IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

# Comparison Of Classification Methods On Sentiment Analysis Of Political Figure Electability Based On Public Comments On Online News Media Sites

To cite this article: K Sigit *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **662** 042003

IOP Conf. Series: Materials Science and Engineering **662** (2019) 042003 doi:10.1088/1757-899X/662/4/042003

# Comparison Of Classification Methods On Sentiment Analysis Of Political Figure Electability Based On Public Comments On Online News Media Sites

**Sigit K**[1], **Dewi A P**[2], **Windu G**[3], **Nurmalasari**[4], **Muhamad T**[5], **Kadinar N**[6]

[1,2,3,4,5] STMIK Nusa Mandiri, Jakarta, Indonesia.
[6] Universitas Bina Sarana Informatika, Jakarta, Indonesia.

E-mail: Sigit024@gmail.com

**Abstract.** Elections are an important part of the political process so that not a few figures from political parties begin preparing to participate in the process. Electability is one of the issues of concern, various things are done to be able to increase the electability of political figures who participated in the general election. Media has become one of the important tools used to increase electability, one of which is online news media. News coverage in online news media with its real-time nature will very quickly get comments from readers and can be used as an assessment of political figures in the form of sentiment analysis. However, it is not easy to analyze sentiments from various comments on online news media, because comments that contain text have irregularities, especially in Indonesian texts. Text mining is one way that can be used to overcome this. Pre-processing of text in text mining is an important part of getting basic information contained in comments. This study uses the Indonesian text pre-processing using the Gata text-mining framework. Then proceed with extracting more in-depth information using the Naïve Bayes classification method and Support Vector Machine which is optimized using Particle Swarm Optimization. The tests carried out with both methods get the results that, the Particle Swarm Optimization based Support Vector Machine method is the best method in the process of classifying sentiments analysis of political figures with an accuracy of 78.40% and AUC 0.850. The results of this study get an effective algorithm in classifying positive and negative comments related to political figures from online news media.

## 1. Introduction

Elections are an important part of the political process so in Indonesia. being a general election in Indonesia in 2019, not a few political figures begin to get ready to participate in the process. Various attempts by political leaders were made to increase their electability. Electability is full of popularity, a popular person will have high electability [1]. In politics, the role of electability is always challenging to convince voters to choose the political figure they like [2]. The tendency of electability is the most during the election season political compilation is not popular with general elections [2] and can be a benchmark for measuring one's level of electability [1].

Sentiment analysis is a process of comparison to uncover and categorize opinions expressed in text, to determine whether responses to a particular topic are positive, negative, or neutral [3][4]. Sentiment analysis also called mining is a field of study that analyzes opinions, sentiments, opinions, judgments, and emotional judgments on attribute entities discussed in a paper [5]. Analyzing sentiments in political texts can be a representation of election results, the electability of candidates, or real political representation in the State [3]. Related to web technology, followed by political activity in online media is an aspect that is a consideration to analyze the sentiments of the electability of political figures through written posts and comments. Research related to sentiment analysis of political figures originating from online media is mostly done by researchers with various methods used [6][7][8][9]. Conducting sentiment analysis originating from text posting and commenting on online media is not easy, because comments that contain text have irregularities, depending on Indonesian texts that have distinctive features from other languages [10].

This study applies machine learning technology to analyze electability sentiments of political figures originating from comment text data on online news media in Indonesia. Text mining is one method used to extract unstructured information from combining text data [11], which is converted into a numerical value that will be used in various data mining methods [12]. In addition to using the text-mining method, data mining methods with group functions are also used for text data that has been extracted with the aim of getting complete information. The Naïve Bayes algorithm [13] and Support Vector Machine [14] optimized with Particle Swarm Optimization [15] are data mining methods for the classification used in this study.

The main purpose of this research is to learn and get the best and decisive method in classifying and analyzing sentiments. Analysis of reader comments on online news media on the electability of political figures seeking the electoral process. This research is also expected to provide solutions in terms of the sentiment classification analysis of political figures and support the discussion of further research on other research objects, such as sentiment analysis in the level of customer satisfaction in the company.

