ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ALGORITHM APPLICATION SUPPORT VECTOR MACHINE WITH GENETIC ALGORITHM OPTIMIZATION TECHNIQUE FOR SELECTION FEATURES FOR THE ANALYSIS OF SENTIMENT ON TWITTER

[1]MOCHAMAD WAHYUDI, [2]DWI ANDINI PUTRI
[1]Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri)

[1]wahyudi@nusamandiri.ac.id, [2]dwiandini@nusamandiri.ac.id

ABSTRACT

Twitter has become one of the most popular micro-blogging platform, recently. Millions of users can share their thoughts and opinions about various aspects and activites. Therefore, twitter considered as a rich source of information for decision-making and sentiment analysis. In this case, the sentiment is aimed to overcome the problem of automatically classifying user tweets into positive opinion and negative opinion. The classifier Support Vector Machine (SVM) used in this study is a machine learning technique that is popular text classifiers, as Support Vector Machine (SVM) algorithm is one that has a linear calcification of the main principles for determining the linear separator in the search space that can best separate the two classes different. But the Support Vector Machine (SVM) has the disadvantage that the appropriate parameter selection problem. The tendency in recent years is to simultaneously optimize the features and parameters for Support Vector Machine (SVM), so as to improve the accuracy of classification on Support Vector Machine (SVM). Genetic Algorithm has the potential to produce better features and becomes optimal parameters at the same time. This research generate text classification in the form of positive and negative tweets on twitter. Measurement accuracy is based on Support Vector Machine (SVM) before and after using a Genetic Algorithm. Evaluation was performed using 10 fold cross validation while accuracy is measured by the confusion matrix and ROC curves. The results of the study showed an increase in accuracy of Support Vector Machine (SVM) from 63.50% to 93.50%.

Keywords*:* *Sentiment Analysis, Twitter, Support Vector Machine (Svm), Classification Text*

1. INTRODUCTION

The last few years have seen a surge inveryattr active in a computational method that has an influence on opinion mining, subjectivity detection, and sentiment analysis. (Balahur, Mihalcea, & Montoyo, 2014). Twitter has become one of the most popular microblogging platform recently. Millions of users can share their thoughts and opinions about various aspects and activities for their micro-blogging platform. Therefore twitter considered a rich source of information for decision-making and sentiment analysis. The emergence of social media has provided a place for web users to share their thoughts and express their opinions on different topics in an event. Twitter has nearly have 600 million users and more than 250 million messages per day. This quickly has become a gold mine for organizations to monitor their brand reputation and with extracting and analyzing the sentiment of tweets posted by the public abouttheir markets and competitors (FH Khan, Bashir, and Qamar, 2014)

Another major area of research focuses on the development of Twitter sentiment analysis approach designed specifically to tweet. Tweets is a unique genre of communication, and the unique features and properties has been questioned in the application and effectiveness of the more traditional approach to sentiment analysis. Tweets are very short text units, a maximum of 140 characters long, and is characterized by a relaxed, compact language with extensive use of slang, abbreviations, acronyms, and emoticons. Tweets also contain hashtags, user references, and embedded links to other sites that contain additional information referenced, more complicated sentiment analysis. (Ghiassi, Skinner, and Zimbra, 2013)

There are few studies that have beendone in terms of sentiment classification on twitter are available online including, sentiment analysisonthe classification witterby using

Support Vector Machine (SVM) (Passonneau, 2011). Opinion sentiment an analysis on movie reviews using the classifier Support Vector Machine (SVM) and Particle Swarm Optimization (Basari etal., 2013). Sentiment clacification to review travel destination using Supervised Machine Learning approach (Ye, Zhang, &Law, 2009).

Support Vector Machine (SVM) is a linear calcification algorithm which has the main principle for determining the linear separator in the best search space that can separate two different classes. (Schoefegger, Tammet, & Granitzer, 2013) excess in the SVM algorithm that is able to identify the separate hyperplane that maximizes the margin between two classes (Chou, Cheng, Wu, and Pham, 2014). SVM is a special case of a family of algorithms which we refer to as a regulator of linear classification method and a powerful method for risk minimization (Weiss, Indurkhya, and Zhang, 2010). However Support Vector Machine (SVM) has a shortage of the appropriate parameter selection problem. (Basari et al., 2013)

Feature selection affects several aspects of the pattern of classification, classification accuracy, the time required for learning classification functions, the amount of sample needed for learning andcosts associated with the features according to Yang and Honavarin Zhao (Zhao, Fu, Ji, Tang, and Zhou, 2011). Incertain application problems, notall ofthese featuresare equally important. Better performancecan be achieved by removingsome of the features. Thus, wecaneliminate thenoisedata, the datathat isirrelevantand redundant.

