Sentiment Analysis of Law Enforcement Performance Using Support Vector Machine and K-Nearest Neighbor

1st Sean Semuel Istia
Department of Information Technology
Satya Wacana Christian University
50711 Salatiga, Central Java, Indonesia
Istiasean@gmail.com

2nd Hindriyanto Dwi Purnomo
Department of Information Technology
Satya Wacana Christian University
50711 Salatiga, Central Java, Indonesia
hindriyanto.purnomo@staff.uksw.edu

***Abstract*— Sentiment analysis or opinion mining is a method to group opinions or reviews into positive or negative. It is important sources for decision making and can be extracted, identified as well as evaluated from online sentiments reviews. This research discussed sentiment analysis in law enforcement on a law case in Indonesia. The analysis uses Support Vector Machine and K-Nearest Neighbor (KNN) for data classification integrated with Particle Swam Optimization (PSO) to increase their performance. The experiment results show that PSO increase the performance of both algorithm**

of the paper is PSO method make value SVM with PSO where value C = 1.0 and Epsilon = 1.0 accuracy 100% while for algorithm KNN with PSO 93.08%. This result show SVM algorithm more accurate than KNN algorithm by using PSO optimization. The performance of law enforcers in the trial case get more positive responses from the people of Indonesia in accordance with their comments or tweets in social media.

Keywords—classification opinion, KNN, SVM, PSO, Sentiment, opinion mining, K-Fold, text mining

I. Introduction

Social media such as Facebook and Twitter are commonly used by its user to express their opinions on various topics. Even professional such as reviewer of news, product and film, as well as blog owners use social networking to express their thought. Information in the media can be categorized into two main types, facts and opinions. Opinions are usually subjective expressions that describe a person's feelings, judgments or feelings toward their entities, events and properties [1]. Opinion mining is very useful in various fields such as commercial product reviews, social media analysis, movie reviews, *etc.* Related to opinion mining, semantic analysis is a valuable technique in creation of recommended systems [2]. Sentiment analysis is done to check the positive, negative and neutral opinion of users regarding products and also to check its popularity and importance in the market. Many issues related to the sentence's classification have been solved with the help of machine learning approach [3].

In this study, sentiment analysis for law enforcement is proposed. The case study used in this research is a law case in Indonesia that has been discussed a lot in the media. Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) are used to classify the data from social network. Particle Swam Optimization (PSO) is added to the SVM and KNN to increase their performance. The PSO algorithm is used to tunes the optimal parameters in the prediction region and it can be used to overcome the problem of local optimum [4].

II. Literature Review

Research on sentiment analysis is growing. Sentiment analysis can be used to assist the study of public policy as well as to provide time and work efficiency a for news providers in classifying news and assisting news seekers to get the daily political news discourse they want [5]. Hidayat [5], conduct research in sentiment analysis on political discourse on online media using Naïve Bayes and Support Vector Machine (SVM). The author stated that the average accuracy for Naïve Bayes method is 59.98% while and the average accuracy for SVM is 90.50%. This research reveals that Support Vector Machine perform better in term of accuracy than Naive Bayes. Lidya et al analyzed sentiment in Indonesian text using Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) [6]. In this paper, the authors claimed that small value of K on K-Fold Cross Validation will result on low accuracy, while increase the value of K will increase its accuracy. The processing time of SVM is longer than KNN, because the vector matrix on SVM is larger than KNN so that iteration for validation is big enough.

Mohammad et al, [7] proposed techniques to label the sentiment accurately with two methods: sentiment classification algorithm (SCA) based on k-nearest neighbor (KNN) and based on support vector machine (SVM). Authors design their classifier with few features like n-gram feature, pattern feature, punctuation feature, keyword based feature and word feature. The paper has been focused on dividing the tweets into positive and negative sentiment and show that sentiment classifier algorithm (SCA) performs better than SVM. Shaki et al, [8] conduct similar research with [7] where the authors used KNN to replace SVM. Using data set up to 6000 tweets, the research show the accuracy of binary, ternary and multi-class sentiment analysis using the KNN algorithm. The execution time using KNN is also less than the other existing techniques mentioned in the paper.