There are several previous studies that used a Naïve Bayes and Support Vector Machine classification model, including:

1. Conducted a sentiment analysis of the 2017 DKI Jakarta Governor Election which was widely discussed in cyberspace, especially on Twitter social media. The method used in this study, for data preprocessing using tokenization, cleansing, and filtering, to determine the sentiment class with the Lexicon Based method. The classification process uses the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM) methods [16].
2. Conducted a study on the analysis of President Jokowi's data sentiments with preprocessing, normalization and stemming using the Naïve Bayes and SVM methods. Search techniques in this study use Boolean searching with the operator "AND". The data that has been obtained is labeled positive, neutral and negative by the researcher then corrected by linguists. After that, preprocessing is done, either changing the word is not standard into standard or commonly called normalization using a dictionary and search for the root word, stemming with the help of the Literary Master application. Furthermore, it is also done tokenization of N-Gram, Unigram, Bigram, and Trigram against the sentence, then removes the commonly used words and does not have valuable information in a context or commonly called stopword removal, and retains emoticons because emoticons are symbols that show expression someone into writing [6].
3. Conducted a sentiment analysis study on the 2017 DKI Jakarta Governor Election with a predictive and descriptive approach. The main objective of this research is to analyze the sentiments of the 2017 DKI Jakarta Governor election on Twitter social media with a predictive and descriptive approach. Data is collected from Twitter using each candidate's username as a search query. For predictive approaches, machine learning algorithms such as the Naïve Bayes Multinomial and Support Vector Machine are used to classify datasets. While the descriptive approach, Time Series Graph and Word Clouds are used to gain deeper insights and find a link between Twitter sentiment and the results of the election itself [17].

## 2. Methods

Research on the comparison of classification methods in the analysis of Indonesian social media sentiment related to the election of the governor of DKI Jakarta. Sentiment analysis has challenges such as comments usually ambiguous, language barrier problems, harsh words, excessive comments, and sentiment clauses. This research tries to distinguish the issue of sentiment classification from Indonesian social media regarding governor elections in DKI Jakarta. Several steps are taken to overcome the problem that includes preprocessing. Preprocessed sentiments are then grouped into positive, negative, and neutral. The classification methods used in this study are the Summation Method, Average on Tweet, Average on Tweet with the threshold on objective scores, Weighted Average, and Naïve Bayes. The experimental results show that Naïve Bayes produces high precision values, high recall values and high accuracy values for neutral and positive sentiments, but Naïve Bayes does not produce good results for negative sentiments [18]. Research on comparative orientation analysis using the technique of Support Vector Machine and Naïve Bayes on film user reviews. Sentiment analysis is used to recognize and know the users' responses to a film. Data is collected through user movie reviews online in the Internet Movie Database archive. Data processing is done by using RapidMiner Tools. The results of the evaluation of Support Vector Machine and Naïve Bayes algorithms in the movie user review dataset, it is shown that drama films have a high degree of accuracy among other film genres. The evaluation results show that Support Vector Machine techniques have the best accuracy than Naïve Bayes [19].

## 3. Results And Discussion

In this study using the Cross-industry standard process for data mining, known as CRISP-DM method, the CRISP-DM stages in Figure 1 [20], above are as follows:


**Figure 1.** Method of CRISP Data Mining.Figure was adopted from Kenneth Jensen, 2016.

### *3.1 Business Understanding*

At the business understanding stage, an understanding of the object of research is carried out. Understanding of the object of research is done by digging up information through several online news media sites on the object of research. Motivation in this phase is that the news presented is usually in the form of text on digital media which is grouped based on the contents of the discussion of each news category. Online media is not only limited as a medium for reading news articles but can also be used to see the issues that occur and can even be used to see the electability of a political figure. This sentiment analysis is done to find classification methods that can help determine positive and negative news article comments. At this stage, an understanding is also made to find the best classification method so that it can help during the data processing process which will be done by comparing the results of the algorithms used and to improve the performance of the classification method can be done by using feature selection.

### *3.2 Data Understanding*

At the data understanding stage, the process of retrieving raw data is carried out in accordance with the required attributes. Data obtained from news media sites online www.detik.com, www.tribunnews.com, www.kompas.com, www.merdeka.com, www.viva.co.id and www.kompasiana.com. The data were collected from February 16, 2018, until May 31, 2018. The primary data obtained were 1480 commentary data from news articles on online news media, namely Detiknews with 1393 comments data, Tribunnews with 8 comment data, Kompas with 72 comment data, Merdeka with 3 comment data, Viva with 3 comment data and Kompasiana with 1 comment data. After the data is collected then the cleansing process is carried out, obtained 1459 comments data. This study uses comment data in Bahasa (Indonesian language).

