Sentiment Analysis Model Based On Youtube Comment Using Support Vector Machine

Fiktor Imanuel Tanesab[1], Irwan Sembiring[2] and Hindriyanto Dwi Purnomo[3]

[1, 2, 3] Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana, Jl. Diponegoro 52-60, Salatiga 50711, Indonesia

[1]viktortanesab@gmail.com, [2]irwan@staff.uksw.edu, [3]hindriyanto.purnomo@staff.uksw.edu

ABSTRACT

Opinion mining or comment toward attitude evaluation, individual entity, are usually called sentiment. Everyone is free to give opinion related with the present opinions on youtube. Hence people have a free will to express their opinion regarding the performance. Due to the raise of many critics that appear in a short amount of time, there a needs to conduct research on opinion mining. In this research, opinion mining is applied on the peformance of Ahok as a governor. The sentiment analysis is used to find a pattern or a certain character of Ahok. Support Vector Machine is used to classified the opinion into positive class, neutral class and negative class. 1000 recorded data is used as a sample data. Preprocessing phase is needed before classifying the data. The preprocessing phase consist of preprocessing the data, tokenizing, cleansing and filtering. In order to determine the percentage of the class sentiment, Lexicon Based method is used. The experiment shows that the proposed method are calculating the percentage weight in this research had used Lexicon Based and Confusion Matrix to know the result of weighting percentage of analysis to SVM. It had been found the result as follows : accuracy 84%, precision 91%, recall 80%, TP rate 91.1 and TN rate 44.8%.

Keywords: *Youtube, Analysis Sentiment, Support Vector Machine, Opinion Mining, Lexicon Based.*

1. INTRODUCTION

Opinion mining is a form of opinion analysis towards a pattern or mood of a person or a certain topic, these things are often called sentiment.[1] In order to filter every statement or opinion sentiment from the public, the most and easy media that is being used is the internet. Many comments on Ahok's remarks have been made on the youtube. The process of searching or tracing the natural language to find patterns or moods of society against certain products, people or topics is called Sentiment Analysis. Sentiment analysis is also often referred to as the opinion of mining.[1] The sentiment analysis has received considerable attention since the research of Pang, Turney, Goldberg and Zhu. Sentiment analysis techniques can support many decisions in many scenarios. This study uses three class attributes, which are positive, neutral and negative, because in the internet the comments that appear can be positive, neutral and negative comments.[2]

The result of the existing information from the existing opinion data, then performed data processing with sentiment analysis. Sentiment Analysis to be the right solution or method to the problems that have been described previously, that is by separating the opinion into a positive, a neutral or a negative, so it is concluded that opinion mining is more dominant in this research. Related to this research, the limitation of data retrieval time is from 2015 until 2016.

The Support Vector Machine (SVM) is suggested in this research. The data used in this research is data of comment on social media youtube and those are taken from Ahok's performance. SVM has some determined criteria to see the accuracy level of Governor Ahok's performance determination.

2. RELATED RESEARCH

Sentiment Analysis describes its keywords "opinion mining" in two ways, namely 1). Analysis and evaluative text, and 2). Predictive rating tracking.[3] Sentiment analysis is a process of picking up and favorability of a natural language.[3] Research A survey of opinion mining and sentiment analysis by Bing Li and Lei Zhang from the University of Illinois at Chicago suggests that sentimental analysis or opinion mining is a study of people's opinion on attitude, emotions, entities, individuals, issues, events, topics and attributes.[4] Sentiment analysis has a purpose to process a decision-making by extracting textual opinions containing positive, neutral and negative opinions by counting the weight of opinions contained.[4][5]. With the opinion mining (analytical sentiment) can extract the data in textual and automatically process the data so that it gets an information of opinion or a particular event.[4][5][6].

Support Vector Machine (SVM) is a very rapidly growing method and is widely used in research studies related to analysis sentiment. Support Vector Machine is a methodological study of opinion mining and sentiment analysis techniques.[7] Opinion extraction of public figures based on sentiment analysis in twitter.[8] Recognizing contextual polarity an exploration of features for phrase-level sentiment analysis.[9] The importance of neutral examples for learning sentiment.[10] Support vector machines supported k-means clustering in document classification.[11]

Based on the related research, the researchers proposed the method of Support Vector Machine (SVM) to see the performance of Ahok. SVM is used in the process of data retrieval, analysis, until the conclusion.

