ISSN 2356-4407

www.STIKI.ac.id

Bersama memajukan ICT Indonesia

PROCEEDING

IC - ITECHS 2014

The 1st International Conference on Information Technology and Security

Malang, November 27, 2014

Published by:

Lembaga Penelitian dan Pengabdian pada Masyarakat

Sekolah Tinggi Informatika dan Komputer Indonesia

PROCEEDING

The 1st International Conference on Information Technology and Security (IC-ITechs)

November 27, 2014

Editors & Reviewers:

Tri Y. Evelina, SE, MM Daniel
Rudiaman, S.T, M.Kom Jozua
F. Palandi, M.Kom

Layout Editor:

Eka Widya Sari

LEMBAGA PENELITIAN & PENGABDIAN KEPADA MASYARAKAT
Sekolah Tinggi Informatika & Komputer Indonesia (STIKI) – Malang
Website: itechs.stiki.ac.id E-mail: itechs@stiki.ac.id

PROCEEDING

The 1st International Conference on Information Technology and Security (IC-ITechs) November 27, 2014

ISSN 2356 - 4407

viii + 276 hlm; 21 X 29,7 cm

Reviewers & Editors:

Tri Y. Evelina, SE, MM
Daniel Rudiaman, S.T, M.Kom
Jozua F. Palandi, M.Kom

Layout Editor:

Eka Widya Sari

Published by:
LEMBAGA PENELITIAN & PENGABDIAN KEPADA MASYARAKAT
Sekolah Tinggi Informatika & Komputer Indonesia (STIKI) – Malang
Jl. Raya Tidar 100 Malang 65146, Tel. +62-341 560823, Fax. +62-341 562525
Website: itechs.stiki.ac.id E-mail: itechs@stiki.ac.id

GREETINGS

Head of Committee IC-Itechs

For all delegation participants and invited guest, welcome to International Conference on Information Technology and Security (IC-Itechs) 2014 in Malang, Indonesia.

This conference is part of the framework of ICT development and security system that became one of the activities in STIKI and STTAR. this forum resulted in some references on the application of ICT. This activity is related to the movement of ICT development for Indonesia.

IC-Itechs aims to be a forum for communication between researchers, activists, system developers, industrial players and all communications ICT Indonesia and abroad.

The forum is expected to continue to be held continuously and periodically, so we hope this conference give real contribution and direct impact for ICT development.

Finally, we would like to say thanks for all participant and event organizer who involved in the held of the IC-Itechs 2014. We hope all participant and keynote speakers got benefit from this conference.

ISSN 2356 - 4407
The 1st International Conference
on Information Technology and Security

LIST OF CONTENT

News Sentiment Analysis Using Naive Bayes And Adaboost

Erna Daniati
Universitas Nusantara PGRI Kediri
ernadaniati@gmail.com

Abstract

Most of the data that exists today is in the form of digital data. If we sell shares or write a book even sell products online, it always involves an electronic devices. Since the paper transactions is occurred in digital form, a lot of data are available to be analyzed. One of data or information in digital form is information about the news. Information provided by the news provider's website contains a variety of things such as economics, politics, sports, and the others. The news has a variety of interesting patterns to be analyzed. The pattern can be used to predict the sentiment which contained in a few words, phrases, or sentences from a paragraph of news. This research discusses about sentiment analysis to the word or phrase which is used as test data to produce several classes such as positive, negative, and neutral sentiment. The method that used to weighting is using TFIDF word, then labeling of sentences is using sign numbers multiplication. Next, training and testing are using Naive Bayes method with a combination of AdaBoost.

Keywords *: news, sentiment, tfidf, Naive Baye, AdaBoost*

Introduction

Most of the data that exists today is in the form of digital data [1]. If we sell shares or write a book even sell online products, it always involves an electronic devices. Since the transaction occurs in the form of digital paper, many of big data is available for analysis. One type of data or information that stored in digital form is information about the news. Information provided by the news provider's website contains a variety of things such as economics, politics, sports, and the others. The news has a variety of interesting patterns for analysis.

The pattern can be used to predict the sentiment which contained in a few words or phrases of paragraphs news. The tested word or phrase can contain sentiment in the form of positive, negative, and neutral from some of the captured information. Some taken sentiment of a sentence can be used as a graphic composition sentiment. It makes anyone who sees it will know how big the resulting composition sentiment that made ??him decide to take action related to the chart provided.