### *3.3 Data Preparation*

The data preparation stage is a stage with a data preparation process that aims to obtain clean data and is ready for use in research. In the early stages of text mining, the text preprocessing stage will be performed, at this stage, researchers will use the Tools Gata Framework Textmining. The following are the steps taken in the text preprocessing.

#### *3.3.1 Case Folding*

Not all text documents are consistent in using capital letters. Therefore, the role of Case Fold is needed in converting all text in a document into a standard form (usually lowercase or lowercase letters), as shown in table 1.

**Table 1.** Comparison of Text Before and After the Case Folding Process

| Before | After |
|---|---|
| "DARIMANA LOGIKANYA SI AHY PALING CUCOK DAN ELEKTABILITAS TINGGI UTK CAWAPRES JOKOWI.. ANAK MASIH HIJAU DAN TIDAK PUNYA PENGALAMAN POLITIK DAN EKSEKUTIF NOL BESAR !!! NGACO..!!! #2019TETAPJOKOWI" | "darimana logikanya si ahy paling cucok dan elektabilitas tinggi utk cawapres jokowi.. anak masih hijau dan tidak punya pengalaman politik dan eksekutif nol besar !!! ngaco..!!! #2019tetapjokowi" |

#### *3.3.2 Tokenizing*

The Tokenizing step is the step of cutting the input string based on each word that composes it, as seen in table 2.

**Table 2.** Comparison of Texts Before and After the Tokenizing Process

| Before | After |
|---|---|
| "darimana logikanya si ahy paling cucok dan elektabilitas tinggi utk cawapres jokowi.. anak masih hijau dan tidak punya pengalaman politik dan eksekutif nol besar !!! ngaco..!!! #2019tetapjokowi" | "darimana logikanya si ahy paling cucok dan elektabilitas tinggi utk cawapres jokowi.. anak masih hijau dan tidak punya pengalaman politik dan eksekutif nol besar !!! ngaco..!!! #2019tetapjokowi" |

*3.3.3 Tagging*

Eliminate the mention of the user name mentioned (@username), delete hashtag (#hastag), delete punctuation, delete numbers because only text data is used, delete links (http: //), delete special characters such as symbols or emoticons, omit words foreign words because they only take words that speak Indonesian. For the word ahead, the word "no" will be normalized by using the underscore "_" to have a clear meaning, for example, "not good" to "not_good" because the word "no" is negative while the word "good" is positive. the results of the tagging process can be seen as in table 3.

**Table 3.** Comparison of Texts Before and After the Tagging Process

| Before | After |
|---|---|
| “darimana logikanya si ahy paling cucok dan elektabilitas tinggi utk cawapres jokowi.. anak masih hijau dan tidak punya pengalaman politik dan eksekutif nol besar !!! ngaco..!!! #2019tetapjokowi” | “darimana logikanya si ahy paling cucok dan elektabilitas tinggi utk cawapres Jokowi anak masih hijau dan tidak_punya pengalaman politik dan eksekutif nol besar ngaco” |

*3.3.4 Filtering*

Filtering Stage is the stage of taking important words from the token results. Can use a stoplist algorithm (throwing less important words) or wordlist (save important words). At this stage for slang words (in Indonesian) must be normalized into the standard form, for example, "to" be "for". The results of the filtering process can be seen in table 4.

**Table 4.** Comparison of Texts Before and After the Filtering Process

| Before | After |
|---|---|
| “darimana logikanya si ahy paling cucok dan elektabilitas tinggi utk cawapres Jokowi anak masih hijau dan tidak_punya pengalaman politik dan eksekutif nol besar ngaco” | “darimana logikanya si ahy cocok elektabilitas cawapres Jokowi hijau tidak_punya pengalaman politik eksekutif nol ngaco” |

*3.3.5 Stemming*

The stemming process is the process of finding the root of the word resulting from the filtering process. The root search for a word or commonly referred to as a basic word can reduce index results without having to eliminate meaning. The results of the stemming process can be seen in table 5.

**Table 5.** Comparison of Texts Before and After the Stemming Process

| Before | After |
|---|---|
| “darimana logikanya si ahy cocok elektabilitas cawapres Jokowi hijau tidak_punya pengalaman politik eksekutif nol ngaco” | “darimana logika si ahy cocok elektabilitas cawapres Jokowi hijau tidak_punya pengalaman politik eksekutif nol ngaco” |

The data preparation process uses the Gata Framework Textmining pre-processing tool, as shown in figure 2.

