USTADZ ABDUL SOMAD LECTURE SENTIMENT ANALYSIS USING SUPPORT VECTOR MACHINE ALGORITHM COMPARISON OF COMPARATIVE FEATURES SELECTION

Dedi Aridarma[1(*)], Rifki Sadikin [2], Bobby Suryo Prakoso[3], Heru Sukma Utama[4],

Master of Computer Science Study Program
Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri
www.nusamandiri.ac.id
[1]dediaridarma8@gmail.com; [2]rifki.sadikin@gmail.com; [3]14002107@nusamandiri.ac.id;
[4]14002126@nusamandiri.ac.id
(*) Corresponding Author

Abstract— Religious lectures are activities that are identical to the presentation of religion, delivered orally by someone who has more religious knowledge than that delivered to the public with the aim that the knowledge delivered can be understood. Ustadz Abdul Somad is one of the lecturers who is known by various levels of society, but not all of his lectures can be accepted by the public there are likes or dislikes obtained from various positive and negative comments on social media. To overcome this problem Sentiment Analysis is used by applying the Support Vector Engine Algorithm method. The purpose of this study is to compile using the Particle Swarm Optimization and Information Gain selection features. The results for Particle Optimization Feature Selection produce 80.57% Accuracy, 85.45% Precision, and 79.52% Recovery, Advantages of Feature Selection Information resulting in 79.78% Accuracy, 78.47% Precision, and a recall of 78, 43%, Based on the results of this study it can be concluded by using the Particle Swarm Optimization feature selection in terms of accuracy when compared to using the Information Gain selection feature.

Keyword*: Support Vector Machine, Particle Swarm Optimization*, *Information Gain*

Abstrak*— Ceramah agama merupakan kegiatan yang identik dengan penyajian keagamaan, disampaikan secara lisan oleh seorang memiliki ilmu agama lebih kemudian disampaikan kepada masyarakat dengan tujuan ilmu yang disampaikan dapat dipahami. Ustadz Abdul Somad salah satu penceramah yang sudah dikenal berbagai lapisan masyarakat, akan tetapi ceramah beliau tidak semua dapat diterima oleh masyarakat ada yang suka atau tidak suka yang didapat dari berbagai macam komentar positif maupun negatif yang ada di media sosial. Untuk memecahkan permasalahan tersebut digunakan Analisis Sentimen dengan menerapkan metode Algoritma Support Vector Machine. Adapun tujuan penelitian ini ialah melakukan kompirasi menggunakan selection fiture Particle Swarm Optimization dan Information Gain. Hasil penelitian untuk Selection Feature Particle Swarm Optimization menghasilkan Accuracy sebesar 80,57%, Precision sebesar 85.45%, dan Recall sebesar 79,52%, Untuk Selection Feature Information Gain menghasilkan Accuracy sebesar 79.78%, Precision sebesar 78.47%, dan Recall sebesar 78,43%, Berdasarkan hasil penelitian ini dapat disimpulkan dengan menggunakan feature selection Particle Swarm Optimization lebih baik dalam tingkat akurasi apabila dibandingkan dengan menggunakan feature selection Information Gain.*

Kata Kunci : *support vector machine, particle swarm optimization*, *information gain*

INTRODUCTION

Lectures can be held anytime and anywhere. There are no specific conditions or specific places to do it (Fitriani, 2017). The time and place to conduct lectures are not limited and anyone can give a lecture (Sakti, 2016) .

From the problems regarding the rejection of Ustadz Abdul Somad's lectures in several regions, there are many different public opinions delivered through comments on social media, it is necessary to do a sentiment analysis approach so that it can find out the public's perspective on blocking the lecture of Ustadz Abdul Somad (Pratama et al., 2018). The method used is proposed by many researchers to be used in sentiment analysis using the Support Vector Machine algorithm (Chandani et al., 2015). The purpose of this study is to optimize and improve the value results with a good degree of

DOI: 10.33480/pilar.v16i1.702

accuracy with the addition of Information Gain feature selection (Bimantoro & ‘Uyun, 2017) and *Particle Swarm Optimization*(PSO) (Noor, 2018).