News sentiment analysis is one of the materials related to natural language processing, text analysis, and computational linguistic approaches to identifying and extracting information in the source material [2]. Basic sentiment analysis is classifying polarity in the document, sentence, or the level of aspects. It means wheater the opinion in the document, sentence or aspects of entities are positive, negative, or neutral. Further more, the polarity of sentiment classification can be emotional status such as angry, sad, happy, and upset.

This research discusses about sentiment analysis to the word or phrase that is used as test data so it produces several classes such as positive, negative, and neutral sentiment. The method to weighting term is using TFIDF. Then, training and testing data are using Naive Bayes method with a combination of AdaBoost. AdaBoost boosting techniques is also

compared with no use of these techniques to prove whether AdaBoost can improve or not the accuracy of the testing of a entered phrase or word.

Literature review

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file "MSW_USltr_format". Opinion Mining or Sentiment Analysis is approach to analyze opinions that extract various information sources such as blogs, comments in forums, product overview, policies or multiple topics within social networks [3]. In this case, there are some big challenges in sentiment analysis. One is a Word Sense which is a classic NLP encountered problem frequently. This problem is conducted in former research to review the efficient techniques, rate of progress, and future research directions in terms opinion mining and sentiment polarity classification. There are some things to be done when sentiment analysis occured. These are the detection of subjectivity, negation, and the sentiment classification based on the features. Classification may use several techniques such as Naive Bayes compared, Maximum Entropy, and Support Vector machines.

Survey of research methods in sentiment analysis is also performed in previous study [2]. That study discussed the corpus used in the form of blogs, micro-blogging and data sets. The study also compared several classification techniques such as Naive Bayes, maximum entropy, support vector machines, K-nearest neighbor and Winnow. It produced some review of the classification, summary, and real-time applications. The data provided in this study proved that the classification model has several advantages from one to the other depending on a variety of different types of features distribution.

The development of research in terms of knowledge about humanity and social attempt to mine the text of government documents, literature, magazines, news and social media content [1]. New area of research is a large part of the topic of big data. It is using data and information in big scale which generated from large amounts of human daily activities and its behavior. Large data can be employed for business activities such as predicting election, determining business decision, and further. This previous research was about the opportunities and challenges of text mining which is the basic of sentiment mining that will be used by the library and its officers. In basic level, the librarian needed to start thinking about copyright, how to developed a collection and provided access to a corpus of news content, whether text mining role be to teach and train researchers on the strengths and weaknesses of the new database and other sources, and how they able developed traditional role of support locations and information retrieval.

Then, there is also previous research that discusses the development and use of applications to classify Indonesian text(supervised), by applying naive Bayes method [4]. It was tested with two inputs namely using and without stop word removal from the Naïve Bayes algorithm and method. The algorithm was held cross validation testing for 10 times (10 folds validation). It was done by dividing the data into 10 sub-sample test. The ratio of test data was started from 10% and then up to 10% times for each time testing until 90%. Each ratio test was performed 10 times and the desired output is value of the its average accuracy. From the experiments that have been done, the implementation of naive Bayes method in the classification of news had good accuracy on test data. It was proved in test data from the website which generated accuracy with a high percentage of more than 87% for large training data (100 articles). Classification can run well on the training data more than 150 documents.

Theory Basic

Natural Language Processing

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable. Natural Language Processing (NLP) is a field of research and application that explores how computers can be used to understand and manipulate natural language texts or talks to do things that are useful [5]. NLP researchers aim to gather knowledge about how humans understand and use the language so that the appropriate tools and techniques can be developed to create computer systems which understand and manipulate natural language to perform the desired tasks. The foundations of NLP lies in a number of disciplines, the computer and information science, linguistics, mathematics, electrical engineering and electronics, artificial intelligence and robotics, psychology, and others. NLP application includes a number of subject areas, such as machine translation, natural language processing and summarizing text, the user interface, multilingual and cross-language information retrieval (CLIR), voice recognition, artificial intelligence and expert systems, and so on.

The field of computational linguistics (CL), together with engineering domains Natural Language Processing (NLP), have become popular in recent years [6]. It has grown rapidly from a relatively obscure addition of AI and formal linguistics to an evolving discipline. It has also become an important area of industrial development. The focus of research in the CL and NLP has shifted over the past three decades from a small prototype studies and theoretical models for learning and robust processing system which applied to large corpora.

