Sentiment analysis of facebook comments on Indonesian presidential candidates using the naive bayes method

Syahriani1*, A A Yana2, T Santoso3
1,2,3Computer Science, STMIK Nusa Mandiri, Jakarta, 10420, Indonesia
E-mail: syahriani.yii@nusamandiri.ac.id

Abstract. Sentiment analysis is a very interesting research object to study. Sentiment analysis itself is one branch of textmining whose research focuses on the opinion of a text document. In this study, the author examines the sentiment analysis of facebook commentary of 2 Indonesian presidential real candidates in the 2014. Henceforth in this study, the two candidates are referred to as Presidential Candidate 1 and Presidential Candidate 2, where the numbering sequence is adjusted to the numbering of real election data of president in Indonesia. Here the author chooses to use Facebook comment data on several statuses posted on the 2 official accounts of the Indonesian presidential candidates, because Facebook is a social media that is widely used by Indonesians, it is evident that Facebook’s social media ranks 3rd in Indonesia. In the process of classifying this text using the Naive Bayes method because this method is very simple, has good performance in many domains and this method is very simple. But the Naive Bayes method itself, has the disadvantage of being very sensitive to too many features, which can lead to low classification accuracy. To overcome the problems that exist in the Naive Bayes method, this study uses a combination of feature selection methods, namely information gain and genetic algorithm, the two additional methods serve to improve the accuracy in Naive Bayes classifier. In addition, in this study the author uses smile-forming character conversion, pre-processing of documents such as transform case, tokenization, Indonesian stopwords, Stemming, Token weighting, then classification and confusion matrix testing. This research produces a positive or negative text classification from Facebook comments. Then the measurement is based on the accuracy of Naive Bayes before and after the addition of the feature selection method. The evaluation process uses 10 fold cross validation. From the results of the implementation and testing, the Naive Bayes method with feature selection has an accuracy level of sentiment classification of 83.67% from the previous results of 60.00% here the researcher also displays the ROC curve.

1. Introduction
According O’Leary and Daniel in Akundi et.al, The term social media implies to the use web based and mobile technologies to communicate and interact based on user generated content and dialogue. Enabled by several accessible platforms and characteristics, social media significantly changed the way individuals, organizations and communities communicate[1]. Facebook is one of the social media that is widely used by Indonesians, it is proven that Indonesia in 2019 has been ranked as the third country and user as well as the largest Facebook ad target in the world with 130 million active users every month the fact is obtained from http:/tekno.kompas.com. Because of this many politicians use social media as a platform to carry out
campaign actions for their parties. One of the facts is the political campaign that was conducted during the 2014 Indonesian presidential election. Facebook is used as a tool to neutralize the activities carried out by these politicians, one of which is as a means of promotion of presidential candidates, so Facebook users can like and follow updated news from the status of the politician he followed. Because of this, the researchers felt very interested in conducting research on sentiment analysis of the comments written by followers of the two Indonesian presidential candidates. There are several similar studies on sentiment analysis, namely about sentiment analysis in film review comments that exist in the Digg social network using the classification of Naïve Bayes, Decision Tree, Maximum-Entropy and K-Means Clustering [2]. In addition there is a film review sentiment analysis and some products from Amazon.com using the classification of Support Vector Machine and Artificial Neural Network[3]. Based on Troussas, et al. "Naïve Bayes algorithm is an algorithm that has quite good results in classifying sentiments compared to the perceptron algorithm[4]. Although Naïve Bayes is considered to be quite good in the classification process, it has the disadvantage of being very sensitive in feature selection[5]. The filter method consists of document frequency, mutual information, information gain, and chi-square. None of the four methods is widely accepted as the best feature selection method for sentiment classification or text categorization, however, information gain is often superior to the others[3]. According to Gunal [6] one of the wrapper methods that can be used in feature selection is Genetic algorithm (GA). Judging from previous research, the researchers in this study will try to conduct experiments using the Naïve Bayes method with information gain for the filtering process in text classification and genetic algorithms for feature selection, in order to improve the results of sentiment analysis accuracy.

2. Literature Review

2.1. Textmining

Text mining is a process to extract interesting and significant patterns to explore knowledge from textual data sources [7]. Text mining is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics[7]. According Sukaya and Biruntha in Salloum et.al, text mining is like an intelligence system which is extracting proper words or sentences from the improper words and then transforming those words into the particular suggestions. Text mining is basically a new field having the main purpose of data recovery, machine learning, information mining and computational linguistics[8].