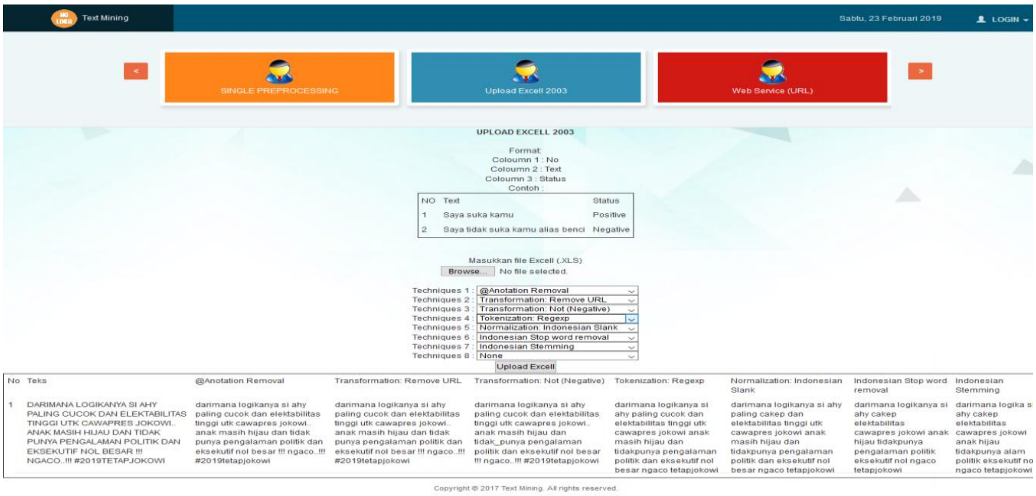


**Figure 2.** Gata Framework Textmining. http://www.gataframework.com/textmining/

### *3.4 Modelling*

Is the mining technique selection phase by determining the algorithm to be used. The tool used is RapidMiner version 8.2. The model test results are classifying positive news articles and negative news articles using the Naive Bayes and Naive Bayes algorithm based on Particle Swarm Optimization (PSO) and Support Vector Machine and Support Vector Machine based on Particle Swarm Optimization (PSO) to get the best accuracy value. In Figure 3 the design of the RapidMiner model is used.

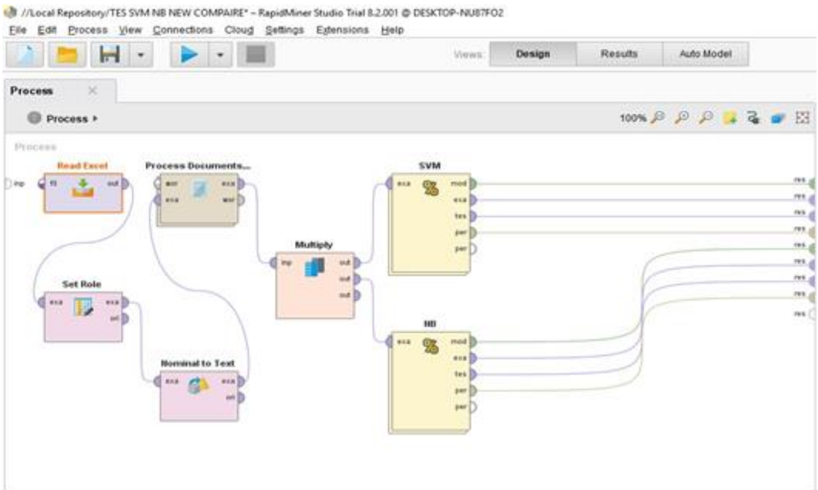


**Figure 3.** Model Design Comparison of Support Vector Machine and Naïve Bayes Algorithms

Figure 4 below is a design of the RapidMiner model for the Naive Bayes algorithm based on Particle Swarm Optimization (PSO) and Support Vector Machine based on Particle Swarm Optimization (PSO).