In the paper by Nirmala et al, [9], the SVM parameter C, σ and ε are tuned using PSO. The experiment results show that the approach not only able to select important features but also to yield high accuracy for sentiment classification. PSO affect the accuracy of SVM after the hybridization of SVM-PSO. SVM-PSO method obtains better result than the SVM on the benchmark dataset of Movies reviews dataset.

978-1-5386-7082-8/18/$31.00 ©2018 IEEE

III. Research Model

A. Sentiment Analysis

Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining [4].

The steps commonly found in the text classification of sentiment analysis are: 1. Define dataset domains: collection of datasets that surround a domain, e.g movie review dataset, product review dataset, and so on. 2. Pre-processing: the initial processing stage is generally done by Tokenization process, Stopwords removal, and Stemming. 3. Transformation: the process of representation of numbers calculated from textual data [10].

B. Pre-Processing

Pre-processing is the process of preparing and cleaning the data for classification [1]. Stages of Text Preprocessing used in this research are Cleansing, Case Folding, Tokenizing, Normalization (Replacement and Eliminate a repetitive characters), Filtering and Stemming.

C. Term Weiighting

Term weighting aims to evaluate the relative importance of different terms. There are three components in a term weighting scheme: local weight, global weight and normalization factor [11]. This stage is largely a weighting technique in text mining using Term Frequency–Inverse Document Frequency (TF.IDF). TF.IDF applies the weighting of the combination of both the multiplication of local weight (term frequency) and global weight (global inverse document frequency) [10]. The Determination of Term Frequency (TF) is valued by looking at the number of words that appear. The TF-IDF can be formulated as follows:

$$idf = \log (N/df) \qquad (1)$$

Where N – number of sentences, df – the number of repeated words.

$$w(t,d)=tf\,(t,d) * idf \qquad (2)$$

Where $w(t,d)$ – sentence weight d against word t, tf – term frequency, t – calculated words, d – sentence weight d, idf – Inverse Document Frequency.

D. Support Vector Machine

Support Vector Machines (SVM) is a supervised learning method that analyzes data and recognizes patterns used for classification. SVM has the advantage of being able to identify a separate hyperplane that maximizes the margin between two different classes. The classification process using SVM begins by converting text into vector data. Vector in this research that is weight. These weights are often combined into a TF.IDF value, simply by multiplying together [12].

For the determination of each sentence is positive or negative, calculations are made to find the hyperplane first with the formula:

$$f(\phi(x)) = sign(w.\phi(x)+b) \qquad (3)$$

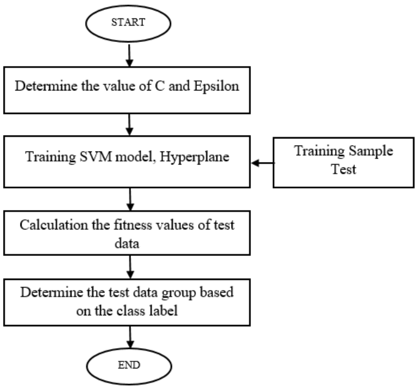
Where $f(\phi(x))$ – Results Classification of test data, w – weight, b – biased, $\phi(x)$ – Kernel test data calculation.

$$K(x,x_i) = (x \,.\, x_i + 1)^2 \qquad (4)$$

Where $K(x, x_i)$ – kernel, x – training data, x_i – training data to i.

With the determination of hyperplane, the test data can be calculated according to the weight of the test data whether including positive or negative class.

