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.

This paper suggests a method to implement sentiment analysis using naïve Bayesian method on digital articles and newspapers. The sentiment analysis focuses on the probability of whether news media give positive or negative review on some particular political figures.

2. BASIC MODEL AND ASSUMPTION

The model starts from what has been suggested in [2] that consider "who" is speaking, "to whom" is speaking, and "what" as variables. This paper adopts those variables and uses them to determined sentiment probability. The values of those variables are updated according to what the system has learned from training data T. Variables "who" (w), "whom" (h) and "what" (a) store their values in a form of matrix M and N as knowledge base set. M is used to stores polarity (p) of w towards h, and Nstores how many times w give such statement toward h $(s_{(w,h)})$. The default values are $\forall(p_{(w,h)}, s_{(w,h)}) = 0$. These values change with T so when

$$\begin{cases} (p_{(w,h)} \wedge s_{(w,h)} > 0) \rightarrow likely\ positive\ sentiment \\ (p_{(w,h)} \vee s_{(w,h)} = 0) \rightarrow likely\ neutral\ sentiment \\ (p_{(w,h)} \wedge s_{(w,h)} < 0) \rightarrow likely\ negative\ sentiment \end{cases} \quad (1)$$

Training data T is a list of independent articles C_i. Each articles can contains one or more political figures "keyword", therefore

$$|(w, h, a) \in C_i| \geq 0 \quad (2)$$

The sentiment of article C_i is cast by $p_{(w,h)}$ and $s_{(w,h)}$ in C_i. Each $p_{(w,h)}$ is determined by the value of a which correspond to unique word o in database D.w, h and o are handled as a token like in [3,7]. o can be "negative" words $(-o)$ such as corruption, convict, and dispute, or "positive" words $(+o)$ like honest, improve, and hope. Each appearance of $p_{(w,h,a)}$ will also increase value of $s_{(w,h)}$ by one. If each $o \in D$ have a value of integer a then

$$\begin{cases} -o \rightarrow (a = -1) \rightarrow (p_{(w,h,-o)} < 0) \wedge (s_{(w,h)} > 0) \\ +o \rightarrow (a = +1) \rightarrow (p_{(w,h,+o)} > 0) \wedge (s_{(w,h)} > 0) \end{cases} \quad (3)$$

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 a 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 equation (2) since all articles that discuss one or more politician will have $|(w,h,a) \in C_i| > 0$ and the others $|(w,h,a) \in C_i| = 0$.

Assumption 2

For each statement o there are always person w who declares, and person h whose the o are declared to. Thus whether exist o and w then there is always h, and whether exist o 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 a always have references to $p_{(w,h)}$.

Assumption 3

Though assumption 2 will hold for most statements o in C_i, there are some possibilities that it will not. There are some cases where there is no reference of w but o 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

$$\begin{cases} (b_z = who\ keyword) \wedge (b_z \neq w_y) \rightarrow w_{y+1} = b_z \\ (b_z = whom\ keyword) \wedge (b_z \neq h_y) \rightarrow h_{y+1} = b_z \end{cases} \quad (4)$$

This assumption ensure that each o 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 o thus change value of$a = -1$ for $p_{(w,h,-o)}$ to $a = +1$. If there exist two or more g in one sentence then polarity of o will be changed as many times as g appear.

Assumption 6

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

Assumption 7

This paper assumes that all articles published by news media have some sentiment tendency towards politician, and all articles have the same degree of significance. Therefore polarity of C_i which correspond to h change the probable polarity of news media towards h. Thus every $p_{(w,h,a)}$ will change $p_{(news\ media,h)}$with the same value of a and increase $s_{(w,h)}$ and $s_{(news\ media,h)}$ by one.