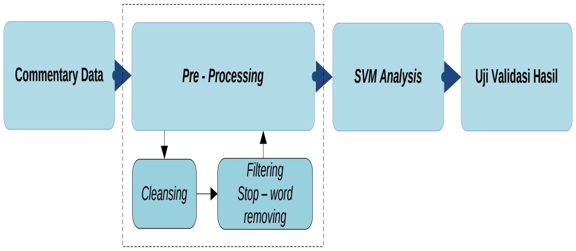


Fig 1. The process of Sentiment Analysis

3. MODEL TO ANALYZE THE SENTIMENT

At this stage of this research, the data was gained by using snipping data as many as 1000 record data. They were taken randomly and in accordance to the performance of Ahok. The data's are taken from 2015 until 2016 shortlists. Those data were taken and classified into three classes of positive opinion, neutral opinion and negative opinion. The data were extracted and analyzed using SVM method. The use of data in the research is the text data of comments in Indonesian and retrieved from YouTube.

A previous action was finishing the process of grouping the documents, such as preparing the data text which is often called pre-processing. There are several stages, such as 1). Comment Data; 2). Pre-Processing; 3) Tokenizing; 4) Determine Sentiment with Lexicon Based.[12]

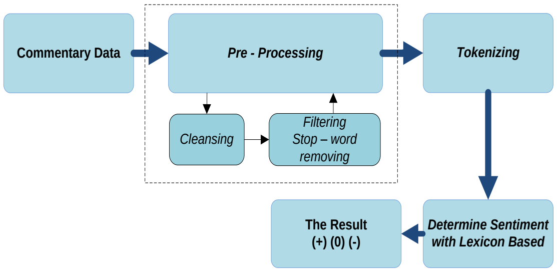


Fig. 2. The model and finishing stage at the SVM Sentiment Analysis.

1) Commentary data: comments data were obtained by snipping method from YouTube. The snipped data is the commentary authentically spoken in Indonesian. Dataset uses as many as 1000 text comments and related to the performance of Ahok.
2) At the pre-processing stage, there are some steps of cleansing and filtering. The process of sorting a sentence aims to reduce the raw sentence so that the required data can be gained through this stage.[14] Cleansing is also has a purpose to remove links and symbols.[15] The process of retrieving the important words is done by Stop-word removing, often also referred to as irrelevant term with the main subject of the database, even though the word often appears in the document.[16] The following are examples of some stop-word removing sentences. [14][15][16]

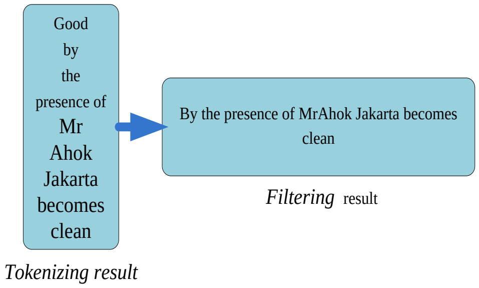


Fig. 3. Stop-word removing

The cleansing model will be described as follows:

Table 1: Cleansing process

Comments	Cleansing result
Whoaa...persistent ....and transparent.... Mr. Ahok	Mr. Ahok is persistent and transparent

3) Tokenizing : A process of cleaning and cutting some words based on the word types intended for punctuation, space, and omitted if there are symbols or anything that is not bound to letters.[17] [20]

Table 2. Tokenizing.

Comments on a text	Tokenizing result
By the presence of Mr. Ahok, Jakarta becomes clean	The presence of Mr Ahok Jakarta becomes clean

4) Determine Sentiment with Lexicon Based : The purpose of lexicon based method is to determine the sentiments of an opinion sentence that appears. The process of determination is done by summing *n* the polarity score of the opinion word, *p* that commented on the feature *f*. The polarity score of an opinion word p will be 1 if it is a positive opinion word, and is worth -1 if the word is a negative opinion word, and is based on a lookup table.[17] Words containing positive, negative and neutral in a sentence, will be calculated by the weight of the value contained in the sentence is done by summing the value of the opinion word that appears. If the value of the opinion in the sentence is 1, the sentence value of the sentence is positive, if the value of the sentence in the sentence is 0, the sentence value of the sentence is neutral, if the value of the sentence in the sentence is -1, and the sentiment value of the sentence is negative.[17]