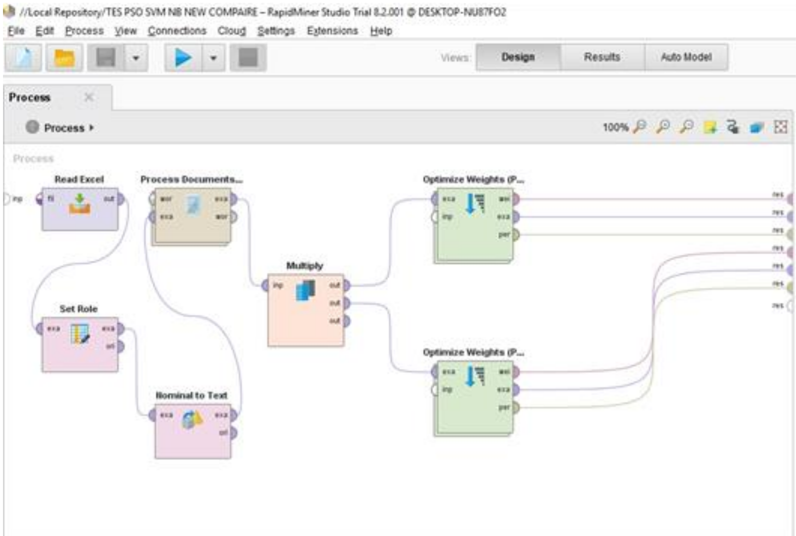


**Figure 4.** Model Design Comparison of Support Vector Machine based PSO Algorithms and Naïve Bayes based PSO

*3.5 Evaluation*

The evaluation phase aims to determine the usefulness of the model that has been successfully made in the previous step. For evaluation, 10-fold cross-validation is used. Following is the design of the Support Vector Machine algorithm model testing process used.

