SIMPLE TEXT MINING FOR SENTIMENT ANALYSIS OF POLITICAL FIGURE USING NAÏVE BAYES CLASSIFIER METHOD

Yustinus Eko Soelistio *, Martinus Raditia Sigit Surendra †
System Information, Faculty of Information and Communication Technology, Multimedia Nusantara University
Jl.Scientia Boulevard, Gading Serpong, Tangerang, Banten-15811, Indonesia
email : * yustinus.eko@umn.ac.id, † sigit@umn.ac.id

ABSTRACT

Text mining can be applied to many fields. One of the application is using text mining in digital newspaper to do politic sentiment analysis. In this paper sentiment analysis is applied to get information from digital news articles about its positive or negative sentiment regarding particular politician.

This paper suggests a simple model to analyze digital newspaper sentiment polarity using naïve Bayes classifier method. The model uses a set of initial data to begin with which will be updated when new information appears. The model showed promising result when tested and can be implemented to some other sentiment analysis problems.

Keywords : text mining, naïve Bayesian, sentiment analysis

1. INTRODUCTION

Indonesia is one of the big democratic nation. Almost everyday there are news about politician that cover many topics include corruption and regional election. Mass media has important role in delivering news therefore can influence public opinion. For example, one news media give positive review on one candidate while others give negative one.

Nowadays news media can deliver their content through digital media. This accessibility opens new opportunity to analyze news content with text mining. Digital news media can be considered as unstructured data. This huge amount of data available on the web creates today an information overloading problem [5].

Text mining has been implemented in many applications such as [1,2,3,4,6,8]. One of the suggested implementation is for analyzing readers’ sentiment on some particular news. Research result by [1] suggest that naïve Bayesian classifier and support vector machines can be used to identify readers opinion, either positive or negative, on English movie’s review and Indonesian daily news.
\[\begin{align*}
-\alpha &\rightarrow (a = -1) \land (p_{(w, h, a)} < 0) \\
+\alpha &\rightarrow (a = +1) \land (p_{(w, h, a)} > 0) \\
\end{align*}\]

Values of \(p_{(w, h, a)}\) are stored in \(w \times h\) matrices of matrix \(M\) and value of \(s_{(w, h)}\) are stored in \(w \times h\) matrices of matrix \(N\). \(w\) is the “who” where the statement come from in an article \(C_i\), and \(h\) is the “to whom” or “to who” \(w\) give his/her statement to. Since there are many combinations of structures in a sentence then seven assumptions will be set and used through out this paper.

**Assumption 1**
There are only two types of articles in the news, first is articles which discuss about one or more politician, and second articles that do not say anything about politician (even though the article is about politic). This assumption will hold true since all articles that discuss one or more politician will have \(\{(w, h, a) \in C_i \mid s_{(w, h)} > 0\}\) and the others \(\{(w, h, a) \in C_i \mid s_{(w, h)} = 0\}\).

**Assumption 2**
For each statement \(\alpha\) there are always person \(w\) who declares, and person \(h\) whose the \(\alpha\) are declared to. Thus whether exist \(\alpha\) and \(w\) then there is always \(h\), and whether exist \(\alpha\) and \(h\) then there is always \(w\). This assumption makes sure that change in \(p_{(w, h)}\) will always by the value of \(a\), and \(\alpha\) always have references to \(p_{(w, h)}\).

**Assumption 3**
Though assumption 2 will hold for most statements \(\alpha\) in \(C_i\), there are some possibilities that it will not. There are some cases where there is no reference of \(w\) but \(a\) is present, such may happen in the first sentence of \(C_i\). Assumption 3 will guarantee that assumption 2 will always be true by assigning \(w\) to the news media where the article appears, hence the default value of \(w = news\ media\). 