Table 3: Lexicon Based Process

Sentiment	Value
Positive	1
Neutral	0
Negative	-1

Based on four basic SVM criteria in table 3, it will be evaluated, Precision and Recall performance of the experiments that have been tested in order to predict the correct data and false data.[19] The evaluation was done with Confusion Matrix where True Positive rate (TP rate) is a positive class that successfully classified as positive class, True Negative rate (TN rate) is a negative class that successfully classified as class negative, False Positive rate (FP rate) is a negative class Classified as positive and neutral classes, False Negative rate (FN rate) is a positive class that is classified as a negative and neutral class.[17]

Table 4: Confusion Matrix

		Real Value	
		True	False
Predicted Value	True	TP (True Positive) Correct result	FP (False Positive) Unexpected result
	False	FN (False Negative) Missing result	TN (True Negative) Correct absence of result

(1)

On the equivalent of Confusion Matrix (1) the calculation of the real percentage are based on the obtained results, where the model of the calculation refers to the results of weighted sentence. The sum of the data were calculated using the lexicon based method, so as to precision, recall and accuracy. The result can be identifies as in the formula below.

$$\text{percison} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{TR rate} = \frac{\text{TP}}{\text{Total positive sesungguhnya}}$$

$$\text{TR rate} = \frac{\text{TN}}{\text{Total negative sesungguhnya}} \tag{3}$$

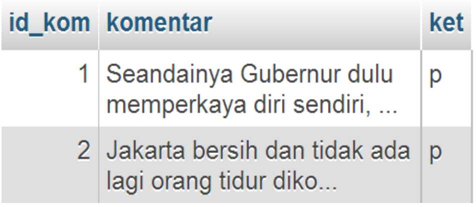
The equation process(2) is done through calculation of weighting accuracy, in order to see the result of a True Positive rate (TP rate) and a True Negative rate (TN rate) so the result of Ahok's performance analysis could be identified. Equation (3) is aimed to calculate the percentage weight of Ahok's performance.

4. RESULT AND DISCUSSION

The choice in the process of data retrieval is a text which criticize the performance Ahok as a governor, which was found on YouTube. The process of taking a certain comment from the text had used a snipping method. The data is taken randomly based on recent video results that appears in YouTube. The data is grouped into three classes namely, positive class, neutral class and class negative after the data was snipped from the source. The process of completing this research is using the Support Vector Machine method.

The process of classification in this research are described in several stages, such as 1). Data preparation. The data used is a text data file. Each record of the data is based on each class (negative, neutral and positive). 2). Pre-processing, is stop - word remove (cleansing) or often referred as normalization, with the aim of removing or changing every sentence that is not bound to standard sentences. 3). Tokenizing is a process of cleansing and cutting words based on word types which is arranged to remove the unnecessary parts. 4). The lexicon based process is used to determine the sentiments of an opinion sentence. Determination is done by summing the n polarity score of the opinion word p that commented on the feature f. The polarity score of an opinion word p will become 1 if it is a positive opinion word, worth -1 if the word is a negative opinion word and is 0 if the word is neutral. [17]

For example, the output of the snipping the text data of comments from YouTube that has been done and has been stored in the database. As shown in Figure 5 below.



id_kom	komentar	ket
1	Seandainya Gubernur dulu memperkaya diri sendiri, ...	p
2	Jakarta bersih dan tidak ada lagi orang tidur diko...	p

Fig. 4. The result of snipping the comments

4.1 Implementation of Support Vector Machine method

The dataset in this study was collected from YouTube with snipping method of comments in the text. The data is taken in the form of Indonesian text, which is a comment with the performance keywords of Ahokas a governor. Data is taken randomly from YouTube.

The dataset used is 1000 text comments, the data is divided equally (balanced) of each class, because with unbalanced data, the constructed classification has a tendency to ignore the minority class. [17] Data is divided into three classes namely, positive, neutral and negative. Labeling is done by lexicon based method.

4.2 Lexicon Based Analysis

The result of sentiment analysis performed by Ahok using lexicon based method has been classified into three class attributes.