3. TRAINING SET AND NAIVE BAYESIANMODEL

Previous studies [1,4] showed that Bayesian classifier can be used to classify books references and news in Indonesian. This research adapts and modifies Bayesian classifier models from those researches and [11] so for training set T which consists of C_i therefore $T \ni w,h,a$ and $a = +1 \leftarrow -o$ or $a = +1 \leftarrow +o$ gets

$$P(p_{(w,h)}|s_{(w,h)}) = \frac{p_{(w,h)}}{s_{(w,h)}} \quad (5)$$

where $P(p_{(w,h)}|s_{(w,h)})$ is the probability of $p_{(w,h)}$in given event $s_{(w,h)}$. Since $p_{(w,h)}$ can be either has positive value or negative value and $|p_{(w,h)}| \leq s_{(w,h)}$ then $-1 \leq P(p_{(w,h)}|s_{(w,h)}) \leq 1$. When $P(p_{(w,h)}|s_{(w,h)}) \sim -1$ then w tends to give negative review on h, where $P(p_{(w,h)}|s_{(w,h)}) \sim 1$ states

otherwise. For example when $P(p_{(w,h)}|s_{(w,h)}) = -0.95$ means that w has 95% probability to has negative sentiment towards h. Thus the probability of sentiment polarity of article C_i towards h is

$$P_{c_{(i,h)}} = \sum_{y=1}^{y=n} p_{(w_{(i,y)},h_i)} / \sum_{y=1}^{y=n} s_{(w_{(i,y)},h_i)} \quad (6)$$

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.

	$w_{(1)}$	$w_{(2)}$	$w_{(3)}$	$w_{(4)}$	$w_{(5)}$	$w_{(q)}$
$h_{(1)}$	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)
$h_{(r)}$	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 $s_{(w,h)}$.

Equation (5) and (6) can be represented as pseudocode form below

```
Function current_article_polarity(name h)
   Set list polarity_towards_h =
   get_all_sentiment_probability_towards_h_from
   _database_T
   Set integer all_polarity_towards_h = 0
   Set integer total_event_towards_h = 0
   DOWHILE (polarity_towards_h NOT = NULL)
      all_polarity_towards_h =
       all_polarity_towards_h +
       polarity_towards_h.value
      total_event_towards_h =
       total_event_towards_h + 1
      polarity_towards_h =
       polarity_towards_h.next
   ENDDO
   Return
   all_polarity_towards_h/total_event_towards_h
END

Function probability(name w, name h)
   save_in_database_T(w,h) =
   polarity(w,h)/event(w,h)
   Return polarity(w,h)/event(w,h)
END

Function polarity(name w, name h, boolean o)
   event(w,h)
   IF (o = positive) THEN
      Return 1
   ELSE
      Return -1
END

Function event(name w, name h)
   totalEvent = totalEvent + 1
END
```

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. 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, cleanser, helper, analyzer, and display.

Reader and parser separate each words and punctuation mark from the text. Then these words will be filtered by cleanser. 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.

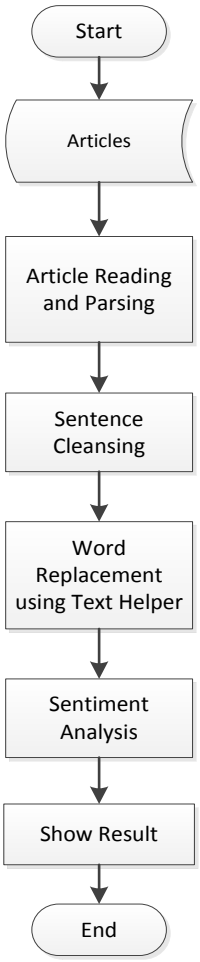

Figure 2. System's activity diagram

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 result system 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.