**Assumption 4**
Article \(C_i\) can have two or more \(w\) and \(h\) therefore the system keep track of \(w_y\) and \(h_y\) by changing their values to the most recent politician keyword found. For example let say \(b_z\) is words in \(C_i\) then \(b_z \in C_i\) and 

\[\begin{align*}
(b_z = \text{who keyword}) &\land (b_z \neq w_y) \rightarrow w_{y+1} = b_z \\
(b_z = \text{whom keyword}) &\land (b_z \neq h_y) \rightarrow h_{y+1} = b_z
\end{align*}\]

This assumption ensure that each \(\alpha\) give the correct \(a\) to \(p_{(w, h)}\).

**Assumption 5**
Every negation keyword \((g)\) in a sentence such as “no” and “not” will change the polarity of \(\alpha\) thus change value of \(a = -1\) for \(p_{(w, h, a)}\) to \(a = +1\). If there exist two or more \(g\) in one sentence then polarity of \(\alpha\) will be changed as many times as \(g\) appear.

**Assumption 6**
To distinguish between +\(\alpha\) and sarcasm, the system check +\(\alpha\) with the value of prior \(p_{(w, h)}\). If \(p_{(w, h)} < 0\) then the +\(\alpha\) will be considered as sarcasm, otherwise it will be considered as legitimate positive statement. Polarity of +\(\alpha\) will not be changed thus \(a = +1\) will be added to \(p_{(w, h)}\).

**Assumption 7**
This paper assumes that all news media have some sentiment tendency towards politician. Therefore each \(\alpha\) in \(C_i\) which correspond to \(h\) change the probable polarity of news media towards \(h\). Thus for every \(p_{(w, h, a)}\) will change \(p_{(w, h)}\) with the same value of \(a\) and increase \(s_{(w, h)}\) by one.

### 3. TRAINING SET AND NAIVE BAYESIAN MODEL

Training set \(T\) consists of \(C_i\) therefore \(T \ni w, h, a\) and \(a = +1 \iff -\alpha\) or \(a = +1 \iff +\alpha\). To find out whether a news media give positive or negative review on particular politician, this paper adapts Bayesian model from [1,4,11] and modify it so

\[P(p_{(w, h)} | s_{(w, h)}) = \frac{p_{(w, h)}}{\sum_{s_{(w, h)}} p_{(w, h)}}\]
negative sentiment towards \( h \). Thus the probability of sentiment polarity of article \( C \) towards \( h \) is

\[
P_{c(h)} = \frac{\sum_{w=1}^{n} p(w|h) P_{c(w)}}{\sum_{w=1}^{n} P_{c(w)}} \tag{6}
\]

Following equation (3) then positive value \( a \) will add \( p(w|h) \) value, and negative \( a \) will subtract \( p(w|h) \) value. Hence equation (1) concludes that the further away value \( p(w|h) \) from 0 then the higher probability of \( w \) has \( o \) sentiment (either positive or negative) towards \( h \).

| \( h_{1} \) | \( w_{11} \) | \( p(1,1) \) | \( p(2,1) \) | \( p(3,1) \) | \( p(4,1) \) | \( p(5,1) \) | \( p(n,1) \) |
| \( h_{2} \) | \( p(1,1) \) | \( p(2,2) \) | \( p(3,2) \) | \( p(4,2) \) | \( p(5,2) \) | \( p(n,2) \) |
| \( h_{3} \) | \( p(1,1) \) | \( p(2,3) \) | \( p(3,3) \) | \( p(4,3) \) | \( p(5,3) \) | \( p(n,3) \) |
| \( h_{4} \) | \( p(1,1) \) | \( p(2,4) \) | \( p(3,4) \) | \( p(4,4) \) | \( p(5,4) \) | \( p(n,4) \) |
| \( h_{5} \) | \( p(1,1) \) | \( p(2,5) \) | \( p(3,5) \) | \( p(4,5) \) | \( p(5,5) \) | \( p(n,5) \) |
| \( h_{6} \) | \( p(1,n) \) | \( p(2,n) \) | \( p(3,n) \) | \( p(4,n) \) | \( p(5,1) \) | \( p(n,n) \) |

Figure 1. Structure of matrix \( M \)

Initial values of all cells in figure 1 are 0. The values will change after the system goes through \( T \). Each cell stores historical data of sentiment \( w \) towards \( h \). Number of \( w \) and \( h \) do not have to be equal. It is possible that \( w_{q} = h_{r} \) in a case of \( w \) give a statement about him/her self.