Sentiment Analysis

Generally, text categorization classify documents by topic. As usual keyword search, would not be accord to take all kinds of opinions [3]. Therefore, it is necessary to use sophisticated opinion extraction method. Sentiment analysis is a natural language processing technique that helps to identify and extract subjective information in a source. Sentiment analysis aims to determine the attitude of the author towards some topics or contextual polarity of the whole document. Attitude is an assessment or evaluation, affective status, or the intended emotional communication. The basic objective in sentiment analysis is classifying given polarity text at the document level, sentence, or feature/aspect (whether the opinions expressed in the document, sentence or entity feature/aspect in positive, negative, or neutral). In addition to the polarity of sentiment classification, some emotional status such as angry, sad and happy can also be identified. One of the sentiment analysis challenges is determining the opinions and subjectivity of the study. Subjectivity is very context sensitive and often strange expression for everyone. Negation Detection and subjectivity are preprocessing steps are most important to achieve efficient impact opinion.

Natural Language Processing and Programming Languages are a field of Computer Science in which each generated with a long tradition of research. Although it is centered around the theme of language, but there is little interaction (if any) among several grammar. Starting with a particular language text, it does show how a natural language programming system can automatically identify the steps, loops, and comments, and turns it into a platform that can be used as a starting point for writing a computer program. It is expected to be very useful for them who begin to learn how to program runs [7].

NLP can describe syntactic information (eg, part-of-speech tagging, chunking, and parsing) or semantic information (eg, word-sense disambiguation, semantic role labeling, named entity extraction, and anaphora resolution). Text corpora has been described in the

manual with the structure of the data to compare the performance of various systems. Availability of standard benchmark has encouraged research in Natural Language Processing (NLP) and effective system has been designed for all these tasks. This systems are often seen as software components to build real-world NLP solution [8].

Naive Bayes

Bayes theory is named after Thomas Bayes, an English non-conformist priest did early work in probability and decision theory during the 18th century [9]. Given X is multiple rows of data. In Bayesian terms, X is considered evidence. As usual, this is explained by the measurements performed on a set of attributes n. There are several hypotheses H, such that multiple rows of data X belongs to class C is determined. For classification problem, if you want to determine P(H|X), then the probability that hold the hypothesis H is given the evidence or observed data tuple X. In other words, to seek the probability that tuple X belongs to class C, it has to consider that we know the description of the attribute X.

P (H|X) is the posterior probability, or a posteriori probability, of H conditioned on X. For example, our world in the form of multiple rows of data are limited to the customers which described by the attributes age and income, respectively, and that X is a customer in 35 year with an income of $ 40,000. Suppose H is a hypothesis in which the customer will buy a computer. Then, P(H|X) reflects the probability that customer X will buy computer , consider that we know the age of customers and revenue.

In contrast, P(H) is the prior probability, or a priori probability, H. For example, this is the probability that any given customer will buy a computer, regardless of age, income, or other information, for that matter. Posterior probability, P (H|X), is based on more information (eg, customer information) from the previous probability, P (H), which is independent of X.
Similarly, P (X|H) is the posterior probability of X conditioned on H. That is, it is the probability that a customer, X is in 35 years and generates $ 40,000, given that we know the customer will buy a computer.

P(X) is the prior probability of X. In the previous example, this is the probability that a person of some customers are in 35 years and generate $40,000. P(H), P(X|H), and P (X) can be estimated from the data given. Bayes theorem is useful in that it provides a way to calculate the posterior probability, P(H|X), P(H), P(X|H), and P(X). Bayes cooperative equation shown in (1).

$$P(H|X) = P(X|H) \,.\, P(H) / P(X) \qquad (1)$$

Adaboost

Boosting is an approach to machine learning based on the idea to create a highly accurate prediction rule by combining many of the rules are relatively weak and inaccurate [10]. AdaBoost algorithm of Freund and Schapire is boosting algorithm that is first put into practice, and remains one of the most widely used and studied, with applications in various fields. Over the years, various ways have attempts to explain the AdaBoost learning algorithm, namely, to understand why it works, how it works, and when it works (or fails). It is understanding the nature of learning in the corporate foundation both in general and related to specific algorithms and phenomena that the field able move forward. The steps of the Adaboost algorithm is shown in Fig. 1.