2.2. Sentiment Analysis

Sentiment analysis (or opinion mining) is defined as the task of finding the opinions of authors about specific entities[9]. According to Pan in Basari, opinion mining or sentiment analysis comes into play when such lots of data makes it difficult to evaluate them personally[10]. According Liu in Basari, Opinion mining learns individuals’ views, tests, behavior, and feelings toward people, individuals, issues, activities, subjects and their features. Opinion are considerable because they are primary influences of our behaviors[10].

2.3. Feature Selection

Feature selection is used to eliminate irrelevant and redundant features, possibly causing confusion, by using specific methods (e.g., brute-force approach, embedded approach, filter approach, wrapper approach, embedded methods-see, for instances[11]).
2.3.1. Feature Selection Method Used:

- **Genetic Algorithm**
  According to Han and Kamber, Genetic algorithms try to combine ideas of natural evolution. In general, genetic learning starts as follows[12]:
  
  (i) An initial population is made up of random rules. Each rule can be represented by a string of bits. As a simple example, suppose that the sample in a given training set is explained by two Boolean attributes, A1 and A2, and that there are two classes, C1 and C2. The "If A1 And Not A2 Then C2" rule can be encoded as string string "100," where the two leftmost bits represent attributes A1 and A2, respectively, and the rightmost bit represents the class. Likewise, the rule "If Not A1 And Not A2 Then C1" can be coded as "001." If the attribute has k values, where k > 2, then k bits can be used to encode the attribute values. Classes can be coded in the same way.
  
  (ii) Based on the idea of resilience from the most suitable, the newly formed population consists of the most appropriate rules in the current population, as well as the descendants of these rules. Typically, fitness rules are assessed with classification accuracy on a set of training samples.
  
  (iii) Heredity is created by applying genetic operators such as crossovers and mutations. In crossovers, substrings of a pair of rules are exchanged to form a new pair of rules. In mutations, bits chosen randomly in string rules are reversed.
  
  (iv) The process of generating a new population based on previous population rules continues until the population, P, develops where each rule in P meets a predetermined fitness threshold.

Genetic algorithms are easily aligned and have been used for classification like other optimization problems. In data mining, genetic algorithms can be used to evaluate other fitness algorithms.

- **Information Gain**
  Information gain is calculated from the output data or dependent variable y which is grouped based on attribute A, denoted by gain (y, A). Information gain, gain (y, A), of attribute A relative to data output y is:

  \[
  Gain(y, A) = \text{entropi}(y) - \sum_{Values(A)} \frac{y_c}{y} \text{entropi}(y_c)
  \]

  where the value (A) is all possible values of attribute A, and yc is a subset of y where A has the value c. The first term in the above equation is total entropy y and the second term is entropy after data separation is done based on attribute A. Information gain is calculated by finding the difference before and after separation. The stages in the process of calculating Information gain are as follows:

  (i) Find the entropy value before separation with the following formula:

  \[
  \text{Entropy (y)} = -\sum \Pi_i \log_2 \Pi_i
  \]

  \(\Pi_i: \text{proportion of data } y \text{ with class } i\)

  (ii) Find the entropy value after separation based on attribute A with the following formula:

  \[
  \text{Totalentropi} = \sum_{Values(A)} \frac{y_c}{y} \text{entropi}(y_c)
  \]

  (iii) Find the information gain value using the following formula:

  \[
  Gain(y, A) = \text{entropi}(y) - \sum_{Values(A)} \frac{y_c}{y} \text{entropi}(y_c)
  \]
2.4. Naïve Bayes Algorithm

The Naive Bayes algorithm is a classification algorithm based on probabilities in statistics put forward by Thomas Bayes that predicts future opportunities based on past opportunities (Bayes theorem). This method is then combined with "naive" where the conditions between attributes are independent of each other. In the dataset, each dataset has attributes and 1 (one) class label, then the probability of data entering a class label can be defined in the following steps [12]:

- Known D is the training data and class label. Each data is represented in the form n attribute vector dimensions X = (x1, x2,... , xn)
- Suppose there are m number of classes, C1, C2, ..., Cm. The Naive Bayes method will predict whether X belongs to the class that has the highest probability posterior value. Naive Bayes will predict X will enter the Ci class if and only if: P (Ci | X) > P (Cj | X) for 1 ≤ j ≤ m, j <> i. Then P will be maximized P (Ci | X). The most class of Ci is called the maximum posteriori hypothesis which is calculated using Equation (3).