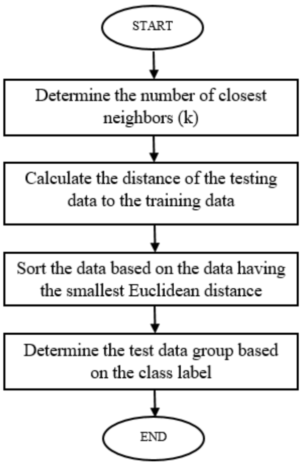
Fig. 1. Flowchart Support Vector Machine


START
Determine the value of C and Epsilon
Training SVM model, Hyperplane
Training Sample Test
Calculation the fitness values of test data
Determine the test data group based on the class label
END

E. K-Nearest Neighbor

KNN is a method to classify new object based on (k) [13]. The goal of KNN is to classify objects based on attributes and training samples. Training sample is the number of data and responses collected based on the case of research, while attributes are one document in accordance with the number of training samples made in the word matrix and has been done weighting.

Fig. 2. Flowchart K-Nearest Neighbor


START
Determine the number of closest neighbors (k)
Calculate the distance of the testing data to the training data
Sort the data based on the data having the smallest Euclidean distance
Determine the test data group based on the class label
END

The value of k should not be greater than the amount of training data. The distance of training data closest to the object to be classified can be calculated using the Euclidean Distance method is formulated as:

$$D(a,b)= \sqrt{\textstyle\sum_{k=1}^{d} \; (a_k\text{-}b_k)^2} \qquad (5)$$

Where *D (a,b)* – scalar distance from two data vectors, *a* – testing data, *b* – training data.

The result of Euclidean Distance calculation is then sorted and taken as much *k* data in order of data quantity for determination of class data.

F. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population-based search algorithm and initialized with a population of random solutions called particles [14]. PSO has advantages such as simple, few parameters, convergence speed and easy to realize, so that PSO is more applied in the field of function optimization, neural network training, pattern classification and traditional optimization algorithm. The particle swarm optimization is valued randomly in search and movement space through D space [15]. The following equations:

$$v^T_{i+1} = w.v^T_j + c_1.r_1(p^i_j - x^i_j) + c_2.r_2.(p^g_j - x^i_j) \quad (6)$$

$$x^i_{j+1} = x^i_j + v^i_{j+1} \quad (7)$$

Where x^i_j and v^i_j – are the position and velocity of particle *i* in the search space. The position and velocity of particle will be updated in every movement (iteration). *r1* and *r2* – are random number in the range [0,1]. c1 and c2 – are coefficient of cognitive learning and social learning. p^i_j –is the best situation of particle *i*, p^g_k–are the global particle best among the population.

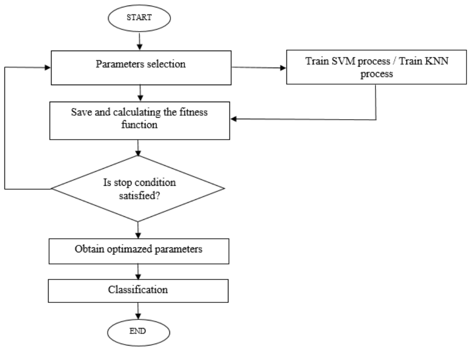
- Particle Swarm Optimization based SVM

  In this study, PSO is used to tune the parameters of Support Vector Machine. Several experiments were conducted to obtain best accuracy. The experiments on Support Vector Machine method and Particle Swarm Optimization are used to optimize the value of C and Epsilon parameters. In result, good parameter value will improve the accuracy of the SVM.

- Particle Swarm Optimization based KNN

  In this study, PSO is used to determine the value of *k* to optimize the performance of KNN classifiers. The appropriate value of *k* will increase the accuracy of the KNN.

Fig. 3. Flowchart PSO based SVM and KNN


START
Parameters selection
Train SVM process / Train KNN process
Save and calculating the fitness function
Is stop condition satisfied?
Obtain optimazed parameters
Classification
END

IV. The implementation of sentiment Analysis

A. Data Collection

In this study, data is taken from social media Twitter.com with the theme of trial of Jesicca Kumala Wongso. The keywords used are related to the performance of law enforcement. It is used to capture comments text in Twitter.com. Data is taken randomly based on the tweets on twitter. In the process of snipping, 1200 data tweets are grouped into two classes positive and negative.