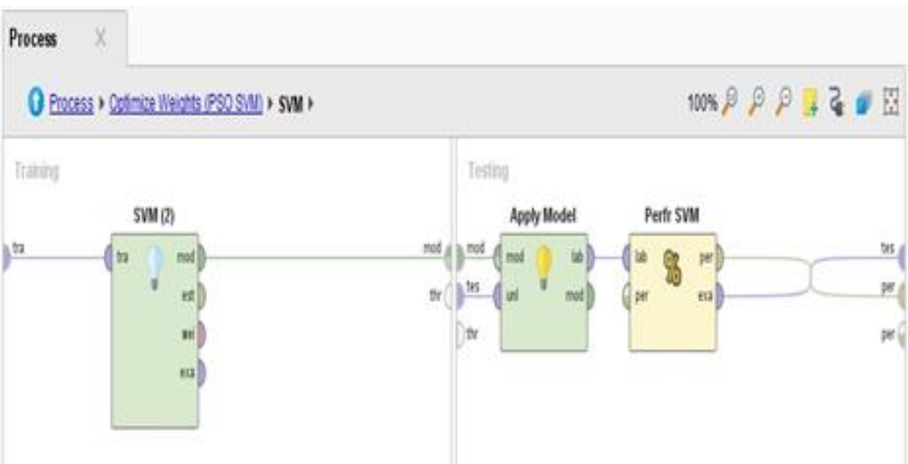


**Figure 5.** Design of a 10-Fold Cross Validation Process for Support Vector Machine

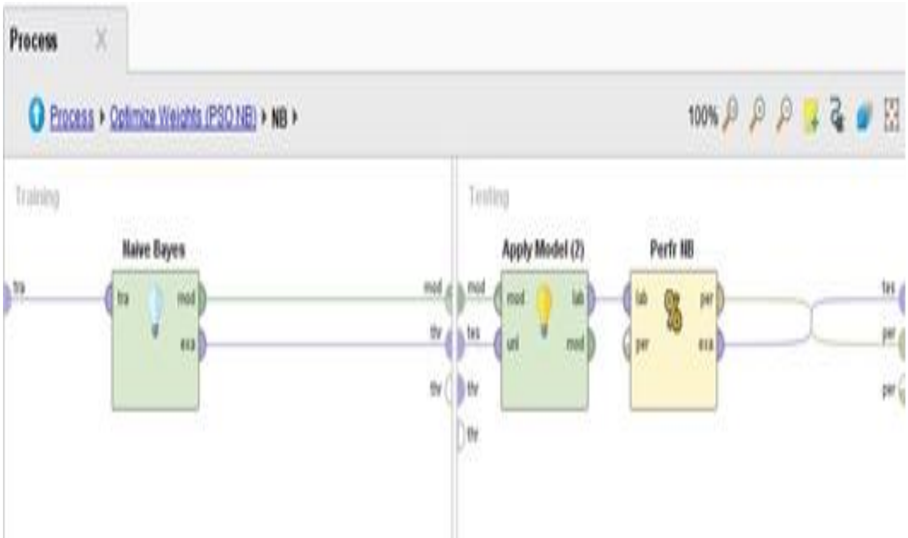


**Figure 6.** Design of a 10-Fold Cross Validation Process for Naïve Bayes

Figure 5 and figure 6 describes the design process in the operator's Support Vector Machine and Naïve Bayes cross-validation in Figure 4. In this test, the data used is clean data that has passed the preprocessing text stage. Data is taken from the Read Excel operator, this is done because the dataset is stored in Excel (.xls). Process documents from files to convert files to documents. The validation process consists of training data and test data. Subprocess training is used to train the model. The trained model is then applied to subprocess testing. Furthermore, after the data has been tested, a cross-process has been carried out where the test data is then used as training data or vice versa, the previous training data used is now the testing data. This process is repeated 10 times for each part so that the parts of the ten parts have been testing data for the Naive Bayes algorithm and Support Vector Machine. At this stage also uses the Role Set which functions to determine the fields in the class then use Particle Swarm Optimization (PSO) so that the resulting accuracy is higher than the results of the previous modeling. Classification Category has value: 1). 0.90 - 1.00 = excellent classification. 2). 0.80 - 0.90 = good classification. 3). 0.70 - 0.80 = fair classification.4). 0.60 - 0.70 = poor classification. 5). 0.50 - 0.60 = failure. As for the comparison of the accuracy and Area Under Curve (AUC) results of the two algorithms that have been used can be seen in table 6.

**Table 6**. Comparison of Accuracy and AUC

| Algorithms | Accuracy | AUC |
|---|---|---|
| Support Vector Machine | 76.09% | 0.848 |
| Support Vector Machine + PSO | 78.40% | 0.850 |
| Naïve Bayes | 68.21% | 0.672 |
| Naïve Bayes + PSO | 74.98% | 0.708 |

*3.6 Deployment*

At the deployment stage, application design will be carried out using data sets on data sets about the electability of political figures. This stage is done by creating a tool built with the VB.net programming language and MySQL database. Figure 7 shows the main menu page for the analysis of sentiments from the selection of political figures.

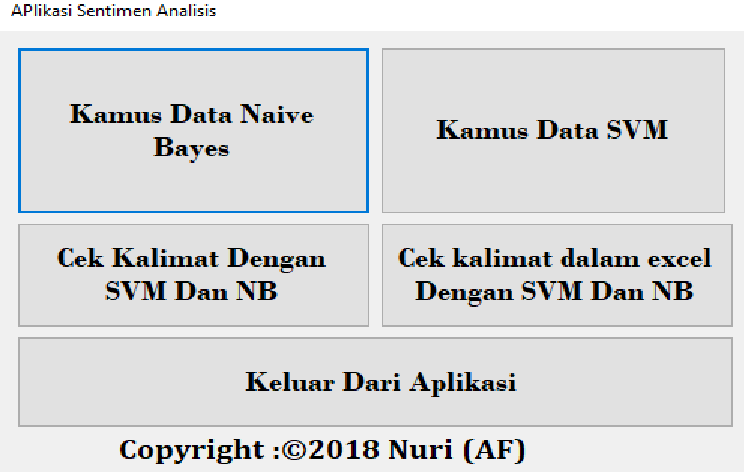


**Figure 7.** Display of the Main Menu Application

Figure 8 shows a data dictionary page from the calculation of the Naive Bayes model and Support Vector Machine,