\[
P(C_i | X) = \frac{P(C_i | X)P(C_i)}{P(X)}
\] (4)

- Because P (X) is fixed for all classes, only P (X | Ci) P (Ci) must be maximized. If the prior probabilities of the class are unknown, it is generally assumed that all classes are equal P (C1) = P (C2) = ... = P (Cm). Keep in mind that the prior probabilities of classes are estimated with P (Ci) = | Ci, D | / | D |, where | Ci, D | is the amount of training data included in the Ci class in dataset D.
- For datasets that have many attributes, the computational complexity will be very high, so it needs to be reduced by assuming all class conditions are mutually independent (independence). This assumes that the values between attributes do not affect each other, so they can be defined: Where the probabilities of P (x1 | Ci), P (x2 | Ci), ..., P (xn | Ci) can be obtained easily from data training. Where xk is the value in the attribute A for dataX. For each attribute, we must see whether the attribute value is categorical or a continuous value.
- If Ak is categorical, then P (xk | Ci) is the amount of xk data that has classCi in D training data divided by | Ci, D |, the sum of all Ci class data recorded in training D.
- If Ak is continuous, such as age data, other numerical data, which cannot be categorized, then the data must be made in a range of values, using Gaussian Distribution with Equations (4) and (5).

\[
g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\] (5)

So we get

\[
P(x_k | C_i) = g(x_k, \mu C_i, \sigma C_i)
\] (6)

- To predict class labels for data X, P (X | Ci) P (Ci), then predictions are made for each classCi. The Naive Bayes method will predict the class for X is Ci if and only if (Equation 6). P (X | Ci) P (Ci) > P (Cj) for 1 ≤ j ≤ m, j <> i ............. (6) In the sense the prediction of the class with the greatest probability is calculated by Equation (7).

\[
P(X | C_i) = \prod_{k=1}^{n} P(x_k | C_i) = P(x_1 | C_i)XP(x_2 | C_i)X...XP(x_n | C_i)
\] (7)
3. Result And Discussion

3.1. Data Profiles

This research uses data from Rachmat. Research C and Lukito in 2015 in .csv format. The data is status and commentary data from the 2014 Presidential Election which bet 68 status (3400 comments) during the campaign period. Only researchers took 300 comments (training data) in which: 15 statuses for Presidential Candidates 1 and 15 statuses for Presidential Candidates 2, then researchers took 10 comments randomly and made 150 comments (testing data). Because in the dataset there are 3 categories of results, namely: positive, neutral and negative, all comments included in the neutral results category are made for the negative category. The neutral category is made into a negative category that is approved by Liu who states that a neutral class can be approved by a negative class [11].

3.2. System Implementation

3.2.1. Subprocess Stage

Handling smile conversion is performed on a variety of common smile-forming characters, namely: ‘:)’, ‘(y)’, ‘_’, ‘(Y)’, ‘:*’, ‘like’, ‘Like’, ‘=/)’, ‘:D’, ‘YES’, ‘yes’ and ‘Yes’, where the smile character is converted into words according to the type of character’s smile, please refer to table 1.

| No | Smile character | Conversion |
|----|----------------|------------|
| 1  | :D             | Smile      |
| 2  | :3             | Smile      |
| 3  | =))            | Smile      |
| 4  | :)             | Smile      |
| 5  | (y)            | Like       |
| 6  | Like           | Like       |
| 7  | LIKE           | Like       |
| 8  | Like           | Like       |
| 9  | Yes            | Like       |
| 10 | YES            | Like       |
| 11 | Yes            | Like       |
| 12 | (Y)            | Like       |
| 13 | =/             | Neutral    |
| 14 | :(             | Sad        |
| 15 | :(             | Sad        |
| 16 | 8)             | Eyeglasses |
| 17 | 8)             | Eyeglasses |
| 18 | *=             | Kiss       |
| 19 | =:P            | Sticking tongue out |
| 20 | =:-P           | Sticking tongue out |
| 21 | =:p            | Sticking tongue out |
| 22 | =:p            | Sticking tongue out |
| 23 | =/             | Disappointed |
| 24 | =/             | Disappointed |
Look at the table below is the result of the conversion of smile characters into words.

| Smile Characters | Words          |
|------------------|---------------|
| O.o              | Shocked       |
| <3               | Love          |
| -.               | Quiet         |

**Table 2.** Comparison of Text Before and After the Conversion Process

| The text before the smile character conversion process- yang dukung X like (y) is carried out |
|---------------------------------------------------------------|
| The text before the smile character conversion process- yang dukung X Suka Suka is carried out |

**Information:**

X = Presidential Candidate 1
Y = Presidential Candidate 2

- Preprocessing Stage:
  - Transform Case
    - The stages used to convert text data into all lowercase letters.