In addition to the selection of features, setting the parameters of Support Vector Machine (SVM) have an important influence on the accuracy of the classification. Incompatibility of a parameter setting can cause a low classification results according to Keerthi and Lin in Zhao (Zhao et al., 2011). SVM optimal classification accuracy is obtained by finding the optimal parameter settings. The parameters to be optimized is the error parameter C and the kernel function parameters such as Gaussian kernel parameter γ for the Gaussian kernel function. Grid search is an alternative and direct search approach. However, the search capability of this approach is very low according to Hsu, Chang, and Lin in Zhao (Zhao et al., 2011). In addition, the search grid can not perform feature selection. In some literature, feature selection method based on Genetic Encryption (GA) has been proposed by Raymer, Punch, Goodman, Kuhn, and Jain in Zhao (Zhao et al., 2011)

The trend in recent years is to simultaneously optimize the part features and parameters for Support Vector Machine (SVM), so as to improve the accuracy of classification on Support Vector Machine (SVM). Genetic algorithms have the potential to produce better features and become the optimal parameters at the same time. (Zhao et al., 2011) In this study the algorithm of Support Vector Machine (SVM) and Genetic Encryption method as a feature selection method to be applied to classify the existing tweet on twitter to improve the accuracy of sentiment s

2. LITERARY REVIEW

A. Sentiment Analysis (Sentiment Analysis)

According Haddi (Haddi, Liu, & Shi, 2013) Sentiment Analysis is treated as an assignment of a classification to classify text orientation to be positive or negative. According Medhat (Medhat, Hassan, & Korashy, 2014) Sentiment Analysis can be considered as a classification process that has 3 main classification level is the level of the document, at the sentence level and at the level of aspects.

According toMoraes (Moraes etal., 2013) measures that are commonly foundin text classification sentimenta n alysisare:

1) *Define domain datasets:* The collection of datasets that covers adomain, for example dataset movie reviews, product reviews datasets, datasets tweets and soforth.

2) *Pre-processing:* Initial processing stage which is generally carried out by the process of tokenization, stop words removal, and stemming.

3) *Transformation:* Process representation figures calculated from textual data. Binary Representation commonly used and only count the presence or absence of awordin the document. How many times aword appears in a documentis also used as aweighting scheme of textual data

B. Support Vector Machine (SVM)

According to Widodo (2013) Support Vector Machine (SVM) is an algorithm that works using a nonlinear mapping to transform the original training data to a higher dimension. In this new dimension, will seek to separate hyperlane linearly and with appropriate nonlinear mapping dimension to high enough, the data from the two classes can always be separated by the hyperlane. SVM find this hyperlane using Support Vector and Margin.Menurut Basari (Basari, Hussin, Ananta, & Zeniarja, 2013) in determining a weighting value of positive and negative class Support Vector Machine (SVM) is determined based on if the value is greater than the weight is equal to 0 then classified into positive class and vice versa if the tilapia weight of less than 0, it can be classified into the negative class. Here is a function of the weight calculation formula on Support Vector Machine (SVM) (Wu, 2009).

$$W^{*} = \sum_{i=1}^{m} \alpha_{i}^{*} y_i X_i$$

The variables and parameters Algorithm Support Vector Machine (SVM) forcalcification:

1.χ={χ1, χ1, .., χm} as training samples
2.y={y1, .., ym} C{± 1} as the label training data
3. Thekernel=type of kernel function
4.par=kernel parameter
5.C=konstanta cost
6.A=[α1, .., αm] as a multiplier and bias

C. Selectionfeature(Feature Selection)

According Gorunescu (Gorunescu 2011) feature selectionis used to eliminateir relevant features and repetitive, which maylead tochaos, by using certain methods. According to John, Kohavi, and Pflegerin Chen (Chen etal., 2009) feature selection method sin machine learningone of them:

1)*Wrapper*: According to Chen (Chen etal., 2009) wrapper using classification accuracy of several algorithms as a function evaluation. Because the wrapper should test aclassifier for each subset of features tobe evaluated, usually moretime-consuming, especially when the number ofitshigh. According Kohaviin Yang (Yang, Liu, Zhou, Chawla, &Albert, 2010) wrapper feature repeatedly evaluate compliance with inductive algorithms. The refore, the features selected by the wrapper approach might bemore suitable for inductive algorithm, and producesa high classification accuracy. According Gunal (Gunal, 2012) one of the wrapper methods that can be used in the selection of the features isthe Genetic Algorithm (GA). This research will be discusse din more detailon the Genetic Algorithm as Genetic Algorithm (Genetic Algorithm)

According Zukhri (Zukhri, 2014) Optimization is the process of completing a specific problem that is at the most favorable conditions from a view point. Which solved the problem closely related to the data that can be express edinone or several variables. According Zukhri (Zukhri, 2014) Genetic Algorithm is a heuristic method was developed based on the principles of genetics and the process of natural selection Darwin's theory of evolution. Optimization method developed by John Holl and around 1960 and popularized by one of his students, David Gold bergin the 1980s. Completion of the search processin the algorithm takes place just as the election of individuals to survivein the evolutionary process.

D. Review from related research study

There are several studies using Support Vector Machine algorithm as clasification in text sentiment classification on twitter, including three related studies discussed in this study have different models, but the classifier Support Vector Machine (SVM) has been shown to have the high estaccuracy amongo ther classifiers. Genetic algorithms can be used as an optimization model that can be used to produce the accuracyof Support Vector. Machine (SVM) is higher. Comparison of three related research can beseen in Table1.