Related research is about the use of cosmetics products that are circulating in the market, not all have good quality, and according to consumer needs. From several reviews of consumer opinion that have bought and tried cosmetic products, a sentiment analysis approach is used using the Support Vector Machine (SVM) method and the addition of Particle Swarm Optimization and Genetic Algorithm selection features (Kristiyanti, 2015) so that it can help consumers in making choices for buying cosmetic products..

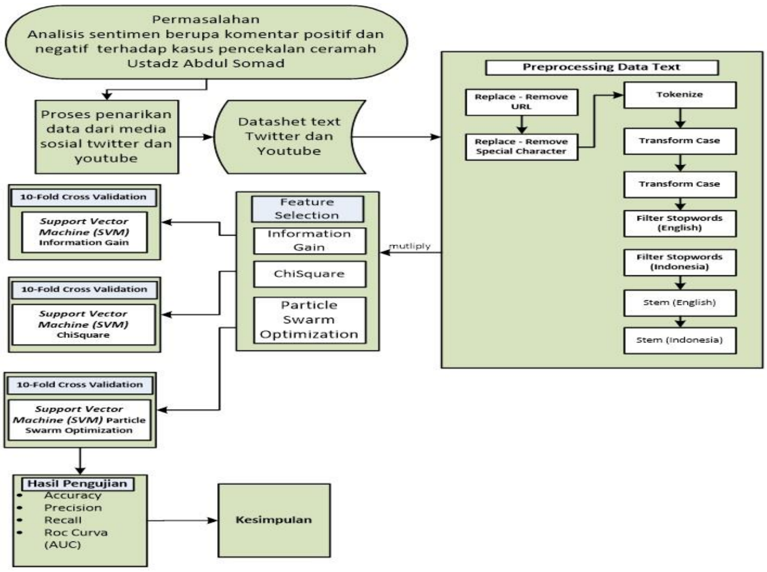
Sentiment analysis research conducted (Arifin, 2016) through public opinion about tourist destinations. The method used is the SVM algorithm and the use of the SVM algorithm with the addition of two feature selections. So that it can be found the best value comparison between the use of SVM and with the use of SVM in combination with feature selection of Particle Swarm Optimization and Genetic Algorithm.

Culinary connoisseurs in the Tegal area provide various comments through social media. Various opinions about culinary places in Tegal have carried out sentiment analysis by applying the SVM algorithm model with the addition of PSO selection features, to provide appropriate recommendations for culinary connoisseurs by displaying the results of optimal accuracy obtained from using the SVM model using PSO feature selection.(Somantri & Apriliani, 2018).

Research sentiment analysis of data taken about the earthquake in North Sumatra. This research uses a Support Vector Machine algorithm based on PSO feature selection (Indrayuni, 2016). The purpose of this study is to obtain the actual predicted value of the earthquake results by comparing the use of Support Vector Machine with the use of Support Vector Machine using PSO feature selection by calculating the average error that occurs through the amount (RMSE)(Noor, 2018).

MATERIALS AND METHODS

To support this research, it is explained about the framework of thinking about how and where the data in this study was obtained, and explains the flow of the data process so that it can produce the research objectives described in Figure 1 below.