**Table 3.** Comparison of Text Before and After Doing the Transform Case Process

| Text before the transform case process | Oke pak boss!! |
|----------------------------------------|----------------|
| Text before the transform case process | oke pak boss!! |

**Tokenization**

All words contained in a comment will be removed from the punctuation mark, as well as symbols or anything that is not a letter if contained in the comment. Following is an example of the results of the tokenization process in rapidminer.

**Table 4.** Comparison of Text Before and After the Tokenization Process

| Text before tokenization process | Oke pak boss!! |
|----------------------------------|----------------|
| Text before tokenization process | oke pak boss!! |
Stopwords Removal
This process will eliminate irrelevant words, such as the conjunctions "and", "or", "but" and others. There are no definite rules to determine the stop word to be used, we can adjust the stopword according to the case being resolved. Here is an example of the results of the stopwords removal process in rapidminer:

Table 5. Comparison of Text Before and After Doing the Stopwords Removal Process

| Text before the stopword removal process is carried out | Text before the stopword removal process is carried out |
|--------------------------------------------------------|--------------------------------------------------------|
| oke pak boss                                           | boss                                                   |

Stemming
Stemming is the process of changing affixed words into basic words. Here is an example of the results of the process of stemming in rapidminer:

Table 6. Comparison of Text Before and After Stemming Process

| Text before stemming process | boss |
|-----------------------------|------|
| boss                        |      |

Classification Stage
This stage is to determine a sentence included in the positive class or negative class based on the probability value of the Bayes formula. Researchers only display 10 comments from a total of 300 comments data in the training data.

Table 7. Boolean Comment Vector Tables with Classification Results Class Labels

| Comment                                                                 | Class     |
|-------------------------------------------------------------------------|-----------|
| Y insya Allah jadi presiden 2014 :), yang like berarti setuju. hehe    | Positive  |
| Hanya Untuk Indonesia Raya...!!! (y)                                    | Positive  |
| Y insya Allah jadi presiden 2014 :), yang like berarti setuju. Hehe    | Positive  |
| You are the best. for INA,.Pak X !                                    | Positive  |
| Mari kita selalu berdoa utk kelancaran no.2 menjadi RI1-2 demi perubahan positiv untuk Indonesia | Positive  |
| Hidup Pak Y ..                                                          | Negative  |
| Menta Duit 200RB Pak Y                                                 | Negative  |
| yang dukung Y mana like nya :D                                         | Negative  |
| Hanya yg IQ jongkok !! Yg menginginkan ada memimpin negri ini           | Negative  |
| Sepakat sama pernyataan pak X. Untuk presiden tetap Y.                 | Negative  |
Information:
X = Presidential Candidate 1
Y = Presidential Candidate 2

Below is a design model created using RapidMiner 7.5, see Figure 1.

Figure 1. Naive Bayes Model

3.3. Model Optimization with Combined Feature Selection Methods
This research combines the method of selecting filter and wrapper features. The filter feature uses the information gain method and genetic algorithm of the wrapper. The data processed will be weighted by information gain to improve the accuracy of the Naive Bayes classification. This research uses the operator select by weight by selecting the parameter weight relation = top k, where k = 10. Then produce the top 10 attributes. The 10 selected attributes will display their weights. See the table below:

| Attribute | Weight |
|-----------|--------|
| beliau    | 1      |
| bencana   | 1      |
| bocoran   | 1      |
| departemen| 1      |
| indonesia | 1      |
| maju      | 0      |
| mobil     | 1      |
| Nama      | 1      |
| pencitraan| 1      |
| pilih     | 1      |

For Genetic algorithms, the adjusted indicators are population size = 50, p initialize = 0.6, p crossover = 0.7, and p generate = 1.0. Whereas the one tested to improve accuracy is the p value of the crossover. The indicator table and test results can be seen in the table below.
Table 9. Indicators and Testing Results

| P. Crossover | Accuracy  |
|--------------|-----------|
| 0.6          | 83.33%    |
| 0.7          | 83.67%    |
| 0.8          | 83.67%    |
| 0.9          | 83.67%    |
| 1.0          | 83.33%    |

If there are other indicators that have also been changed, it can cause data processing to become longer. Below is the Naïve Bayes design model with the method of selecting Information Gain and Genetic Algorithm features.