Matrix \( N \) has the same structure as matrix \( M \) with the difference \( p(w|h) \) cells. In matrix \( N \), values of \( p(w|h) \) are substituted with values of \( s(w|h) \).

4. SYSTEM DESIGN AND TEST MODEL

This paper creates a system design to test the model. The test will use articles \( C \) from Indonesia’s digital newspaper \( K \) to create \( T \). To test the accuracy, the system will try to determine \( P(P_{c(w)}|s(w|h)) \) from another article \( C \notin T \). Set of \( o \) and \( a \) are determined beforehand by human interpreter.

Before the test, a general system flow is established. The general system flow consists of five modules which are reader and parser, cleaner, helper, analyzer, and display.

Reader and parser separate each words and punctuation mark from the text. Then these words will be filtered by cleaner. The filter removes special words and adverbs from the sentence by comparing them with a set of pre define filter words such as ‘seorang’ (‘a’, ‘an’ in English), ‘adalah’ (‘is’ in English), ‘yang’ (‘that’ in English). As example a sentence like ‘ABC adalah seorang koruptor’ (ABC is an corruptor) will become ‘ABC koruptor’.

Sentences that have been cleansed will be scanned by helper to find non-common pronoun or politician names such as alias or pronoun. This module will change those words and replace them with system’s keywords. These keywords will be used by analyzer to locate \( w \), \( h \), and \( o \).

The system reads all sentences separately hence all sentences are evaluated independently. Analyzer process will be triggered by a condition where the system found special word. These special words can be categorized into negative and positive types.

Figure 3 is process flow when the system found negative word like koruptor (‘corruptor’ in English), and tersangka (‘suspect’ in English). The result of process in figure 3 are \( w \), \( h \), and \( p(w|h) \). Using this system result can create matrix \( M \) and \( N \).

The process flow of positive type is similar to the process flow of negative type. The only different is that the keywords \( o \) and the opposite value (+1) to be assigned for positive words.
The algorithm is tested using Java and MySQL database to save the data. The test uses articles from K digital newspaper as training data T and evaluated articles C_i.

From T, the system form matrix M and N with w = kpk, 0 and h = andi, kpk, deddy. w = 0 is the default value which equals to w = K.

Figure 4 shows the result of
\[ P(p_{(kpk, andi)} | s_{(andi, kpk)}) = \frac{-1}{1} \]
\[ P(p_{(andi, kpk)} | s_{(andi, kpk)}) = \frac{-6}{1} \]
and \[ P(p_{(kpk, deddy)} | s_{(kpk, deddy)}) = \frac{-1}{1} \]. By employing formula (6) then \[ P_{C_{(andi, kpk)}} = 1 \], and \[ P_{C_{(deddy, kpk)}} = -1 \]. Thus it means that this particular article from K has probability of 100% giving negative sentiment to Andi and Deddy, and has probability of 100% giving positive sentiment to kpk. These result will be saved as system’s knowledge base T and will be used to analyze C_{i+1}.

Figure 4. Result after running data T

Table 1. Matrix M after running data T

| Article Index | Sentence Index | Who | Whom | Value |
|---------------|---------------|-----|------|-------|
| 1             | 1             | kpk| andi| 1     |
| 1             | 5             | 0  | andi| 1     |
| 1             | 7             | 0  | andi| 1     |
| 1             | 8             | 0  | andi| 1     |
| 1             | 8             | 0  | kpk | 1     |
| 1             | 10            | 0  | andi| 1     |
| 1             | 12            | 0  | andi| 1     |
| 1             | 14            | 0  | andi| 1     |
| 1             | 17            | kpk|      |       |