Permasalahan
Analisis sentimen berupa komentar positif dan negatif terhadap kasus pencekalan ceramah Ustadz Abdul Somad
Proses penarikan data dari media sosial twitter dan youtube
Datashet text Twitter dan Youtube
Preprocessing Data Text
Replace - Remove URL
Replace - Remove Special Character
Tokenize
Transform Case
Transform Case
Filter Stopwords (English)
Filter Stopwords (Indonesia)
Stem (English)
Stem (Indonesia)
mutliply
Feature Selection
Information Gain
ChiSquare
Particle Swarm Optimization
10-Fold Cross Validation
Support Vector Machine (SVM) Information Gain
10-Fold Cross Validation
Support Vector Machine (SVM) ChiSquare
10-Fold Cross Validation
Support Vector Machine (SVM) Particle Swarm Optimization
Hasil Pengujian
Accuracy
Precision
Recall
Roc Curva (AUC)
Kesimpulan

Source: (Aridarma, 2019)

Figure 1: Research Thinking Framework

The data used were obtained as many as 633 data in the form of public comments contained on Twitter and Youtube about the banning of a lecture by Abdul Somad, taken from December 2017 to September 2018. Data retrieval is done by crawling and scrapping data. The data obtained is stored in the form of Microsoft Excel format, the data obtained will be labeled. as shown in table 1 below.

Table 1. Data Text Comments

source	comment	Sentiment
youtube	Jadi Ustad Abdul Somad itu gk usa ikut2 politik dssar pembuat ujaran kebenciyan	negative

source	comment	Sentiment
youtube	Ustad Abdul Somadz penyebar FITNAH!!!!	negative
youtube	sehat selalu buat pak Ustad Abdul Somad jangan berhenti berdakwah pak Ustad Abdul Somad	positive
youtube	Sabar.iklas..ustad..semoga dalam dakwa.yg tulus....allah melindungi ustad somad	positive
youtube	Mantaappp Ustadz ,,, berkarakter , cerdas , berwibawa . Muslim ga boleh lembek	positive
twitter	Modar kowe yg nolak Ustad Abdul Somad @id	negative
twitter	UAS seorang ulama, wajib konfirmasi perihal tsb, sebaik koordinasi dengan panitia & kepolisian, biar semua clear	positive
twitter	Sabar y pak ustad..Allah slalu bersama org yg baik seperti pak ustad somad..smg org ini menyadari kesalahanya,.dan d beri hidayah oleh Allah SWT.	positive
twitter	Insya Allah ceramahnya lbh sejuk dan membawa kepada kedamaian	positive
twitter	Pak...itu cm segelintir orng...saya orang hindu malu sm pak usd...atas insiden d bali.	positive
DST	Bali itu neraka dunia.. Cbak lah orang bali yg mengusir ulama ke aceh ..kalau bukan mati masok rumahh sakit y	negative

Source:(Aridarma, 2019)

This study uses the Support Vector Machine algorithm for the classification process. Support Vector Machine is a technique for making predictions, both predictions in the case of regression and classification (Awais et al., 2015).

Theoretically analyzed, use a concept of computational learning and simultaneously produce a good job (F. C. Li, 2009). Training data need to be used to estimate the function of classification problems. The classification function is described as follows

$f : R^N \rightarrow \{1,-1\}$ where k N is a dimensional pattern. i x and the class label i y, where

$(x\,, y\,), \ldots\ldots, (x\,, y\,) \in\ R^N\, x\{1,-1\}$..........................(1)

According to the above equation, the SVM classifier must meet the following formula:

$W^T\phi\,(x_i) +\ b \geq +1\ if\ y_i = +1$ ....(2)

$W^T\phi\,(x_i) +\ b \leq -1\ if\ y_i = -1$..............................(3)

Which is equivalent to

$y_i[W^T\phi\,(x_i) +\ b] \geq 1, i = 1, 2\ \ldots k$..........................(4)

Non-linear function ϕ maps the original space to an as a high-dimensional feature space. The hyperplane will be built by the inequalities that are produced and are defined as follows:

$W^T\phi\,(x_i) + b = 0$..(5)

Two classes will be distinguished by an optimal hyperplane A data comparison process is needed to avoid a larger range of numbers that dominates a smaller range of numbers. This can be useful to avoid difficulties during numerical calculations and help improve accuracy. Each variable can be distinguished linearly in the range [0,1] by the normalized formula process..