Table 5: Lexicon Based Analysis Result

Sentiment Performance		
Sentiment	Result	Real data
Positive	678	744
Neutral	192	22
Negative	105	234
Matched data		805
Unmatched data		195

4.3 Confusion Matrix Analysis

The calculation of accuracy value of analysis towards the SVM method’s result that was done using need to have the accuracy, precision and recall performance evaluation from the experiment with confusion matrix method. The evaluation is done using Confusion Matrix includes the True Positive Rate (TP rate), True Negative Rate (TN Rate), False Positive Rate (FP Rate) and False Negative Rate (FN rate) as an indicator. The TP rate is the percentage of the positive class which was classified as the positive class, whereas the TN rate is the percentage of the class negatively classified as a negative class. FP rate is class negative which is classified as positive class. The FN rate is a class positive that is classified as a negative class. [18]

Table 6: Confusion Matrix Result Analysis

		Real Value	
		Positive	Negative
Prediction	Positive	tp678	fp129
	Negative	fn66	tn105

In the confusion matrix equations (1) a real percentage is calculated based on the obtained results. In equation (2), an accurate weighting calculation is performed to see the results of the true positive rate (TP rate) and the True Negative rate (TN rate). [18] The result of his analysis is the performance of Ahok is as follows:

Table 7: The classified results of Ahok’s Sentiment Performance

Sentiment Classified Result				
Accuracy (%)	Precision (%)	Recal l (%)	TP rate (%)	TN rate (%)
84%	91%	80%	91.1%	44.8%

Table 7.providesinformation on the accuracy, recall, TP rate and TN rate of each test result data based on 1000

comment data. From the process of data pre-processing produces a number of tokens which are then used as input classification process using SVM method.

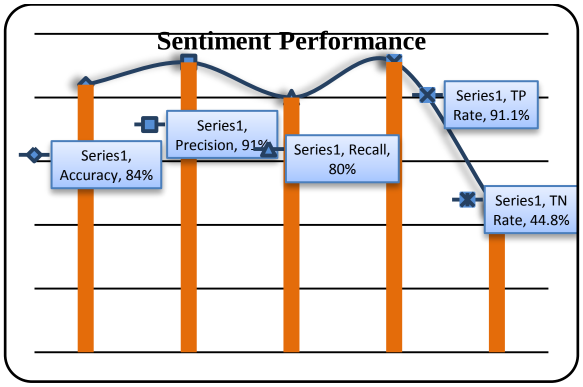


Fig. 5. Ahok's Sentiment Performance

In Figure 6.The accuracy, precision, recall, TP rate and TN rate values were tested with the Support Vector Machine (SVM) method. Value of accuracy 84%, precision 91%, recall 80%, TP rate 91.1% and TN rate 44.8%. TN rate is inversely proportional to TP rate, where TN rate is obtained percentage 44.8%, while TP rate 91.1%

Accuracy value is one of the assessment parameters of the method that has been used, the accuracy value obtained from the amount of data (1000 data) successfully classified correctly according to the class of sentiment of the total amount of data classified.

5. CONCLUSION

In summary, Support Vector Machine (SVM) is being used to view the performance sentiment analysis of Ahok. There are four processes that has been conducted namely, Data Comments, Pre-Processing, Tokenizing and Determine Sentiment with Lexicon Based. [12] Calculating the percentage weight in this research had used Lexicon Based and Confusion Matrix to know the result of weighting percentage of analysis to SVM.

Sentiment analysis can be used to find out how far the performance of Ahok based on the results gained from netizen's comment on YouTube. The result of classification of weighted values according to Support Vector Machine (SVM) method has brings us to the conclusion that the value of True Positive rate is 91.1% based on the comments taken from 2015 until 2016.

For the further development of this research, further researchers need to take into account that data recording should be high in numbersto achieve the accuracy of the results and conclusions on the opinion mining analysis.

REFERENCES

[1] G. Vinodhini., RM. Chandrasekaran ., "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Issue Volume 2, 2012.

[2] Anto Satriyo Nugroho., Arief Budi Witarto., Dwi Handoko., "Support Vector Machine – Teoridan Aplikasinya dalam Bioinformatika1", 2003.

[3] Bo Pang., Lillian Lee., "Opinion Mining and Sentiment Analysis" Vol 2, 2008.

[4] Bing Li,. Lei Zhang., "A survey of opinion mining and sentiment analysis" University of Illinois at Chicago, 2012.

[5] Aamera Z.H. Khan,. Mohammad,. Atique,. V. M. Thakare,. "Combining Lexicon-based and Learning -based Methods for Twitter Sentiment Analysis" National Conference on Advanced Technologies in Computing and Networking –ATCON, 2015.

[6] Imam Fahrur Rozi., Sholeh Hadi Pramono., Erfan Achmad Dahlan., "Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi" Jurnal EECCIS Vol. 6, 2012.

[7] M.Tech., "Methodological Study Of Opinion Mining And Sentiment Analysis Techniques", International Journal on Soft Computing (IJSC) Vol. 5, No. 1, 2014.