B. Pre-Processing

Pre-processing in this case will be done in accordance with the Indonesian language.

1) Cleansing. In this phase, all characters of non alphabetion is removed.

TABLE I. Cleansing

Before Cleansing	After Cleansing
Hukum indonesia jadi rusak&konyol #SidangJesicca	*Hukum indonesia jadi rusak konyol*

2) Case Folding. The next step is changing all the letters into all lowercase letters (lower case).

TABLE II. Case Folding

Before Case Folding	After Case Folding
Hukum indonesia jadi rusak konyol	hukum indonesia jadi rusak konyol

3) Tokenizing. In this step, a sentence is splits into each word that composes it.

TABLE III. Tokenizing

Before Tokenizing	After Tokenizing
hukum indonesia jadi rusak konyol	hukum indonesia jadi rusak konyol

4) Normalization. All slangs is normalized to becomes normal words.

TABLE IV. Normalization

Before Normalization	After Normalization
kalo	kalau
sampe	sampai
hakim	hakim
memutuskan	memutuskan
bersalah	bersalah
dan	dan
dihukum	dihukum
pindah	pindah
warna	warna
negara	negara
gue	saya
Before Eliminate	**After Eliminate**
jpu	jpu
mulaaaai	mulai
mikir	pikir
keras	keras

5) Filtering. Words that are appear too often, having general relevance and showing less relevance to the text are removed.

TABLE V. FILTERING

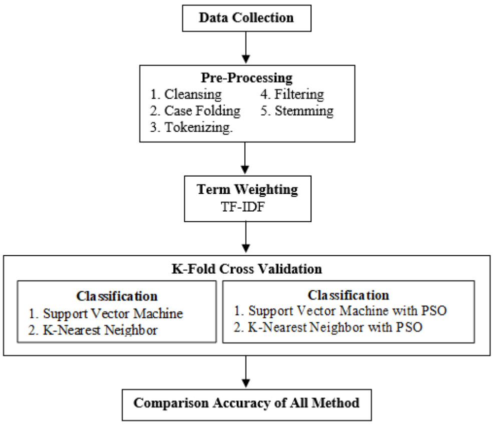
Before Filtering	After Filtering
hukum indonesia jadi rusak konyol	hukum indonesia rusak

6) Stemming. It is the process of combining or separating every variant of a word into a word.

TABLE VI. STEEMING

Before Steeming	After Steeming
hakim memutuskan bersalah	hakim putus salah

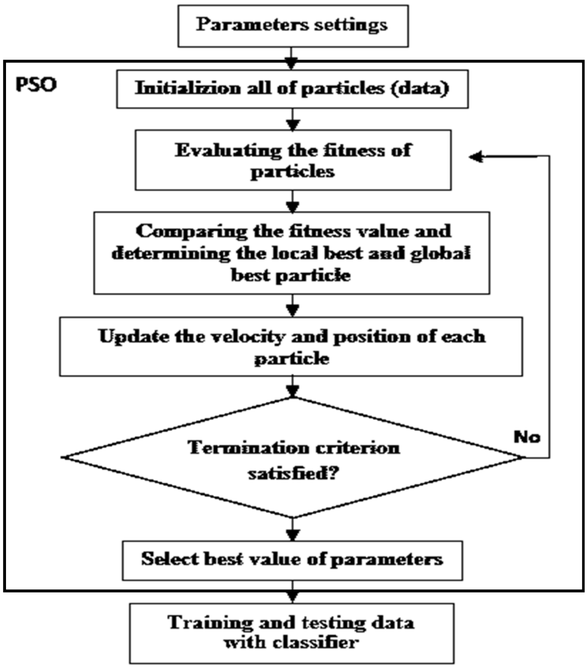
Fig. 4. The Proposed Method



C. Applying PSO to SVM and KNN

PSO is used to select the appropriate parameters on the Support Vector Machine and K-Nearest Neighbor method. Several experiments were conducted to obtain best accuracy. In this research, the parameters of SVM are and Epsilon the parameter of KNN is *k*. After weighting the data, parameters settings be required to data classification. The use of PSO on SVM and KNN will improve the tuning parameters of both methods.