Table1: Comparison to Related Research

Tittle	Preprocessing	Feature Selection	Classifier	Accuracy
Sentiment Analysis of Twitter Data (Passonneau, 2011)	- Replace all the emoticons by emoticon dictionary - Replace all URLs with a Tag "U" - Replace targets with a Tag "T" - Replace all negations by Tag "NOT" - Replace a sequence of repeated characters - Tokenizing all the tweet - Dentify stop words	- POS-specific prior polarity features - tree kernel	Support Vector Machine (SVM)	75,39%
Opinion Mining of Movie Review using Hybrid Method of Support	- fillter data -Data cleansing - xtract to text file	- Case Normalization - Tokenizing - Steamming - Generate	Support Vector Machine (SVM)	77%

Vector Machine and Particle Swarm Optimization (Basari et al., 2013)		n-Gram		
Sentiment classification of online reviews to travel destinations by supervised machine learning approaches (Ye et al., 2009)	Convert all characters to lowercase	n-Gram	Support Vector Machine (SVM)	80%
Proposed Model	- Tokenization - Stop Removal words - Stemming - Generate 2-gram	- N-Gram - Genetic Algoritma	Support Vector Machine (SVM)	?

By reviewing studies above, it can be seen that the Support Vector Machine (SVM) classifier is best to solve the problem of sentiment analysis. In this study used a Support Vector Machine classifier algorithm (SVM) and the method used optimization techniques are Genetic algorithm as a feature selection method to improve the accuracy of classifiers.

3. REVIEW THE RESEARCH OBJECT

The object of research conducted in this thesis is on twitter sentiment analysis and optimization techniques in the context of feature selection with the following explanation:

b. Sentiment on Twitter

According to Khan (K. Khan, Baharudin, and Khan, 2014) Micro-blogging site Twitter is a rich source of information and diverse. This is due to the nature of the micro blog where people can send messages in real time about their opinions on various topics, discuss the issues that are popular, they also can argue about the complaints and express positive sentiments for prouk they use in daily life. Even some of the companies manufacturing the products have studied the reactions of users through twitter. However extent on twitter sentiment data make it difficult to analyze and understand the sentiment to classify tweet on twitter as sentiment positive, negative or neutral in real time. According Dehkharghani (Dehkharghani, Mercan, Javeed, & Saygin, 2014) Twitterisa popular micro blogging and social networking websites with a registered user base of about 650 million per year in 2013 that allows users to send text messages at mostonly 140 characters (tweet). Twitter users send or discuss a message (tweet) about the subject in everyday life. Andit can be seen that in the past few years has been widely used twitter political parties to launch acampaign against the community

c. Feature Selection (Selection feature)

According Medhat (Medhat et al., 2014) the task of sentiment an alysis has been considered as a classification problem, the first step in a sentiment calcification problem is to extract and select features on the text. Here penjelsan of several feature selection is currently:

1) Terms Presence and Frequance: These features are individual words or N-Gram and the number of frequencies that often appear as giving weight to the words into a binary value (zero if the message has appeared, and one said otherwise) or using a frequency weighting is istilh to show interest relative to the features.

2)*Part Of Speech (POS)*: The discovery of an adjective because it is someone important indicator of an opinion.

3) Opinion Words and Phrases: Are words commonly used to mengekspersikan opinions including those of good or bad and like datau hate. From the other hand some phrases expressing an opinion without using words opinions.

4) Negations: The emergence of negative words that can change the orientation of opinion as well not be on par with the bad

d. Validation and Evaluation of Data Mining Algorithms

According to Han (Han & Kamber, 2007) confusion matrix is a very useful tool for analyzing how well the classifier can identify tuples of different classes. In some confusion matrix known as True positive terms that refer to the positive tuples that are correctly labeled by the classifier, while True negative is negative tuples that are correctly labeled by the classifier. There is also a false positive which is a negative tuple incorrectly labeled by the classifier, and false negative is a positive tuples that are not properly labeled by the classifier. ROC curve will be used to measure the AUC (Area Under the Curve). ROC curve divides the positive results in the y-axis and the negative results in the x-axis (Witten, Frank, & Hall, 2011). So the larger the area under the curve, the better the prediction results. which can change the orientation of opinion as well not be on par with the bad.

4. RESEARCH FRAMEWORK

This study starts from the problems in the text calcification on twitter which consists of approximately 140 characters using Support Vector Machine (SVM), in which the classification has a shortage of the appropriate parameter selection problem, due to the incompatibility of a parameter settings may cause the results of the classification low. The dataset used in this study is taking on twitter tweet of data obtained from http://rs.peoplebrowsr.com/ to be tested by using the 100 tweets positive and 100 negative tweet with time over 5 months ago. Preprocessing performed by tokenize, stopwords removal, stemming, generate 2-Gram. and wrapper feature selection method with Genetic Algorithm. While pengkalsifikasi used is Support Vector Machine (SVM). 10 fold cross validation testing will be conducted, the accuracy of algorithm will be measured using a confusion matrix and ROC curves. RapidMiner Version 5.3 is used as a tool to measure the accuracy of the data of experiments conducted in the study.