Research Methods

In this research uses some methods procedurally. Several methods of the study are as follows:

- The study of literature. This phase is done by finding and collecting some literature such as journals, theses, books and articles both offline and online.
- Analysis. This phase is carried out by analyzing the problems and needs of this research. Problem analysis is done to study and understand the problem domain well for overall analysis of problems, opportunities, and constraints.
- Design. It is composed of two step: logical and physical design.
- Implementation. This phase is done by realizing or building system design that has been done.
- Testing. The main objective in this phase is comparing some of the features and accuracy of existing clasifier in this study with previous research so found advantages and disadvantages of.

In the design phase has an overview of the system. General description of the system is performed in Fig. 2. In Fig. 2, there is a process of crawling news website. Crawling is the process of finding a web address and store documents that presented in the website. Crawling in this research only takes a document in the form of news. So, before this process occurs, it is necessary to provide some website address with just the news category. Crawling will explore the website address that has been provided and stores some words and phrases in system. This process is done repeatedly to get the latest data as well. Repetition of this process requires a regulated pause once every 3 minutes so it is necessary a trigger based on a specified timer. This process will also check the uniqueness of each document so there is not redundancies in document storage.

In Fig. 2, there are also several important phases. The phase of the preprocessing and training. The preprocessing phase consist of two steps. The first step in the early phase of training before and there is a second step in the testing phase. At the beginning of the preprocessing phase is tokenization. This process aims to eliminate the existing punctuation in a word or sentence. This punctuation need to be removed because it can not be conducted word processing. After punctuation had been omitted, the next step is breaking a few sentences into words. Breaking sentences into words is done by identifying the presence of spaces between words. Once the identification is done, the sentence is broken down (splitting) based on space so generate words.

Given: $(x_1,y_1),\ldots,(x_m,y_m)$ where $x_i \varepsilon X$, $y_i \varepsilon Y$ {-1,+1}
Initialize: $D_1(i) = 1/m$ for $i=1,\ldots m$.
For $t=,\ldots,T$;

- Train weak learner using distribution D_t.
- Get weak hypothesis $h_t : X \rightarrow$ {-1, +1}.
- Aim: select h_t with low weighted error:

$$\varepsilon_t = Pr\sim D_t [h_t(x_i) \neq y_i]$$

- Choose $\alpha_t = \frac{1}{2} \ln[(1 - \varepsilon_t) / \varepsilon_t]$
- Update, for $i = 1,\ldots,m$:

$$D_{t+1}(i) = D_t(i)\exp(-\alpha_t y_i h_t(x_i)) \cdot (Z_t)^{-1}$$

Where Z_t is normalization factor (chosen so that D_{t-1} will be a distribution).
Output the final hypothesis:

$$H(x)= \text{sign}[^T\Sigma_{t=1} \alpha_t h_t(x)]$$

Fig. 1. Adaboost Algoritm

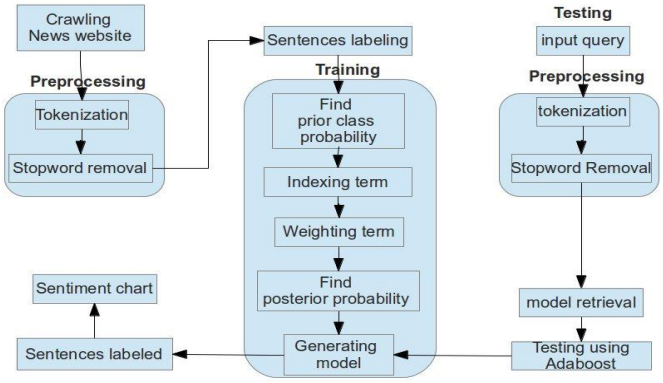

Fig. 2. General description of the system

The next step in the preprocessing phase after tokenization is stopword removal. It will eliminate the words that less contribute in meaning of sentence. Words Type including stopword is a conjunction, preposition, adverb of time, and auxiliary verbs. The process begins by creating a stopword list in a database. The next step is doing some scanning of words that obtained from tokenization. If there is a word similar to the word in stopword list, it should be removed. Removal of this word will also affect accuracy and improving efficiency at storing in database.