Fig. 5. Applying PSO to SVM and KNN parameters


Parameters settings
PSO
Initializion all of particles (data)
Evaluating the fitness of particles
Comparing the fitness value and determining the local best and global best particle
Update the velocity and position of each particle
Termination criterion satisfied?
No
Select best value of parameters
Training and testing data with classifier

V. RESULT AND DISCUSSIION

A. Data Analysis using SVM and SVM-PSO

In this research, RapidMiner Studio is used to process the text data. For SVM, the values of C and Epsilon are tested from the range of 0.1 to 1.0. The classification results using SVM are shown in the table VIII:

TABLE VII. ACCURACY DATA SVM AND SVM-PSO

C	*Epsi -lon*	*SVM*		*SVM-PSO*	
		Accu -racy	*AUC*	Accu- racy	*AUC*
0.1	0.1	87.89%	1.000 (Pos)	89.21%	1.000 (Pos)
0.3	0.1	88.32%	1.000 (Pos)	90.43%	1.000 (Pos)
0.7	0.1	98.01%	1.000 (Pos)	99.80%	1.000 (Pos)
1.0	0.1	100%	1.000 (Pos)	100%	1.000 (Pos)
0.1	0.3	87.89%	1.000 (Pos)	89.21%	1.000 (Pos)
0.3	0.3	88.60%	1.000 (Pos)	90.43%	1.000 (Pos)
0.7	0.3	95.58%	1.000 (Pos)	100%	1.000 (Pos)
1.0	0.3	100%	1.000 (Pos)	100%	1.000 (Pos)
0.1	0.7	87.89%	1.000 (Pos)	89.21%	1.000 (Pos)
0.3	0.7	88.75%	1.000 (Pos)	90.43%	1.000 (Pos)
0.7	0.7	99%	1.000 (Pos)	100%	1.000 (Pos)
1.0	0.7	100%	1.000 (Pos)	100%	1.000 (Pos)
0.1	1.0	87.89%	1.000 (Pos)	89.21%	1.000 (Pos)
0.3	1.0	89.17%	1.000 (Pos)	92.87%	1.000 (Pos)
0.7	1.0	99%	1.000 (Pos)	100%	1.000 (Pos)
1.0	1.0	100%	1.000 (Pos)	100%	1.000 (Pos)

Fig. 6. Graph of Accuracy Testing *C* Value where *Epsilon* = 1.0

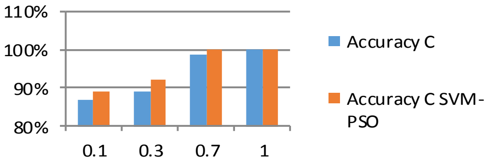
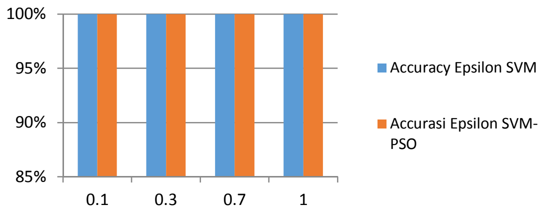


Fig. 7. Graph of Accuracy Testing *Epsilon* Value where *C* = 1.0



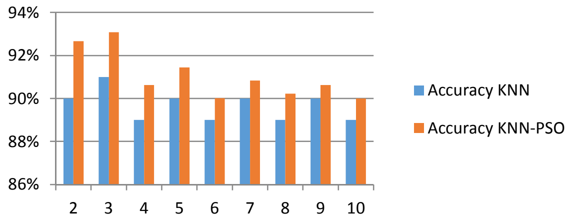
In Table VII, the accuracy of SVM is 100% with the of value C = 1.0, regardless of the value of Epsilon. This is because large value of C can maximize Margin or the distance between the hyperplane and the pattern of each class. On the other hand, the value of epsilon is slightly influencing the accuracy. Based on the experiments, the best values for C and Epsilon are C = 1.0 and Epsilon = 1.0. The table also shows that PSO increase the accuracy in most of the experiments design. At C = 0.7, 100% accuracy are obtained when the value of Epsilon = 0.3,0.7,1.0. It also obvious that the highest accuracy values are obtained when the C = 0.7 and C = 1.