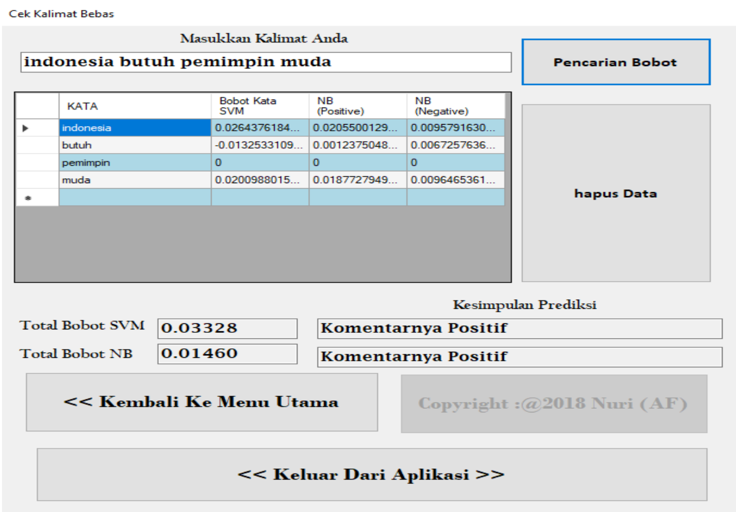


**Figure 8.** Display Sentence Check Menu

Figure 9 display of the Data Dictionary Menu

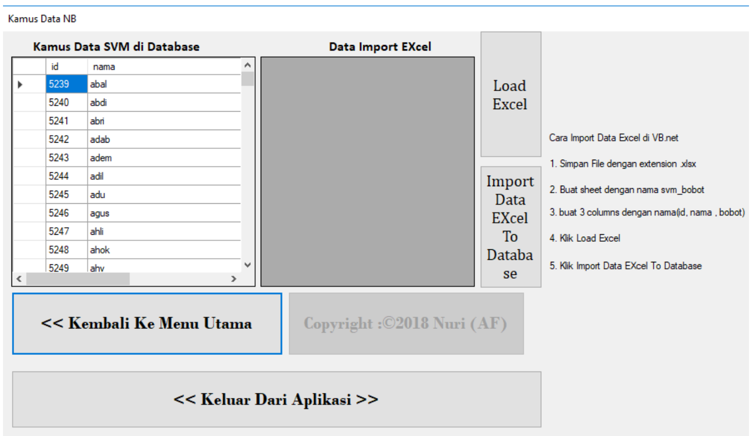

**Figure 9.** Display of the Data Dictionary Menu

Figure 10 is a page view of the sentence check menu.

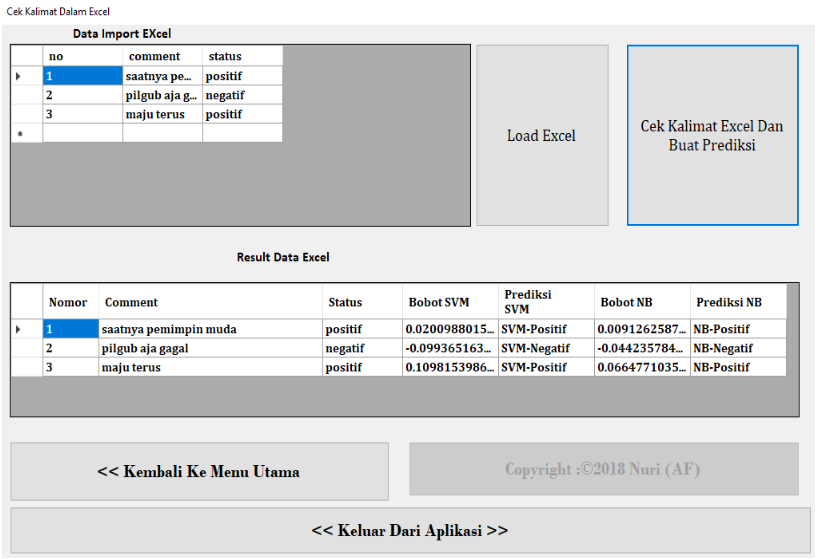


**Figure 10.** Sentences Check Menu Display in Excel

This page is a page to check every sentence that is inputted freely and to know the sentence weight and status prediction of the inputted sentence. Figure 18 is a page view of the sentence check menu in an Excel file. This page is a page to check every sentence contained in an Excel file. Difference between sentence check menus in Excel with sentence check menu is if you use sentence check menu in Excel, the sentence that will be checked can be in large numbers, but if you use sentence check menu, only one sentence can be checked.

## 4. Conclusion

From the results of the research that has been conducted related to the classification of electability sentiment analysis of political figures on online news media comments, it can be concluded that the use of the Gata Textmining framework is very helpful in the commentary pre-processing of Indonesian. The use of Support Vector Machine Algorithm based on Particle Swarm Optimization (PSO) has a better level of accuracy than Naive Bayes based on Particle Swarm Optimization (PSO). While Particle Swarm Optimization (PSO) plays an important role in increasing the accuracy of the Support Vector Machine algorithm. So that the Support Vector Machine algorithm based on Particle Swarm Optimization (PSO) can be a solution for classification in the sentiment of analysis of political figures. Based on the results above, this study also carried out a deployment phase based on the accuracy of the algorithm Support Vector Machine based on Particle Swarm Optimization (PSO) using VB.net and MySQL.