Disscussion

There is a phase of labeling the sentence in Fig. 2. This phase is begun with grouping some word formerly. In this study, the group consists of 23 types of word or part-of-speech. These word types are a common noun, primary numeral, adjective, intransitive verb, transitive verbs, particle, interjection, subordinating conjunctions, adverb, genetive pronoun, proper pronouns, modal, irreguler numeral, negation, WH-Pronouns, coordinating conjunctions, prepositions, locative pronouns, collective pronouns, determiner, number pronoun, ordinal numeral. Some words are also given classes that defined in accordance with the sentiments in this research. The sentiment is positive, negative, and neutral. Thus, a collection of words with kind words and sentiments class will form a dictionary word.

Grouping the word is intended to determine who owned a sentiment word when meeting with other words. Given example, there is noun with the positive sentiment if meet with a negation word then the word will bear grudges both negative. Examples of grouping words based on the word type is indicated in Table 1. Defining sentiment of a sentence in this study is using the rule of number sign multiplication. The rule is indicated in Fig. 3. To clarify labeling or grading on a word in order to be training, then given example sentences in Fig. 4.
Sentence in Fi. 4 is performed stopword removal that does not contribute to the determination of a sentence sentiment. Some kind words including stopword is a preposition and conjunctions. After the sentence in Fig. 4 is performed stopword removal and matching the dictionary, it will generate a new sentence in Fig. 5. The sign number multiplication is applied in Fig. 5 so it generates a multiplication in Fig. 6.

Multiplication results in Fig. 6 is positive. Actually, Fig. 4 there are two clauses. The first sentence is "Wildfire is occurred in electrical shop". Then, last sentence is "fire has been slaked fast" should have a positive sentiment. However, there is a word "but" so done the negation of the positive sentiment that becomes negative. Thus, the positive sentiment generated because there is multiplication of negative words with negative sentiment. The next step in this research is seeking the prior probability of a class sentiment in the labeled corpus. In the previous Bayes equation, there is a notation P(H) which states the probability of class to the overall sentiment of a sentence or corpus. Thus, to determine the probability of class sentiment, it can be written in (2). Thus, the probability of a class can be found by determining

the amount of a certain grudges sentences divided by the number of whole sentences are provided for the training data.

Having determined the probability of class sentiment, the next step is to perform indexing and weighting of the word. Indexing is done by collecting all the different words from the existing corpus. So, all the words contained in a corpus is collected in a dictionary but just in different words. Then, each word is weighted against the documents on which it appears. In this study, the grade is weighted based on the existing sentence. Weighting of each word use tf-idf method. This method aims to calculate the weight of the fatherly words that indicate the level of importance of a word to document. In tf-idf method, there is relationships of emerging word frequency with many documents that load which shown in (3) and (4) [11].

Table 1. Example of Word Dictionary

No.	Word	Part-Of-Speech	Symbol	Sentiment
1.	wildfire	common noun	NN	negative (-)
2.	occurred	intransitive verb	VBI	neutral (n)
2.	in	prepositions	IN	neutral (n)
3.	shop	proper pronouns	NNP	neutral (n)
4.	electrical	proper pronouns	NNP	neutral (n)
5.	But	negation	NEG	negative (-)
6.	fire	common noun	NN	neutral (n)
7.	has	modal	MD	positive (+)
8.	slaked	adverb	JJ	positive (+)
9.	fast	adverb	JJ	positive (+)

positive x positive = positive. positive x negative = negative. negative x negative = positive.		positive x neutral = positive. negative x neutral = negative. neutral x neutral = neutral

Fig. 3 Rules of multiplication number sign

Wildfire is occured in electrical shop but fire has been slaked fast.

Fig. 4 Sentence example

Wildfire(NN/-) occured(VBI/n) in(IN/n) electrical(NNP/n) shop(NNP/n) but(NEG/-) fire(NN/n) has(MD/+) slaked(JJ/+) fast(JJ/+)

Fig. 5 Sentence example after stopword removal

NN/- . VBI/n . IN/n . NNP/n . NNP/n . NEG/- . (NN/n . MD/+ . JJ/+ . JJ/+) = +

Fig. 6 Multiplication of word sentiments

$$P(H) = n \text{ (sentiment sentence x) / n (sentiment)} \quad (2)$$

$$idf\,(j) = log\,(N/df(j))) \quad (3)$$

$$tf\text{-}idf\,(j) = tf(j\,) * idf(j\,) \quad (4)$$

There is the notation *tf* which is frequency of the number of words or terms in (3) that appear in a sentence. The frequency will be multiplied by the frequency index words of document load. The index is performed in (4). There is notations *N* that is the total number of documents in (4), while the notation *df*(*j*) indicates the number of documents that contain the word or term *j*. In this study, a document is represented by a sentence. So, if there is *df*(*j*) then it means the number of sentences with particular sentiment containing word *j*. Likewise, *tf*(*j*) is the frequency of the word that appears in a sentence with a certain sentiment. There are 3 sentiment is positive, negative, and neutral so that every word has a weight corresponding to each of the third sentiments. This will be used to define the probability of each features in the training phase using a Naive Bayes which is indicated by the notation *P*(*X*/*H*).