B. *Data Analysis using KNN and KNN-PSO*

TABLE VIII. ACCURACY DATA KNN AND KNN-PSO

k value	*KNN*		*KNN-PSO*	
	Accuracy	***AUC***	**Accuracy**	***AUC***
2	90.02%	0.973 (Pos)	92.67%	0.975 (Pos)
3	91.04%	0.960 (Pos)	93.08%	0.948 (Pos)
4	89.82%	0.940 (Pos)	90.63%	0.943 (Pos)
5	90.22%	0.920 (Pos)	91.45%	0.923 (Pos)
6	89.41%	0.911 (Pos)	90.02%	0.911 (Pos)
7	90.22%	0.917 (Pos)	90.84%	0.898 (Pos)
8	89.21%	0.911 (Pos)	90.22%	0.898 (Pos)
9	90.63%	0.903 (Pos)	90.63%	0.903 (Pos)
10	89.41%	0.895 (Pos)	90.43%	0.891 (Pos)

Fig. 8. Graph of Accuracy Testing KNN and KNN-PSO



For the KNN, the value of *k* is determined from *k* = 2 to *k* = 10. Based on the experiment, the highest accuracy is obtained when *k* is 2. In this test, the value of *k* = 1 is not considered due to sensitive noise. In KNN-PSO method, the accuracy is above 905 for all value of *k*. The highest accuracy is obtained when *k*= 3.

C. *Comparison Classifier Methode*

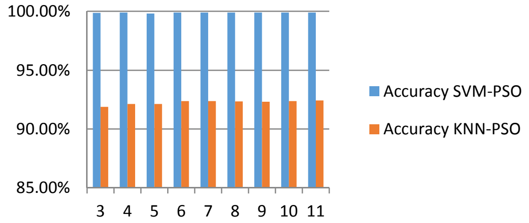
N-fold Cross Validation Technique is used to validate the accuracy of the proposed methods.

TABLE IX. COMPARISON ACCURACY SVM-PSO AND KNN-METHODS

n-fold	Accuracy SVM-PSO	Accuracy KNN-PSO
3	99.86%	91.88%
4	99.91%	92.12%
5	99.82%	92.13%
6	99.91%	92.36%
7	99.91%	92.38%
8	99.91%	92.35%
9	99.91%	92.31%
10	99.91%	92.36%
11	99.91%	92.42%

The validation method is set with the value of *n* = 3 to 11. The highest accuracy is obtained when the value of *n* = 11. The experiment result indicate that the higher the value of n, the better its accuracy.

Fig. 9. Comparison Accuracy SVM-PSO and KNN-PSO in K-FOLD



VI. CONCLUSION

In this research, sentiment analysis on law enforcement is conducted using SVM and KNN. The data used in this research is derived from twitter with the theme a law case occurs in Indonesia. The comments on twitter that are used as data are written in Indonesia language. The data is processed using SVM and KNN. PSO is integrated in SVM and KNN to improve the accuracy of both methods. The experiment results show that SVM is more accurate than KNN. The use of PSO also clearly increases the accuracy of both methods. It can be infer that PSO can improve the process of parameters tuning for both SVM and KNN.

References

[1] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishera, 2012, p.17.

[2] Speriosu, M. et al, "Twitter polarity classification with label propagation over lexical links and the follower graph", Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP, Edinburgh, Scotland, 2011, volume 30, issue 32, page no. 456-475

[3] Thelwall, M., Buckley, K., Paltoglou, G., "Sentiment strength detection for the social web", J. American Society for Information Science and Technology, 2012, volume 63 issue 1, pp- 163–173

[4] Shieh, M.Y. et al., " Applications of PCA and SVM-PSO based real-time face recognition system", Math. Probl. Eng. 2014, 2014, doi:10.1155/2014/530251.