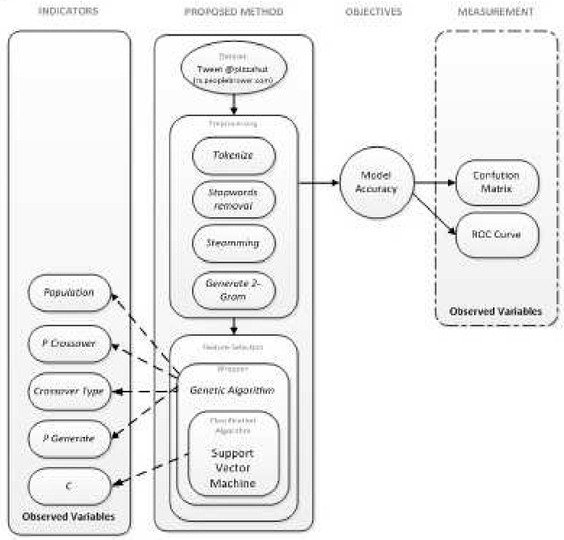


Fig 1. Illustrates The Framework Proposed In This Study

5. RESEARCH METHODOLOGY

a.Data Collection

The dataset used in this study is taking the data tweet on twitter downloaded from http://rs.peoplebrowsr.com/tweetsconsistingof650positiveand634negativetweets. Examples of positive tweets:

"*tweet this #Best Fandom 2014 Directioners and tag me please! make me happy I fucking love pizza and i choose pizza hut do it*".

Examples ofnegativetweets:

*"@pizzahut Fine then, ignore me. I'm going to Burger King anyways, phht. Your loss. Bye. I don't like you. Hmph. *Takes bite out of pizza**".

b.Initial Data Processing

To reduce the length of time data processing, in this study only uses 100 positive and 100 Tweet Tweet negative as training data. Then the data Tweets through preprocessing stage because there are some syntactic features that may not be useful to be processed using the machine learning algorithms, so the data must be cleaned tweet like the @ (at) for a link to the user name, url or website link (http, url, www), # (hashtag), RT (to retweet). This dataset in the preprocessing stage to go through four processes, namely:

c. *Tokenization:* Which collectsall the words that appear an dremoves any punctuation or symbols that are not letters..

d. *Stopwords Removal*: Is the removal of words that are not relevant, such as the, of, for, with, and so on.

e. *Stemming:* Grouping words into several groups that have the same root, such as drug, drugged, and drugs where the root of it all is said drug.

f. *Generate 2-Gram*

Character N-Gram is the nearest whole numbern of the order of the feedback given. For example, 3-grams of words "TERM" willbe"_ _T", "_ TE", "TER" ERM"," RM_"," M__". N-grams with one dimension called unigram, if the two-dimensional called bigram, whereas 3-dimensional called trigrams and if more than 3 dimensions basically calledN-gram.

As for the phase transformation by TF-IDF weighting on each word. Where the process calculates the presence or absence of a word in the document. How many times a word appears in a document is also used as a weighting scheme of textual data

g. Proposed method

The method proposed in this research is to use wrapper feature selection method. Of the type wrapper used is Genetic Algorithm as a feature selection method that the accuracy of the classifier Support Vector Machine (SVM) can be increased. The author uses the classifier Support Vector Machine (SVM) as it is very capable of identifying separate hyperplane that maximizes the margin between the two classes, efficient and is a popular machine learning techniques for text classification, and has good performance. Genetic Algorithm applied writer is using Support Vector Machine (SVM) were tested in the wrapper stage.

h. Evaluation and Validation Results

Validation was performed using 10 fold cross validation. While the measurement accuracy is measured by the confusion matrix and ROC curves to measure the value of AUC. With the confusion matrix, accuracy (SVM) Support Vector Machine before use and after feature selection method using feature selection methods. Table 3.2 The following is a confusion matrix display and calculation formula according Gorunescu (Moraes, 2013):

$$Akurasi = \frac{\mathbf{TP + TN}}{\mathbf{TP + TN + FP + FN}}$$

6. RESULT AND DISCUSSION

a. Classification

Classification process in this research is to determine as entenceas a member of a class of positive or negative class based on the value weighted sum of aformula SVM - value 0. If the value is greater than the weight equal to 0 then classified in to positive classan dvice versaif the weightis less than the same in di goto 0, it can be classified in to negative class. Documents willbe calculated weight value can be seen in table2.

Table 2: Documents The Calculated Value Of Weight.

Document	Atribute						Class
	Love_pizza	love	pizzahut	I_love	I_dislike	price	
Pos13.txt (@pizza hut I LOVE YOUR PIZZA)	0,758	0,354	0,003	0,412	0	0	Positive
Neg94.txt (@pizza hut I dislike you're prices)	0	0	0,001	0	0,001	0,586	Negative

Here is anexample calculation for pos 13. txt document with the following functions:

$$\mathbf{W_{13} = Y_{13} . X_{13}}$$
$$\mathbf{= 1 . (0{,}758+0{,}354+0{,}003+0{,}412)}$$
$$\mathbf{= 1{,}527}$$

Where W13 is the weight for Pos 13. txt document that specifies the classes of positive or negative, while Y13 is the value of the label for the documentin which Y is only worth 1 for the positive class and 1 for a negative class. X13 values of existing attributes in a document you want to be calculated and determined weight class. Weights for Pos 13. txt documentis 1,527, then the document is classified in to positive class.

Here is an example calculation for Neg 94. txt document with the following functions:

$$\mathbf{W_{94} = Y_{94} . X_{94}}$$
$$\mathbf{= -1 . (0{,}433+0{,}001+0{,}586)}$$
$$\mathbf{= - 1{,}020}$$

Where W94 is the weight for pos 94. txt document that specifies the classes of positive or negative, while Y94 is the value of the label for the document in which Y is only worth 1 for the positive class and – 1 for a negative class. X94 values of existing attributes in a document you want tobe calculated and determined weight class. Weights for Neg94. txt documentis-1.02, thenthe document is classified in to negative class. The above calculation can be made a model using Rapid Miner 5. Design a model Support Vector Machine (SVM) can be seenin Figure 2.