Figure 2. Naïve Bayes Model with Feature Selection

3.4. Confusion Matrix

Below will show a comparison of the results of the Naïve Bayes accuracy before adding the feature selection method in table 10 and ROC in figure 3 and after adding the feature selection method in table 11 and ROC in figure 4.

Table 10. Naïve Bayes Confusion matrix model before adding a feature selection method

| Naïve Bayes Accuracy:60.00%+/−6.67%(mikro:60.00%) |
|--------------------------------------------------|
| true positive | true negative | class precision |
| pred. positive | 132          | 15               | 89.80%          |
| pred. negative | 105          | 48               | 31.37%          |
| class recall   | 55.70%       | 76.19%           |
Figure 3. ROC Naïve Bayes

Table 11. Confusion matrix model Naïve Bayes after addition feature selection method

|              | true positive | true negative | class precision |
|--------------|---------------|---------------|-----------------|
| pred. positive | 235           | 47            | 83.33%          |
| pred. negative | 2             | 16            | 88.89%          |
| class recall  | 99.16%        | 25.40%        |                 |

Figure 4. ROC Naïve Bayes with Feature Selection

4. Conclusions and recommendations

4.1. Conclusion
Judging from the data that has been processed, the merging of the method of selecting features such as filter and wrapper can improve the classification accuracy of the naïve bayes method. Facebook comment data can be classified properly into positive and negative classes. The accuracy of the data before using the merger of the naïve bayes method and feature selection was 60.00%, whereas after the merger with the method of selecting the accuracy of the features to 83.67% it could be seen that there was an increase in accuracy by 23.67%. Then in the process of implementing a classification system that uses Indonesian social media data, there is a need for preparation especially in emoticon conversion.
4.2. Recommendations
The use of naive Bayes method with information gain and Genetic algorithm is seen to be able to improve good performance in text classification, hopefully for further research can combine other methods such as chi square, Gini index and mutual information. It is also expected to be able to conduct research in the field of text mining with different data, and hopefully in subsequent studies that use Indonesian social media data, it is necessary to add the written abbreviation conversion preprossing.

5. References
[1] Akundi A Tseng B Wu J Smith E Subbalakshmi M and Aguirre F, 2018 Text mining to understand the influence of social media applications on smartphone supply chain Procedia Comput. Sci. 140 p. 87–94.
[2] Yessenov K, 2009 Sentiment Analysis of Movie Review Comments 6.863 p. 1–17.
[3] Moraes R Valiati J F and Gavião Neto W P, 2013 Document-level sentiment classification: An empirical comparison between SVM and ANN Expert Syst. Appl. 40, 2 p. 621–633.
[4] Troussas C Virvou M Espinosa K J Llaguno K and Caro J, 2013 Sentiment analysis of Facebook statuses using Naive Bayes Classifier for language learning IISA 2013 - 4th Int. Conf. Information, Intell. Syst. Appl. July p. 198–205.
[5] Chen J Huang H Tian S and Qu Y, 2009 Feature selection for text classification with Naïve Bayes Expert Syst. Appl. 36, 3 PART 1 p. 5432–5435.
[6] Günlü S, 2012 Hybrid feature selection for text classification Turkish J. Electr. Eng. Comput. Sci. 20, SUPPL.2 p. 1296–1311.
[7] Fan, W., Wallace, L., Rich S, 2017 Today’s communications of the ACM Commun. ACM 60, 7 p. 5.
[8] Salloum S A Al-Emran M Monem A A and Shaalan K, 2017 A survey of text mining in social media: Facebook and Twitter perspectives Adv. Sci. Technol. Eng. Syst. 2, 1 p. 127–133.
[9] Feldman R, 2013 Techniques and applications for sentiment analysis Commun. ACM 56, 4 p. 82–89.
[10] Basari A S H Hussin B Ananta I G P and Zeniarja J, 2013 Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization Procedia Eng. 53 p. 453–462.
[11] Gorunescu F, 2015 Intelligent System Reference Library 3, 2.
[12] Han, J., Kamber, M., Pei J, 2014 Data mining: Data mining concepts and techniques.