemilu-2019

Normah, N. (2019). Naïve Bayes Algorithm For Sentiment Analysis Windows Phone Store Application Reviews. *SinkrOn*, *3*(2), 13. https://doi.org/10.33395/sinkron.v3i2.242

Nugroho, A. S. (2008). Support Vector Machine : Paradigma Baru Dalam. *Neural Networks*, *8*(Support Vector Machine), 92–99.

Parimala, R., & Nallaswamy, R. (2012). Feature Selection using a Novel Particle Swarm Optimization and It's Variants. *International Journal of Information Technology and Computer Science*, *4*(5), 16–24. https://doi.org/10.5815/ijitcs.2012.05.03

Ramadhani, R. A., Indriani, F., & Nugrahadi, D. T. (2017). Comparison of Naive Bayes smoothing methods for Twitter sentiment analysis. *2016 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2016*, 287–292. https://doi.org/10.1109/ICACSIS.2016.7872720

Sarlan, A., Nadam, C., & Basri, S. (2015). Twitter sentiment analysis. *Conference Proceedings - 6th International Conference on Information Technology and Multimedia at UNITEN: Cultivating Creativity and Enabling Technology Through the Internet of Things, ICIMU 2014*, (November 2014), 212–216. https://doi.org/10.1109/ICIMU.2014.7066632

Saubani, A. (2019). *KPU Tetapkan Dua Pasangan Calon Pilpres*. 2019.

Wahyudi, M., & Kristiyanti, D. A. (2016). Sentiment analysis of smartphone product review using support vector machine algorithm-based particle swarm optimization. *Journal of Theoretical and Applied Information Technology*, *91*(1), 189–201.

Yan, P., & Jiao, M. H. (2016). An improved particle swarm optimization for global optimization. *Proceedings of the 28th Chinese Control and Decision Conference, CCDC 2016*, *8*, 2181–2185. https://doi.org/10.1109/CCDC.2016.7531347