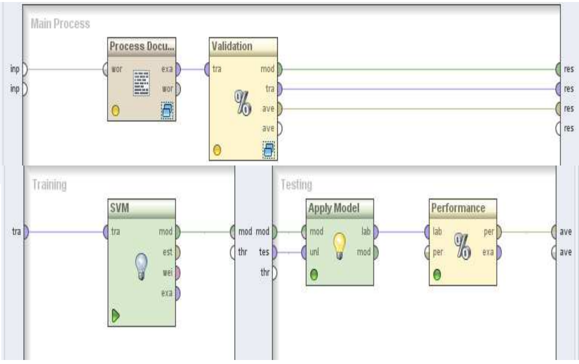


Fig 2. Design Model Of Support Vector Machine (SVM) Using Rapidminer

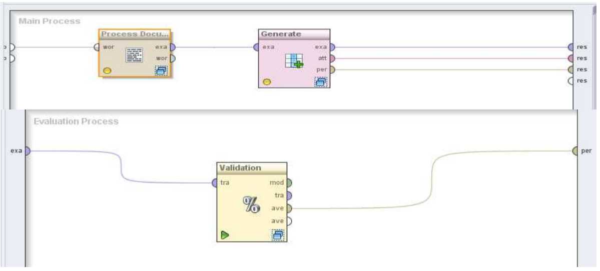
b. *Experiments Against Indicator Model*

Based on experiments conducted by Wahyu (2014) experimental research begins with the adjustment method with genetic algorithm Population size values that are at the default value, then dilanjutkann with multiples of 5 to value Population size that produces the higher the accuracy of which will be used in step experiment next. Then do the adjustment in the value of P Initialize, P Crossover, P Generate with keliapan 0.2 to get a good model. While the method of SVM (Support Vector Machine) based on research Ilhan & Tezel (2013) adjustment is done on the value of the parameter C as a control parameter to a value of 1.0. And the end result the highest accuracy is achieved when the value of the Population size = 15, P Initialize = 0.9, P Crossover = 0.9, P = 1.0 and Generate. Table of indicators and test results can be seen in Table 4.6.

Table 3: Table Indicator And Test Results

Population size	**P Initialize**	**P Crossover**	**P Generate**	**C**	**Accuracy**
5	0.5	0.5	0.1	1.0	80.56 %
10	0.7	0.7	1.0	1.0	92.50 %
15	0.5	0.5	1.0	1.0	86.50 %
15	**0.9**	**0.9**	**1.0**	**1.0**	**93.50 %**
15	1.1	1.1	1.0	1.0	92.50 %
20	0.9	0.9	1.0	1.0	89.43 %

In adjustment indicator on Genetic Algorithm, the highest accuracyis obtained by a combination of populationsize =15, pinitialize = 0.9, p=0.9 crossover, and generate p=1.0. Results accuracy reaches 93.20%. If other indicators also changedits value, cancause the data processingis becoming increasingly longer. Design model of Support Vector Machine (SVM) with Genetic Algorithm can be seenin Figure 4.2.



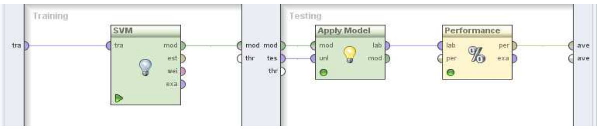


Fig 3. Supoort Model Design Vector Machine And Genetic Algorithm In Rapid Miner

c. *Discussion*

The text classification model on twitter tweets, can be used to follow a strategy of companies and organizations in the decision making of their twitter account, namely by identifying tweet sentiment positive and negative. Tweet of the data that already exists, separated into words, and give weight to each word. Can be seen anywhere words associated with the sentiment that often arise and have the highest weighting. Thus it can be seen that the positive or negative tweets. In this study, the results of testing of the model will be discussed through the confusion matrix to show how well the model is formed. Without using Genetic Algorithm wrapper method, the algorithm of Support Vector Machine (SVM) itself has resulted in an accuracy of 63.50% and the AUC value 0.875. Such accuracy is less accurate, so it needs to be improved further using Genetic Algorithm wrapper method. Accuracy algorithm Support Vector Machine (SVM) increased to 93.50% and the AUC value 0964 as can be seen in Table 4

Table 4: Model Algorithm Support Vector Machine (SVM) Before And After Using Genetic Algorithm

	Algoritma *Support Vector Machine* (SVM)	**Algoritma *Support Vector Machine* (SVM) + Genetic Algorithm**
Successful classification of positive tweets	97	97
Successful classification negative tweets	30	90
Model accuracy AUC	63.50%	93.50%
Model accuracy AUC	0.903	0.940

d. *Measurement with the Confusion Matrix*

Measurements with aconfusion matrix here will show the comparison of the results of the model's accuracy Support Vector Machine (SVM) before being added Genetic algorithm method that can be seenin Table 4 and after adding

Genetic algorithm method that can be seenin Table5.