[8] Nur Hayatin., Mustika Mentari., AbidatulIzzah., "Opinion Extraction of Public Figure Based on Sentiment Analysis in Twitter", Journal of Engineering, Vol. 1, No. 1, 2014.

[9] Theresa Wilson., Janyce Wiebe,. Paul Hoffmann,. "Recognizing Contextual Polarity, Eksplorasi Fitur Untuk Frase – Level Analisis Sentimen", Association for Computational Linguistics, Vol. 35, 2008.

[10] Moshe Koppel., Jonathan Schler,. "The Importance Of Neutral Examples For Learning Sentiment", Computational Intelligence, Vol. 22, Number 2, 2006.

[11] Ahmad Yusuf., Tirta Priambadha., "Support Vector Machines Yang Didukung K-Means Clustering Dalam Klasifikasi Dokumen", Jurnal Ilmiah Teknologi Informasi (JUTI), Vol.11, No. 1, 2013.

[12] Oman Somantri., Slamet Wiyono., Dairoh., "Optimalisasi Support Vektor Machine (SVM) UntukKlasifikasiTemaTugasAkhirBerbasis K-Means" Telematika, Vol. 13, No. 02, Pp. 59 – 68 ISSN 1829-667X, 2016.

[13] Ni Wayan Sumartini Saraswati., "Text Mining Dengan Metode Naïve Bayes Classifier Dan Support Vector Machines Untuk Sentiment Analysis", Universitas Udayana Denpasar, 2011.

[14] Falahah., Dyar Dwiki Adriadi Nur., "Pengembangan Aplikasi Sentiment Analysis Menggunakan Metode Naïve Bayes (Studi Kasus Sentiment Analysis dari media Twitter) "Seminar Nasional Sistem Informasi Indonesia, 2015.

[15] Nuvirta Monarizqa., Lukito Edi Nugroho., Bimo Sunarfri Hantono., "Penerapan Analisis Sentimen Pada Twitter Berbahasa Indonesia Sebagai Pemberi Rating", Artikel Reguler, Vol. 1, nomor3, 2014.

[16] Elly Susilowati., Mira Kania Sabariah., Alfian Akbar Gozali, "Implementasi Metode Support Vector Machine Untuk Melakukan Klasifikasi Kemacetan Lalu Lintas Pada Twitter", ISSN : 2355-9365 e-Proceeding of Engineering : Vol.2, No.1 Page 1478, 2015.

[17] Ghulam Asrofi Buntoro., "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter", Universitas Muhammadiyah Ponorogo, Article, 2017.

[18] LuhRia Atmarani., I.A. Dwi Giriantari., Made Sudarma., "Sistem Opinion Mining dengan Metode Pos Tagging dan SVM Untuk Ekstraksi Data Opini Publikpada Layanan JKBM" Teknologi Elektro, Vol. 16, No1, 2017.

[19] Ika Menarianti., "Klasifikasi Data Mining Dalam Menentukan Pemberian Kredit Bagi Nasabah Koperasi" Jurnal Ilmiah Teknosains, Vol. 1 No. 1, ISSN 2460-9986, 2015.

[20] Lila Dini Utami., Romi Satria Wahono., "Integrasi Metode Information Gain Untuk Seleksi Fiturdan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes", ISSN 2356-3982 Journal of Intelligent Systems, Vol. 1, No. 2, 2015.

AUTHOR PROFILES:

Fiktor Imanuel Tanesab, was born in Kefamenanu, East Nusa Tenggara. Recently graduated as Bachelor of Computer Science from STIKOM Uyelindo Yelindo Kupang, Indonesia. Fiktor is currently finishing his post-graduate study in Information System master program in the Faculty of Technology and Information, Satya Wacana Christian University.

Irwan Sembiring, Completed his undergraduate program in UPN "Veteran" Yogyakarta, majoring in Information Technology in 2001, pursued higher degree in School of Computer Science and Electronics Gadjah Mada University, Yogyakarta, Indonesia and received Master Computer in 2004. Doctor in Computer Sciences from Gadjah Mada University, Yogyakarta, Indonesia, Now he is a lecturer at faculty of information technology Satya Wacana Christian University, Salatiga Indonesia. His research interests include Network Security and Digital Forensic.

Hindriyantodwi Purnomo,is a senior lecturer at Satya Wacana Christian University, Indonesia. He received his doctoral degree from Department of Industrial and System Engineering, Chung Yuan Christian University, Taiwan. His research interests are in the field of applied soft computing and numerical optimization. His research has been published in several reputable international journals.