Table 2. Matrix N after running data T

| Article Index | Sentence Index | Who | Whom | Value |
|---------------|---------------|-----|------|-------|
| h(K)          | 0             | 0   |      |       |
| h_{kpk}       | 1             | 0   |      |       |
| h_{andi}      | -7            | -1  |      |       |
| h_{deddy}     | 0             | -1  |      |       |

Figure 5 shows the result after running C_{i+1}.

\[ P(p_{(kpk, andi)} | s_{(andi, kpk)}) = \frac{-1}{2} \]
\[ P(p_{(andi, kpk)} | s_{(andi, kpk)}) = \frac{1}{2} \]
and \[ P(p_{(kpk, deddy)} | s_{(kpk, deddy)}) = \frac{2}{2} \].

Following formula (6) then \[ P_{C_{(i+1, andi, kpk)}} = \frac{2}{4} \].
Thus it means that article C_{i+1} has probability of 50% giving positive sentiment to Andi.
One major problem in applying this model is to accurately identify the ‘who’ and the ‘whom’ in the articles on precise context. Finding meaning in context behind each sentence and associate it with whole articles’ context is still a challenge for the future.

5. CONCLUSION

The model proposed in this paper can be used to analyze sentiment of an article in digital news media towards politician. The algorithm in the model is fairly general to be used in many other sentiment analysis problems with only few modifications. The prompt test shows promising results, nevertheless further test with bigger data set and more complex slang language yet has to be done.

One major problem in applying this model is to accurately identify the ‘who’ and the ‘whom’ in the articles on precise context. Finding meaning in context behind each sentence and associate it with whole articles’ context is still a challenge for the future.

REFERENCE

[1] N.W.S Saraswati, Text Mining Dengan Metode Naïve Bayes Classifier Dan Support Vector Machines Untuk Sentiment Analysis, Thesis, Program Magister Program Studi Teknik Elektro, Universitas Udayana, Denpasar, 2011.

[2] F. Neri, C. Aliprandi, F. Capei, M. Cuadros, Tomas, Sentiment Analysis on Social Media, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012, 919-926.

[3] A. Adrifina, J.U. Putri, I W. Simri, Pemilahan Artikel Berita Dengan Text Mining, Proceeding, Seminar Ilmiah Nasional Komputer dan Sistem Intelijen, Universitas Gunadarma, Depok, 2008, 176-181.

[4] A. Nurani, B. Susanto, U. Proboekti, Implementasi Naïve Bayes Classifier Pada Program Bantu Penentuan Buku Referensi Mata kuliah, Jurnal Informatika Vol 3 No 2Universitas Kristen Duta Wacana, 2007, 32-36.

[5] S. Iritano, M. Ruffolo, Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining, IEEE, 2001.

[6] E.J. Fortuny, T.D. Smoedt, D. Martens, W. Daelemans, Media Coverage In Times of Political Crisis: a Text Mining Approach, Expert Systems with Applications, Scienedirect, 2012.

[7] Fatudimu I.T, Musa A.G, Ayo C.K, Sofoluwe A. B, Knowledge Discovery in Online Repositories: A TextMining Approach, European Journal of Scientific Research, ISSN 1450-216X Vol.22 No.2 , 2008, 241-250.

[8] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, Y. Yu, Mining Social Emotions from Affective Text, IEEE Transactions On Knowledge And Data Engineering Vol. 24 No. 9, 2012, 1658-1670.

[9] J.T. Malin, D.R. Throop, C. Millward, H.A. Schwarz, F.Gomez, C. Thronesbery, Linguistic Text Mining for Problem Reports, Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, San Antonio, USA, 2009, 1578-1583.

[10] B.T. Kieu, S.B. Pham, Sentiment Analysis for Vietnamese, Second International Conference on Knowledge and Systems Engineering, IEEE Computer Society, 2010, 152-157.

[11] G.F. Lugier, Artificial Intelligence Structures And Strategies For Complex Problem Solving, Pearson Education Inc., 2009, 182-185.