Table5: Confusion Matrix Model Support Vector Machine (SVM) Prior Tothe Addition Ofgenetic Algorithm Method

Akurasi *Support Vector Machine* (SVM): 63.50% + - 8.38% (mikro 63.50)

	True negative	**True positif**	**Kelas precision**
Pred.negative	30	3	90.91 %
Pred.Positive	70	97	58.08%
Class recall	30.00%	97.00 %	

Table 6: Confusion Matrix Model Support Vector Machine (SVM) After Addition Of Genetic Algorithm Method

Akurasi *Support Vector Machine* (SVM): 93.50% + - 5.50% (mikro 93.50%)

	True negative	True positif	Kelas precision
Pred.negatif	90	3	96.77%
Pred.Positif	10	97	90.65%
Kelas recall	2.14%	97.99 %	

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} = \frac{97-90}{97+90+10+3} = \frac{187}{200} = 0.$$

Figure 4 is the ROC curve for the model Support Vector Machine (SVM) before using Genetic Algorithm and Figure 5 is the ROC curve for the model Support Vector Machine (SVM) after using Genetic Algorithm.

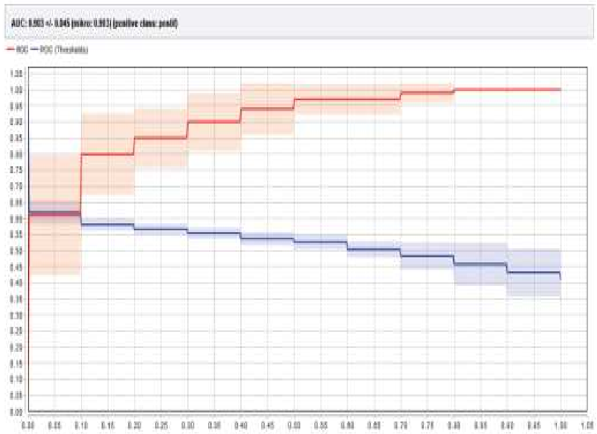


Fig 4. ROC Curve Model Support Vector Machine (SVM) Before Using Genetic Algorithm

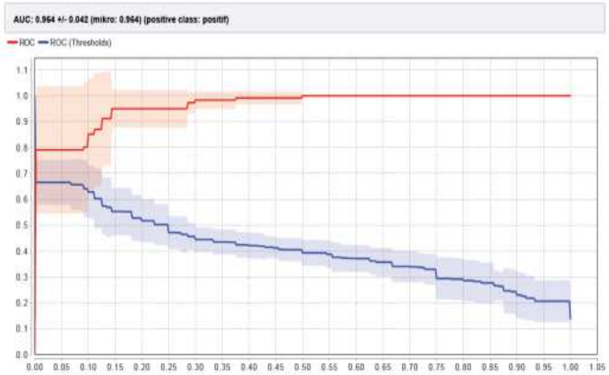


Fig 5.ROC Curve Model Support Vector Machine (SVM) After Using Genetic Algorithm

6. DESIGN AND IMPLEMENTATION

This study will make an application to test existing model susing different datasets and unknown class. Applications are made using Macromedia Dreamweaver software with the PHP programming language. Figure 6 is a flow diagramof the stages of the application process classification
that the author made.

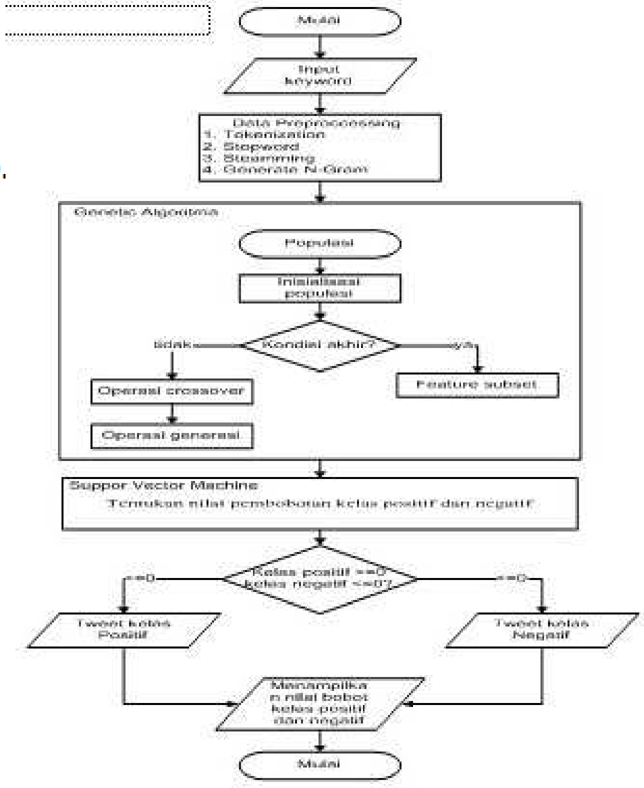
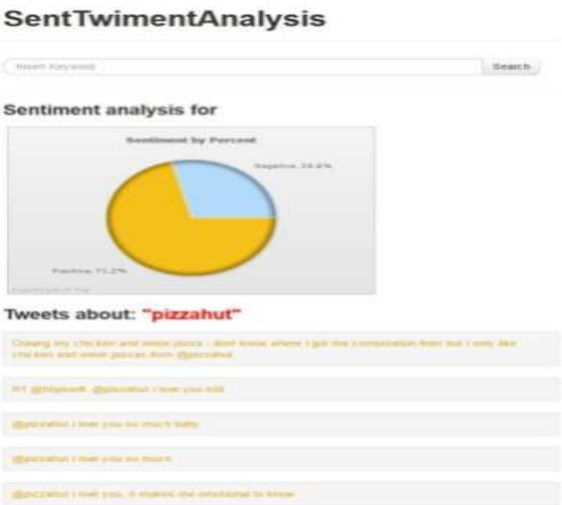


Fig. 6 The folw Diagram Classification Processwith Support Vector Machine (SVM) using Genetic Algorithm Method.