$V^{new} = \frac{v-min^v}{max^v-min^v}$ .... (6)

v : original value

V^{new} : scale value.

max^v : upper limit of feature values

min^v : lower limit of feature values.

In addition to using the Support Vector Machine method, this study adds some features, Swarm Optimization (PSO and Information Gain.

PSO is a method of population search, derived from research for the movement of organisms from groups of birds or fish, such as genetic algorithms (Chandra, 2018), PSO searches using the population (swarm) of individuals (particles) (H. Li et al., 2016) that are updated from iteration to iteration (Fei et al., 2009).

To find the optimal solution, each particle moves in the direction (Ma et al., 2019) of the best previous position (pbest) and the best global position (gbest). The speed and position of the particles can be updated as follows:

$v_{i,m} = W.v_{i,m} + C_1{}^* R^*\ (pbest_{i,m} - x_{i,m}) +$
$C_2\ {}^* R^*\ (gbest_{i,m} - x_{i,m})x_{id} = x_{i,m} + v_{i,m}$ ....(7)

t : shows counter iteration

v_{ij} : particle velocity i on the jth dimension (the value is limited between$[-v_{max}, v_{max}]$

p_{ij} : the position of particle i on j dimensions (limited value) $[-p_{max}, p_{max}]$

$pbest_{ij}$: the position of pbest particle i in the jth dimension

$gbest_{ij}$: gbest position of the jth dimension

w : heavy inertia (balancing global exploration and local exploitation)

$rand_1\ dan\ rand_2$: random functions in the range [0, 1]

β : constraint factor for controlling the weight speed (value is 1)

$c_1\ dan\ c_2$: personal and social learning factors (value 2)

Information gain (Pristyanto et al., 2019) is one method of feature selection that is widely used to determine the limits of the importance of an attribute (Maulana & Karomi, 2015). The information gain value is obtained from the entropy value before separation reduced by the entropy value after separation. Measurement of this value is only used as an initial step for determining the attributes that will be used or discarded (Bimantoro & 'Uyun, 2017). Attributes that meet the weighting criteria that will be used in the classification process of an algorithm, feature selection with information gain is done in 3 stages, namely (Mwadulo, 2016):

1. Calculate the information gain value for each attribute in the original dataset.
2. Determine the desired limit (threshold). This will allow attributes with a weight equal to the limit or greater to be retained and remove attributes that are below the limit.
3. The dataset is improved by reducing the attributes.

Measurement of this attribute was first spearheaded by Claude Shannon in information theory and written asi (Shaltout et al., 2014) :

$$info(D) = -\sum_{i=1}^{c} p_i log_2(p_i) \qquad ....(8)$$

The description of the formula is:

c : the number of values that exist in the target attribute(number of classification classes)

p_i : total until class i

$$info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} x\ info(D_j) ....(9)$$

The description of the formula is:

A : attribute

$|D|$: the sum of all data samples

$|D_j|$: number of samples for value j

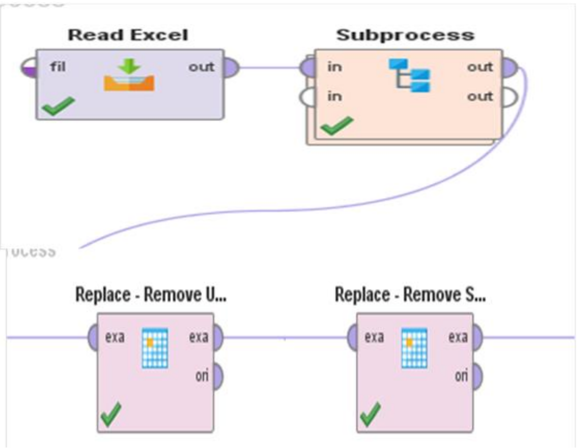
v : a possible value for attribute A

Furthermore, the information gain value that will be used is calculated using the formula below:

$$Gain\ (A) = |info(D) - info_A(D)| \qquad ....(10)$$

RESULTS AND DISCUSSION

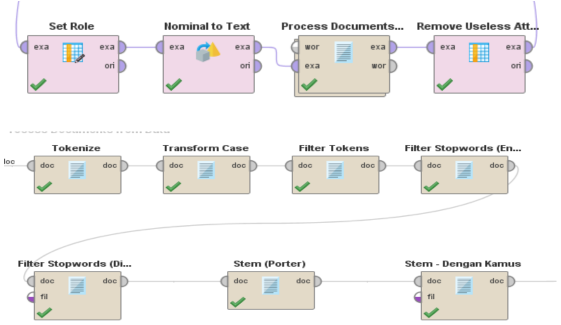
At this stage the datasheet that has been prepared in the form of an Excel file is imported into the data using the Read Excel operator. After that the data is processed using the Subprocess operator, in the Subprocess operator the Replace-Remove URL operator is used to remove links that are still in the comments and the Operator Replace Remove Special Character is useful for deleting the characters contained in, an explanation of this process is in Figure 2, below this:



Source :(Aridarma, 2019)

Figure 2: Process Read Excel, Subprocess, Replace - Remove URL, Replace Remove Special Character

In determining what attributes will be labeled, the Operator Set Role is used and to change the nominal value into text form, the Nominal to Text Operator is used, In the Process Documents Operator from Data, there are stages of the process such as Tokenize which is useful for separating the text into sections with space and punctuation restrictions, Transform Case is the process of changing the text entered into all lowercase letters, Filter Tokens process to limit the minimum and maximum number of characters, Filter Stopwords process to eliminate text that is incompatible with existing text and has been determined in the list of texts contained in the stopword, and Stemming the process of finding the root word of a word by removing all affixes (Rizqi et al., 2017). After the Documents from the Data, the process is performed on the datasheet, the data will be processed using the Remove Useless Attributes operator which aims to eliminate the unused attributes. This process can be seen in Figure 3.

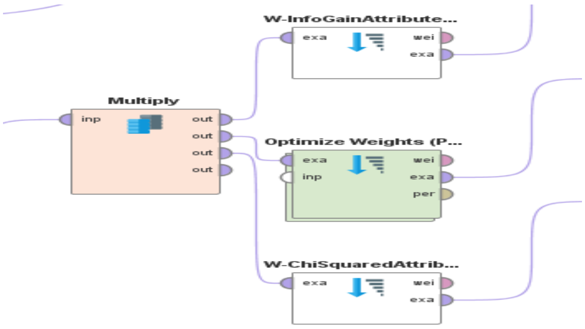


Source : (Aridarma, 2019)

Figure 3. Process Set Role, Nominal to Text, Process Documents from Data, Tokenizer, Transform Case, Filter Tokens, Filter Stop Words, Stemming, Remove Useless Attributes

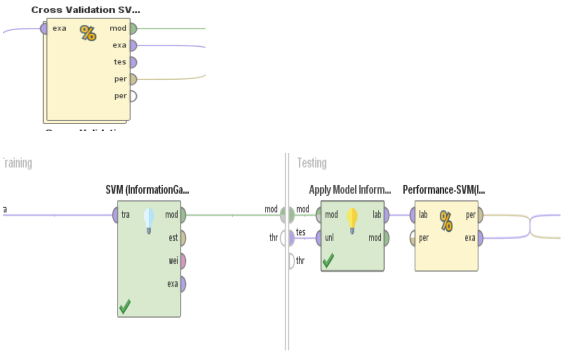
DOI: 10.33480/pilar.v16i1.1165

The Multiply process is then performed to retrieve objects from the input port and send a copy to the output port, which will be continued to determine the selection of features used. Feature selection uses PSO and InformationGain. Each connected port makes an independent copy. Furthermore, the process of cross-validation with k-10 fold cross-validation can carry out standard testing carried out to predict the error rate, can be seen in Figure 4 below.