Features generated from each word is a the weight value from each of the specified class sentiment. Sentiment classes is obtained after the calculation of the executed posterior class. So, any word or term has the 3 features in a weight value of each sentiment. Some of these terms are collected in a dictionary. A dictionary is a model generated from the training phase. The model is used to classify the sentences that do not have labels wheater positive, negative, or neutral sentiment.

Furthermore, testing of sentences which do not have the sentiment label Naive Bayes also uses methods that also involves previously generated models. Tests using a Naive Bayes can be determined using (5). In that equations, each feature is represented by the weight of each sentiment class which determined its posterior probability. After the posterior probability of each feature is generated then it is compared each one with the other features. Class predictions is generated by looking at the largest value of each of its posterior probability. If one class sentiment has the largest posterior probability values among others, the class of the sentence is that sentiment. For example, a sentence does not have class sentiment. After training, It has been generated probability of each feature. Then, the calculation of the posterior probability of the probability of each class is performed so resulting posterior probability of sentiment class positive= 0.0045, negative= 0.0017, and neutral = 0.00011. So, class prediction results of this sentence is positive because the largest value of the posterior probability of class sentiment is positive.

$$classify(x_1,...,x_n) = argmax\ p(C=c) \prod_{i=1}^{n} p(X_i=x_i|C=c) \qquad (5)$$

The resulting class prediction using Naive Bayes is not entirely accurate. Many studies have shown the accuracy of below 100%. However, the accuracy of which is still below the expectation can be increased again. Adaboost method that described earlier can improve the accuracy of these predictions. Before applying this algorithm, the model must be generated from the training process. Adaboot algorithm has been shown in the Fig. 1.

This algorithm will take some distribution *D* of training data. Taken initial distribution D_1 is 1/*m* of the entire training data, where *m* indicates the number of defined classes. If the number of classes in this study is 3 then the initial distribution of D_1 is $^1/_3$. In this distribution is sought its error values. Error value is the total sum of the amount classifier prediction that is correct with the given initial labeled data. In Adaboost algorithm, the calculation to find the error is shown in (6).

$$\varepsilon_t = Pr_{i \sim D_t} [h_t(x_i) \neq y_i] \qquad (6)$$

The amount of error is used to calculate that shown in (7). However, cases which have more than two classes, there are some calculations that are added to (7). Previous study adds that calculation as shown in (8). There is additional *log*(*K* - 1) in (8). Notation *K* is the number of classes in the corpus. The addition of this calculations provide significant results in which this has been done in previous study [12]. In previous study, it has been shown that the error

values decrease with increasing iteration. If using adaboost method typically with more than 2 classes, error is relatively stable value as the number of iterations. When compared with (2), the value of the error is smaller.

$$\alpha_t = ½\ ln(\ (1-\varepsilon_t) / \varepsilon_t) \quad (7)$$

$$\alpha_t = ½\ ln(\ (1-\varepsilon_t) / \varepsilon_t\ ) + ln(K-1) \quad (8)$$

The next step is making changes to the distribution of training data. This change is shown in (10). Equation 10 which contained the notation Z_t is the constant normalization. Constant is calculated by summation of the D_t at each iteration. This is shown in (11). After the iteration is complete, it is obtained varying values of α_t according to the previous iteration. Value of α_t is used to predict the unknown class sentiment of a sentence. Prediction is done by summing α_t that multiplied with the hypothesis for each iteration. Prediction is shown in (12). In (12), the final results of class prediction is 1 or -1. This does not indicate that predictions can only produce 2-class. However, more than 2 classes can be predicted with using one vs. all methode.