Figure 7 is when the application is inputted keyword "pizzahut" and displays sentiment analysis in the pie chart with the number of tweets percentage of positive and negative tweets and tweet-tweet appears related to the keyword. Figure 7iswhenthe applicationis inputtedkeyword "pizzahut".


SentTwimentAnalysis

Insert Keyword | Search

Sentiment analysis for

Tweets about: "pizzahut"

Craving my chicken and onion pizza - dont know where i got the combination from but i only like chicken and onion pizzas from @pizzahut

RT @...: @pizzahut i love you xxx

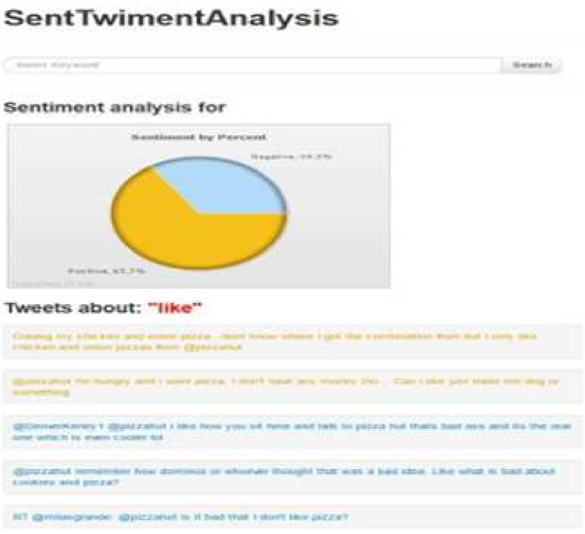
@pizzahut i love you so much baby

@pizzahut i love you so much

@pizzahut i love you. it makes me emotional to know

Fig. 7 Display sentiment analysis applications tweets

Testing applications sentiment analysis tweet with different keywords ie "like" its results can be seen from the figure 8


SentTwimentAnalysis

Insert Keyword | Search

Sentiment analysis for

Tweets about: "like"

Craving my chicken and onion pizza - dont know where i got the combination from but i only like chicken and onion pizzas from @pizzahut

@pizzahut I'm hungry and I want pizza. I don't have any money tho... Can I like just trade me dog or something

@ConnerKenley1 @pizzahut i like how you sit here and talk to pizza hut thats bad ass and its the real one which is even cooler lol

@pizzahut remember how dominos or whoever thought that was a bad idea. Like what is bad about cookies and pizza?

RT @milesgrande: @pizzahut is it bad that I don't like pizza?

Fig.8 Display Sentiment Analysis Applications Tweet With Different Keywords

7. IMPLICATIONS OF RESEARCH

The implications of this research covers several aspects, including:

A. Implications Of System's Aspects

The evaluation results demonstrate the application of Genetic algorithm can improve the accuracy of Support Vector Machine (SVM) and a pretty good method to classify tweets on twitter. Program application made in this study using computer specification requirements such as Intel processors inside core i3, 2GB memory, 250GB hard drive, Windows 7 operating system, using Microsoft software Dreamwaever and PHP programming language.

B. Implications for Managerial aspects

Thus the application of this SentiTweetAnalysis application program can be used to follow a strategy of companies and organizations in the decision making of their twitter account as an evaluation decision, namely by identifying tweet sentiment positive and negative. In the implementation of this program needs to be established SentiTweetAnalysis application of SOP (Standard Operating Procedure) for the standard processes that exist in the system, held socialization or counseling to inform procedures for the use of the SentiTweetAnalysis application, the appointment of the administrator is assigned to manage the application SentiTweetAnalysis, training against SentiTweetAnalysis application users to run in accordance with the SOP has been determined, then perform maintenance (maintanace) on application.

C. *Implications for further research aspects*

Such research can be developed by the unit and the different domains such as camera review, review of the hotel, and so on. Other possible uses classifiers outside Supervised Learning. So it could be different from the general study of existing as using pengkalsifikasian Naive Bayes, Artificial Neural Network, Maximum Entropy and others.

8. CONCLUSIONS

To classify text sentiment analysis with data in the form of data tweet on twitter, one classifier that can be used is the classifier algorithm using Support Vector Machine (SVM). This is because the Support Vector Machine (SVM) is a linear classification algorithm that is able to identify the separate hyper plane that maximizes the margin between the two classes. In addition, Support Vector Machine uples are also popularly used for text classification and has a good performance.

Wrapper method has been shown to improve the accuracy of the classifier Support Vector Machine (SVM) in terms of data processing. Data can be classified tweet well into positive and negative forms. Accuracy Support Vector

Machine (SVM) before using the merger method of Genetic Algorithm akirasi wrapper generated is equal to 63.50% and the AUC value 0.875. Meanwhile, after using the merger method of Genetic Algorithm, accuracy increased to 93.50% and the AUC value 0.940. Improved accuracy reaches 30%. To support this research developed sentiment analysis applications classify positive and negative tweets using the programming language PHP.