Source : (Aridarma, 2019)
Figure 4 Feature Selection Process

The next process is validated using the K-fold cross-validation process. K-fold cross-validation is a technique to estimate the performance of the training model that has been built. This method divides training data and testing data as much ask parts of data. The function of k-fold cross-validation is so that there is no overlapping of the testing data (Jiawei et al., 2013). In k-fold cross validation there is a Support Vector Machine algorithm process. The K-fold cross-validation process can be seen in Figure 5 below.



Source: (Aridarma, 2019)
Figure 5. K-fold cross-validation process, Support Vector Machine, Apply Model, Performance

The results of modeling the Vector Machine support algorithm method by compiling Particle Swarm Optimization, ChiSquare, and Information Gain selection features are shown in table 2 below

Table 2 Research Results

Model	*Accuracy*	*Recall*	*Precision*
Support Vector Machine Selection Fiture Particle Swarm Optimization	80,57%	85.45%	79,52%
Support Vector Machine Selection Fiture Information Gain	79.78%	78.47%	78,43%

Sumber: (Aridarma, 2019)

CONCLUSION

Application of the Support Vector Machine algorithm method by compiling selection features on public comments for Ustad Abdul Somad's lecture resulted in a level using Particle Swarm Optimization obtained an Accuracy value of 80.57%, Precision of 85.45%, and Recall of 79.52% and for Selection Feature Information Gain produces Accuracy of 79.78%, Precision of 78.47%, and Recall of 78.43%, It can be concluded using the Particle Swarm Optimization feature selection is better in the level of accuracy when compared to using the Information Gain selection feature.

REFERENCE

Aridarma, D., Sadikin, R., Prakoso, B. S., & Utama, H. S. (2019). *Independent Research Final Report: Ustadz Abdul Somad Lecture Sentiment Analysis Using Support Vector Machine Algorithm Comparison Of Comparative Features Selection*.

Arifin, Y. T. (2016). Komparasi Fitur Seleksi Pada Algoritma Support Vector Machine Untuk Analisis Sentimen Review. *Jurnal Informatika*, *3*(2), 191–199.

Awais, M., Altaf, B., Member, S., & Yoo, J. (2015). *Epileptic Seizure Classification SoC Using a Non-Linear Support Vector Machine*. 1–12.

Bimantoro, D. A., & ‘Uyun, Is. (2017). *PENGARUH PENGGUNAAN INFORMATION GAIN UNTUK*. *2*(1), 42–52.

Chandani, V., Wahono, R. S., & Purwanto, P. (2015). Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film. *Journal of Intelligent Systems*, *1*(1), 56–60. http://www.journal.ilmukomputer.org/index.php?journal=jis&page=article&op=view&path%5B%5D=10

Chandra, H. A. (2018). Particle Swarm Optimization Pada Metode Knn Euclidean Distance

DOI: 10.33480/pilar.v16i1.702

Berbasis Variasi Jarak Untuk Penilaian. *Technologia: Jurnal Ilmiah*, *9*(1), 59. https://doi.org/10.31602/tji.v9i1.1103

Fei, S.-W., Miao, Y.-B., & Liu, C.-L. (2009). Chinese Grain Production Forecasting Method Based on Particle Swarm Optimization-based Support Vector Machine. *Recent Patents on Engineering*, *3*(1), 8–12. https://doi.org/10.2174/187221209787259947

Fitriani, W. (2017). *Pemanfaatan Kultum dalam Pembinaan Akhlak Siswa di SMPN 1 Indrapuri* [UIN Ar-Raniry Banda Aceh]. https://repository.ar-raniry.ac.id/id/eprint/542/

Indrayuni, E. (2016). Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization. *EVOLUSI : Jurnal Sains Dan Manajemen*, *4*(2). https://doi.org/10.31294/EVOLUSI.V4I2.697

Jiawei, H., Kamber, M., & JianPei. (2013). *Data mining Concepts and Techniques Preface and Introduction*.