[5] Hidayat. A.N, "Analisis Sentimen Terhadap Wacana Politik PadaMedia Masa Online Menggunakan Algoritma Support Vector Machine Dan Naive Bayes", Jurnal Elektronik Sistim Informasi Dan Komputer (Jesik), Vol.1, pp.3-5, 2015.

[6] Lidya S.K, Sitompul, S. Efendi, "Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM) dan K-Nearest Neighbor (K-NN)", in Paper Presented at National Technology and Information Conference, 28 March 2015. Yogyakarta, Indonesia.

[7] Huq R. Mohammad., Ali, Ahmad., Rahman, Anika., “Sentiment Analysis on Twitter Data using KNN and SVM”., (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017.

[8] Shaki, Anjume., Arora, Jyoti.,” Sentiment Analysis of Twitter Data using KNN Classification Technique”., IJSRD - International Journal for Scientific Research & Development| Vol. 6, Issue 03, 2018 | ISSN (online): 2321-0613

[9] Nirmala Devi et al.,“ Sentiment Classification Using Svm And Pso”., International Journal of Advanced Engineering Technology, E-ISSN 0976-3945, Int J Adv Engg Tech/Vol. VII/Issue II/April-June,2016/411-413.

[10] Moraes, R, Valiati, J. F, & Gavião Neto, W. P, “Document Level Sentiment Classification: An Empirical Comparison between SVM and KNN”, Expert Systems with Applications, Vol 40, pp.621–633, 2013.

[11] Lan M, Tan C. L, Su J, & Lu Y, “Supervised and traditional term weighting methods for automatic text categorization”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31 (4), 721 -735, 2009.

[12] Yunita, Norma, “Analisis Sentimen Berita Artis Dengan Menggunakan Algoritma Support Vector Machine Dan Particle Swarm Optimization”, Jurnal Sistem Informasi Stmik Antar Bangsa, Vol 5, no 2, pp.104-112, 2016.

[13] Gorunescu. F, Data Mining : Concepts, Models, and Techniques, Verlag Berlin Heidelberg, Springer, 2011.

[14] Abraham, A., Grosan, C, & Ramos, V. “Swarm Intelligence in Data Mining”. USA: Spinger, XVIII, 268, 2006.

[15] Salehpour, Elham., Iran, J. Vahidi., Hosseinzadeh , Hssan.,“Solving optimal control problems by PSO-SVM”., Computational Methods for Differential Equations http://cmde.tabrizu.ac.ir Vol. 6, No. 3, 2018, pp. 312-325

[16] Patil G, Galande V, et al, “Sentiment Analysis Using Support Vector Machine”, International Journal of Innovative Research in Computerand Communication Engineering, Vol. 2, p.2609, 2014.

[17] Chou, J.-S, Cheng, M.-Y, Wu, Y.-W & Pham, A.-D, “Optimizing Parameters Of Support Vector Machine Using Fast Messy Genetic Algorithm For Dispute Classification”, Expert Systems With Applications, Vol 41, pp. 3955–3964, 2014.

[18] J. Han, M. Kamber & J. Pei, Data Mining Concepts and Techniques, Morg an Kaufmann Publishers is an imprint of Elsevier, 3rd ed, pp.83-91, 2012.

[19] Basari et al, “Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization”, Procedia Engineering, Vol 53, pp.453-462, 2013.

[20] Rozi, Hadi, Achmad, “Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi”, Jurnal EECCIS, Vol. 6, No. 1, pp.621 –633, 2012.

[21] Haddi, E, Liu, X, dan Shi. Y, “The Role of Text Pre-processing in Sentiment Analysis”, Procedia Computer Science, Vol 17, pp.26-32, 2013.