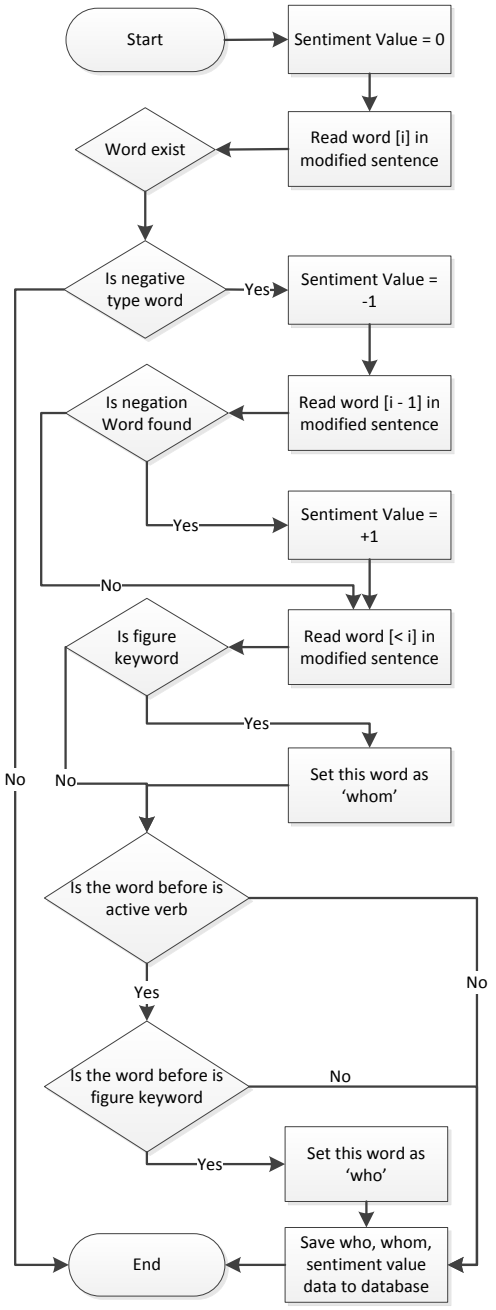

Figure 3. The process flow of negative type words scenario.

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 $h = andi$. $w = 0$ is the default value which equals to $w = K$.

Table 1 and 2 shows the result of P_{C_i} by employing formula (6). Based on assumption 7 then $P_{T_{(andi)}}$ can be calculated by averaging P_{C_i}. Since $C_i \in T$ then $P_{T_{(andi)}} = \frac{\sum_1^n P_{C_{(i,andi)}}}{n} = \frac{-8.5}{10}$.

Thus it signifies K, up to this point, has 85% probability have negative sentiment to Andi. As a result, the general formula for finding sentiment probability of a news media towards one particular politician is:

$$P_{T_{(h)}} = \frac{\sum_1^n P_{C_{(i,h)}}}{n} \quad (7)$$

Table 1. Matrix *M* after running data *T*

i	$P_{C_{(i,andi)}}$
1	-1
2	0.5
3	-1
4	-1
5	-1
6	-1
7	-1
8	-1
9	-1
10	-1

Table 2. Matrix *N* after running data *T*

	$w_{(K)}$	$w_{(kpk)}$
$h_{(andi)}$	34	9

This result was validated by comparing it with human readers. The articles from the training set was given to some respondents to give feedback about sentiments tendencies (-1 for negative, +1 for positive) in the articles for each sentences. The feedback then be calculated using the same formula (6) and give $P'_{T_{(andi)}} = \frac{-6.9}{10}$. The different between $P'_{T_{(andi)}}$ and $P_{T_{(andi)}}$ is the system accurateness. Therefore the system has 81.2% accuracy.

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.

There are two major problems in applying this model. First 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. Second, there is lack of precise measurement method in validating the result. Further studies are still needed.

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. Capeci, 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. Proboyekti, Implementasi Naïve Bayes Classifier Pada Program Bantu Penentuan Buku Referensi Matakuliah, Jurnal Informatika Vol 3No 2 Universitas Kristen Duta Wacana, 2007, 32-36.

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

[6] E.J. Fortuny, T.D. Smedt, D. Martens, W. Daelemans, Media Coverage In Times of Political Crisis:a Text Mining Approach, Expert Systems with Applications, Sciendirect, 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. Luger, Artificial Intelligence Structures And Strategies For Complex Problem Solving, Pearson Education Inc., 2009, 182-185.