Kristiyanti, D. A. (2015). Analisis Sentimen Review Produk Kosmetik Melalui. *Konferensi Nasional Ilmu Pengetahuan Dan Teknologi*, 69–76. http://konferensi.nusamandiri.ac.id/prosiding/index.php/knit/article/view/33

Li, F. C. (2009). Comparison of the primitive classifiers without features selection in credit scoring. *Proceedings - International Conference on Management and Service Science, MASS 2009*, 1–5. https://doi.org/10.1109/ICMSS.2009.5302730

Li, H., Feng, X., Cao, L., Li, E., Liang, H., & Chen, X. (2016). A New ECG Signal Classification Based on WPD and ApEn Feature Extraction. *Circuits, Systems, and Signal Processing*, *35*(1), 339–352. https://doi.org/10.1007/s00034-015-0068-7

Ma, P., Zhang, H., Fan, W., & Wang, C. (2019). Early fault diagnosis of bearing based on frequency band extraction and improved tunable Q-factor wavelet transform. *Measurement: Journal of the International Measurement Confederation*, *137*, 189–202. https://doi.org/10.1016/j.measurement.2019.01.036

Maulana, M. R., & Karomi, M. A. Al. (2015). INFORMATION GAIN UNTUK MENGETAHUI PENGARUH ATRIBUT TERHADAP KLASIFIKASI PERSETUJUAN KREDIT. *JURNAL LITBANG KOTA PEKALONGAN*, *9*(1), 113–123. https://jurnal.pekalongankota.go.id/index.php/litbang/article/view/28

Mwadulo, M. W. (2016). A Review on Feature Selection Methods For Classification Tasks. *International Journal of Computer Applications Technology and Research*, *5*(6), 395–402. https://doi.org/10.7753/IJCATR0506.1013

Noor, A. (2018). Perbandingan Algoritma Support Vector Machine Biasa dan Support Vector Machine berbasis Particle Swarm Optimization untuk Prediksi Gempa Bumi. *Jurnal Humaniora Teknologi*, *4*(1), 31–37. https://doi.org/10.34128/jht.v4i1.37

Pratama, Y. T., Bachtiar, F. A., & Setiawan, N. Y. (2018). Analisis Sentimen Opini Pelanggan Terhadap Aspek Pariwisata Pantai Malang Selatan Menggunakan TF-IDF dan Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(12), 6244–6252.

Pristyanto, Y., Adi, S., & Sunyoto, A. (2019). The effect of feature selection on classification algorithms in credit approval. *2019 International Conference on Information and Communications Technology, ICOIACT 2019*, 451–456. https://doi.org/10.1109/ICOIACT46704.2019.8938523

Rizqi, U., Fatichah, C., & Purwitasari, D. (2017). Pembentukan Tesaurus pada Cross-Lingual Text dengan Pendekatan Constraint Satisfaction Problem. *Jurnal Teknik ITS*, *6*(2). https://doi.org/10.12962/j23373539.v6i2.23686

Sakti, Z. (2016). *Pengertian Ceramah, Jenis, komponen, metode, dan Contohnya*.

Shaltout, N. A., El-Hefnawi, M., Rafea, A., & Moustafa, A. (2014). Information gain as a feature selection method for the efficient classification of influenza based on viral hosts. *Lecture Notes in Engineering and Computer Science*, *1*(October 2016), 625–631.

Somantri, O., & Apriliani, D. (2018). SUPPORT VECTOR MACHINE BERBASIS FEATURE SELECTION UNTUK SENTIMENT ANALYSIS KEPUASAN PELANGGAN TERHADAP PELAYANAN WARUNG DAN RESTORAN KULINER KOTA TEGAL. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *5*(5), 537–548. https://doi.org/10.25126/jtiik20185867