$$D_{t+1}(i) = D_t(i)\exp(-\alpha_t y_i h_t(x_i)) \ .\ (Z_t)^{-1} \quad (10)$$

$$\square_{\square\square} = {}^{m}\Sigma_{i=1}\, D_{t+1}(i) \quad (11)$$

$$H(x) = \text{sign}[{}^{T}\Sigma_{t=1}\, \alpha_t\, h_t(x)] \quad (12)$$

The use of adaboost method can improve the prediction accuracy class. The accuracy of this prediction is shown by the declining value of the prediction error [13]. Fig. 7 shows a comparison of methods adaboost with other classifier. The results of the study in Fig. 7 is the case of text categorization using the 4 methods namely Naive Bayes, TFIDF probability, Rocchio and Sleeping Expert. There are two graphs on the left and right in Fig. 7. The left side of chart is taken from the Reuters news article corpus while the right side of the chart is taken from the Main title AP corpus. This indicates that the increasing class on two different corpus, the use of Adaboost method still produces a low error rate compared with the 3 methods which mentioned above.

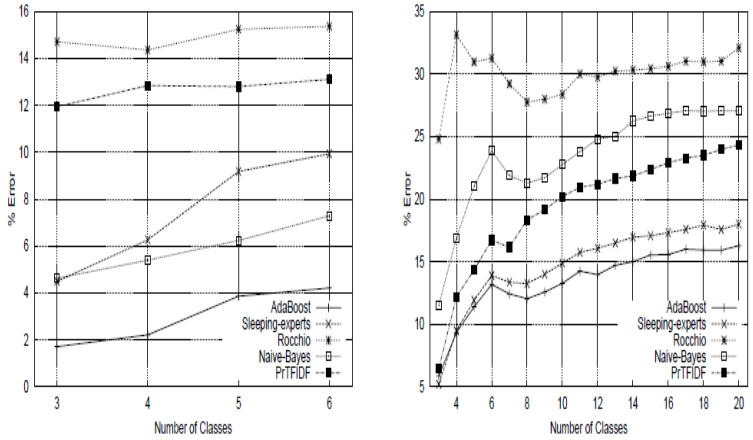

Fig. 7 Comparission of Adaboost, Naive Bayes, Probabilitas TFIDF, Rocchio dan Sleeping Expert [13]

Conclusion and recomendation

This study has been carried out several measures in accordance with the formulation of the problem and the objectives that have been identified in the introduction chapter. The conducted steps have generated some of the conclusions and the lack that be suggestions for future research. The conclusion of this study is:

- Labeling on each sentence using the multiplication sign numbers.

- The system will be built using the Naive Bayes method for training and testing stage later in the testing used a combination of Naive Bayes and Adaboost.
- Merging the Naive Bayes and Adaboost increase than only Naive Bayes.

References

[1] D. Chaney, "Text mining newspapers and news content: new trends and research methodologies", Singapore, Juli 2013.

[2] G.Vinodhini dan R.M. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", India, vol. 2, Juni 2012.

[3] C. Sindhu dan S. ChandraKala, "Survey on Opinion and Sentiment Polarity Classification", India , vol. 3, Januari 2013.

[4] C. Darujati dan A.B. Gumelar, "Pemanfaatan Teknik Supervised untuk Klasifikasi Teks Bahasa Indonesia", Indonesia, vol. 16, pp. 1858 - 4667, Februari 2012.

[5] G.G. Chowdhury, "Natural Language Processing", Glasgow UK, vol. 37, pp. 51-89, 2003.

[6] A. Clark, C. Fox, dan S. Lapping, The Handbook of Computational Linguistics and Natural Language Processing, Malden: Wiley-Blackwell, 2010.

[7] R. Mihalcea, H. Liu, H. Lieberman, "NLP (Natural Language Processing) for NLP (Natural Language Programming)", Mexico City, 7th Intl. Conf., Februari 2006.

[8] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, "Natural Language Processing(Almost) from Scratch", Princeton US, vol. 12, pp. 2493-2537, Agustus 2011.

[9] J. Han dan M. Kamber, Data Mining Concept and Techniques 2nd, San Fransisco: Morgan Kaufman Publisher, 2006.

[10] R.E. Schapire, Explaining AdaBoost, Springer, 2013.

[11] S.M. Weiss, N. Indurkhya, dan T. Zhang, Fundamentals of Predictive Text Mining, London: Springer, 2010.

[12] J. Zhu, H. Zou, S. Rosset, dan T. Hastie, "Multi-class adaboost", vol.2, pp. 349-360, 2009.

[13] Y. Freund dan R.E. Schapire, "A Short Introduction to Boosting", vol.14. pp.771-780, September 1999.