The model established in this study applied on twitter Pizza Hut as a determinant of corporate strategy decision of the opinions expressed by the customer, so that this research can be used also for other organizations and companies that have a twitter account to analyze sentiment happens to the product them to determine subsequent decisions and strategies to be followed.

REFERENCES:

[1] Aydin, I., Karakose, M., & Akin, E. (2011). A multi-objective artificial immune algorithm for parameter optimization in support vector machine. *Applied Soft Computing*, *11*(1), 120–129. doi:10.1016/j.asoc.2009.11.003

[2] Balahur, A., Mihalcea, R., & Montoyo, A. (2014). Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech & Language*, *28*(1), 1–6. doi:10.1016/j.csl.2013.09.003

[3] 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. doi:10.1016/j.proeng.2013.02.059

[4] Campo-Ávila, J. Del, Moreno-Vergara, N., & Trella-López, M. (2013). Bridging the Gap Between the Least and the Most Influential Twitter Users. *Procedia Computer Science*, *19*(Ant), 437–444. doi:10.1016/j.procs.2013.06.059

[5] Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, *36*(3),5432–5435

[6] Chou, J.-S., Cheng, M.-Y., Wu, Y.-W., & Pham, A.-D. (2014). Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification. *Expert Systems with Applications*, *41*(8), 3955–3964. doi:10.1016/j.eswa.2013.12.035

[7] Dehkharghani, R., Mercan, H., Javeed, A., & Saygin, Y. (2014). Sentimental causal rule discovery from Twitter. *Expert Systems with Applications*, *41*(10), 4950–4958. doi:10.1016/j.eswa.2014.02.024

[8] Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, *40*(16), 6266–6282. doi:10.1016/j.eswa.2013.05.057

[9] Gorunescu, F. (2011). Data Mining Concept Model Technique.

[10] Gunal, S. (2012). Hybrid feature selection for text classification ¨, *20*.

[11] Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science*, *17*, 26–32. doi:10.1016/j.procs.2013.05.005

[12] Han, J., & Kamber, M. (2007). Data Mining Concepts and Techniques

[13] Ilmiah, D. K., Akhir, T., Studi, P., Informatika, T., Komputer, F. I., Dian, U., Udinus, P. S. I. (2014). Dokumen Karya Ilmiah | Tugas Akhir | Program Studi Teknik Informatika - S1 | Fakultas Ilmu Komputer | Universitas Dian Nuswantoro Semarang | 2014, 0–1.

[14] Khan, F. H., Bashir, S., & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, *57*, 245–257. doi:10.1016/j.dss.2013.09.004

[15] Khan, K., Baharudin, B., & Khan, A. (2014). Mining Opinion Components from Unstructured Reviews: A Review. *Journal of King Saud University - Computer and Information Sciences*. doi:10.1016/j.jksuci.2014.03.009

[16] Larose, D. T. (n.d.). *METHODS AND*.

[17] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*. doi:10.1016/j.asej.2014.04.011

[18] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña-López, L. A. (2014). Ranked WordNet graph for Sentiment Polarity Classification in Twitter. *Computer Speech & Language*, *28*(1), 93–107. doi:10.1016/j.csl.2013.04.001

[19] Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621–633.

[20] Passonneau, R. (2011). Sentiment Analysis of Twitter Data. *Proceedings of the Workshop on*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Languages in Social Media, (ISBN: 978-1-932432-96-1), 30–38.

[21] Schoefegger, K., Tammet, T., & Granitzer, M. (2013). A survey on socio-semantic information retrieval. *Computer Science Review*, *8*, 25–46. doi:10.1016/j.cosrev.2013.03.001

[22] Suyanto. (2014). Artificial Intelligence Searching-Reasoning-Palnning-Learning. Bandung: Informatika Bandung.

[23] Wang, Z., Shao, Y.-H., & Wu, T.-R. (2013). A GA-based model selection for smooth twin parametric-margin support vector machine. *Pattern Recognition*, *46*(8), 2267–2277. doi:10.1016/j.patcog.2013.01.023

[24] Weise, T. (2009). *Global Optimization Algorithms – Theory and Application –* (Second Edi.). it-weise.de (self-published).

[25] Widodo, Prabowo Pudjo, Rahmadya Trias Handayanto dan Herlawati. (2013). Penerapan Data Mining dengan Matlab. Bandung: Rekayasa Sains.

[26] Witten, H. I., Frank, E., & Hall, M. A. (2011). Data Mining Practical MachineLearning Tools And Technique. Burlington: Elsevier Inc.

[27] Wu, Xindong, Vipin Kumar. (2009). The Top Ten Algorithms in Data Mining.

[28] Yang, P., Liu, W., Zhou, B. B., Chawla, S., & Albert, Y. (2010). Ensemble-based wrapper methods for feature selection and class imbalance learning, 1–12.

[29] Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, *36*(3), 6527–6535. doi:10.1016/j.eswa.2008.07.035

[30] Zhao, M., Fu, C., Ji, L., Tang, K., & Zhou, M. (2011). Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. *Expert Systems with Applications*, *38*(5), 5197–5204. doi:10.1016/j.eswa.2010.10.041

[31] Zukri, Zainudin. (2014) Algoritma Genetika Metode Komputasi Evolusioner untuk Menyelesaikan Masalah Optimasi. Yogyakarta: Andi Offset.