## Acknowledgments

We would like to say thank you for our lecturer, Mr. Dr. Windu Gata, M. Kom has guided in writing this research. We also thanks to our parents who have given support to us and we also very appreciate for knowledge that we had during this research.

IOP Conf. Series: Materials Science and Engineering **662** (2019) 042003 doi:10.1088/1757-899X/662/4/042003

## References


[1] Aryo. Prakoso 2018 Selektifitas Media Sosial Dalam Meningkatkan Elektabilitas Figur. *Ettisal Journal Of Communication*. **3**(2), pp. 1-10

[2] Taibbi. Matt 2019 *In 2019, Let's Finally Retire 'Electability' A favored totem of the campaign press has long been useless, and now is actively misleading*. Retrieved from https://www.rollingstone.com/politics/politics-features/what-is-electability-774196/

[3] Jannti, R., Mahendra, R., Wardhana., W, C. Adriani, M 2018 Stance Classification towards Political Figures on Blog Writing. *2018 International Conference on Asian Language Processing (IALP)*

[4] oxforddictionaries.com. 2018. *Sentiment Analysis*. Retrieved from https://en.oxforddictionaries.com/definition/sentiment_analysis

[5] Liu, B 2015 *Sentiment Analysis Opinions, Sentiment, and Emotion in Text*. (Chicago, USA: Cambridge University Press)

[6] Saputra, N., Adji, T. B., & Permanasari, A. E 2015 Analisis Sentimen Data Presiden Jokowi dengan Preprocessing Normalisasi dan Stemming menggunakan Metode Naive Bayes dan SVM. *Jurnal Dinamika Informatika*. **5**(1), pp. 1-10

[7] Li, Quanzhi. Nourbakhsh, A. Fang, Rui. Liu, Xiaomo. Shah, Sameena 2016 Sentiment Analysis of Political Figures across News and Social Media, *2016 International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation*, UCDC Center, Washington DC, USA

[8] Budiharto, W., & Meiliana, M. 2018 Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. *Journal of Big Data*. **5**(1), pp. 5 -12.

[9] Aliandu, Paulina 2013 Twitter Used by Indonesian President: An Sentiment Analysis of Timeline. *Information Systems International Conference (ISICO)*.

[10] Apriyanti, T., Wulandari, H., Safitri, M., & Dewi, N. 2016 Translating Theory of English into Indonesian and Vice-Versa. *Indonesian Journal of English Language Studies*. **2**(1), pp.38-59.

[11] R. Feldman and J. Sanger *The Text Mining Handbook: Advances Approaches in Analyzing Unstructured Data*, New York, NY: Cambridge University Press, 2007.

[12] Goyal, Shubham 2016 Sentimental Analysis of Twitter Data using Text Mining and Hybrid Classification Approach. *International Journal of Advance Research, Ideas and Innovations in Technology*. **2**(5), pp. 1-15.

[13] Vadivukarassi. M, Puviarasan. N, Aruna. P 2017 Sentimental Analysis of Tweets Using Naive Bayes Algorithm. *World Applied Sciences Journal*. **35**(1), pp.54-59

[14] Al-Amrani. Y, Lazaar. M, El Kadiri. E. K 2018 Sentiment Analysis Using Hybrid Method Of Support Vector Machine And Decision Tree. Journal of Theoretical and Applied Information Technology. **96**(7)

[15] Wahyudi. M, Kristiyanti. D. A 2016 Sentiment Analysis Of Smartphone Product Review Using Support Vector Machine Algorithm-Based Particle Swarm Optimization. Journal of Theoretical and Applied Information Technology. **91**(1), pp. 5 -19

[16] Buntoro, G. A. 2016 Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *Integer Journal Maret*. **1**(1), pp.32-41

[17] Budiono, D. F., Nugroho, A. S., & Doewes, A. 2017 Twitter Sentiment Analysis of DKI Jakarta ' s Gubernatorial Election 2017 with Predictive and Descriptive Approaches. *2017 International Conference on Computer, Control, Informatics and Its Applications*, pp. 89–94

[18] Fatyanosa, T. N., & Bachtiar, F. A. 2017 Classification method comparison on Indonesian social media sentiment analysis. *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, pp.310–315

[19] Rana, S., & Singh, A. 2016 Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques. *International Conference on Next Generation Computing Technologies*, pp.106–111

[20] Kenneth Jensen 2016 *IBM SPSS Modeler CRISP-DM Guide* Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf