Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 116 (2017) 137–143

Procedia Computer Science

www.elsevier.com/locate/procedia

2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017, Bali, Indonesia

News Article Text Classification in Indonesian Language

Rini Wongso*, Ferdinand Ariandy Luwinda, Brandon Christian Trisnajaya, Olivia Rusli, Rudy

Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. K.H. Syahdan No. 9, Jakarta 11480, Indonesia

Abstract

This research intends to find the appropriate algorithm to automatically classify a news article in Indonesian Language. We obtain our dataset which is taken by using a web crawling method from www.cnnindonesia.com. First of all, the document will first undergo some Text Preprocessing method in the form of Lemmatization and Stopwords Removal. The reason we are doing the Text Preprocessing step before anything else is to minimize the noise in the document. Next, we apply Feature Selection onto the document to further separate important words and less important words inside the document. After applying Feature Selection, the document will be classified by the classifier. We are comparing the TF-IDF and SVD algorithm for feature selection, while also comparing the Multinomial Naïve Bayes, Multivariate Bernoulli Naïve Bayes, and Support Vector Machine for the Classifiers. Based on the test results, the combination of TF-IDF and Multinomial Naïve Bayes Classifier gives the highest result compared to the other algorithms, which precision is 0.9841519 and its recall is 0.9840000. The result outperform the previous similar study that classify news article in Indonesian language which obtained 85% of accuracy.

© 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.

Keywords: Classification; *Feature Selection*; TF-IDF; Multinomial Naïve Bayes

1. Introduction

The technological advances in Computer Science field especially in the past few decades have made it possible for huge volume of data available anytime and anywhere. However, with the many types of information available,

* Corresponding author. Tel.: +62-21-5345830; fax: +62-21-5300244.
E-mail address: rwongso@binus.edu

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.
10.1016/j.procs.2017.10.039

grouping existing information is becoming more challenging and thus could lead to information overload [1]. The ability to classify document into certain categories is helpful to face information overload. Automatic document classification is developed as manual work is no longer effective [2]. When done automatically, people won't be required to think about which category a document or text belongs to.

According to the survey done by Aggarwal & Zhai [3], there are some algorithms which can be used for text classification, such as: Chi Square, Information Gain, Term Frequency – Inverse Document Frequency (TF-IDF), Gini Index, Mutual Information, Supervised LSI, Supervised Clustering and Linear Discriminant Analysis (LDA). In a comparative study done by Zhang [4], they evaluate three methods of TF-IDF, LSI and multi-word text representation in information retrieval and text classification. The documents used were Chinese and English documents. Based on their experiment, in Chinese information retrieval, TF-IDF shows the best result followed by LSI and multi-word. In Chinese documents classification, LSI is superior to TF-IDF, while multi-word still give the worst result. In English documents, both for information retrieval and documents classification, LSI outperforms TF-IDF, but multi-word still stay behind.

One of the difficulties in feature selection is the number of data, hence it is better if dimensionality reduction is applied [5]. Dimensional reduction can be achieved by using such method as Singular Value Decomposition (SVD), which is applied in LSI to projects document vectors into an approximated subspace in order to represent semantic similarity. SVD and LSI algorithm is used in phishing email detection research [6]. In recent years, the use of TF-IDF is still popular although the despite the huge number of data extracted from TF-IDF. In 2014, research from [7] use TF-IDF with KNN to categorize 500 online documents from 20_Newsgropup dataset which achieve more than 80% accuracy in average.

According to Aggarwal [3] there are several model of classifier such as: Probabilistic Classifiers, Naive Bayes Classifiers and Linear Classifiers. Naïve Bayes Classifiers dan Support Vector Machine are the classifier that is used the most to solve Document Classification problem and they both provide a quite promising result. Using the right pre-processing treatment, Naïve Bayes can provide promising accuracy result. The research that conducted by Trivedi [8], result of Naïve Bayes classifier's precision and recall are higher. However, in Rennie's research [9], SVM is better than one type of Naive Bayes Classifier, which is Multinomial Naive Bayes Classifier. Research conducted by Ramdass [10] used Naive Bayes classifier to classify newspaper's article. Meanwhile, research by Liliana [11] that used Support Vector Machine as classifier to classify newspaper's article in Indonesian language from Kompas.com and obtained result of 85% in accuracy.

There are many other researches for document classification, but unfortunately, it is still poorly explored for Indonesian language. This study will use combination of methods used in previous researches and implemented them with data source of news article from www.cnnindonesia.com. Thus, this research intends to find which combination of feature selection and classifier that can give best result in order to improve classification for newspaper's article in Indonesian Language.

2. Methodology

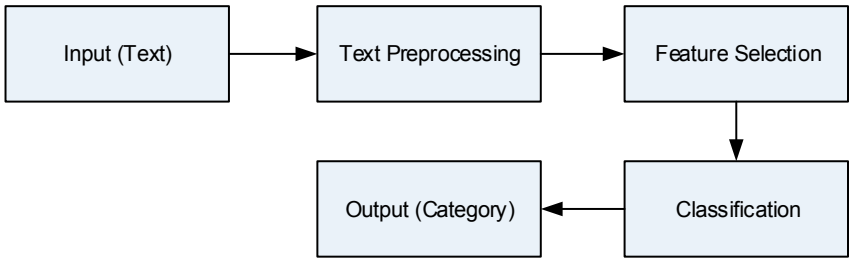

Fig. 1. Methodology

Figure 1 above illustrates the methodology proposed in this research. Input in the form of text is obtained by using web crawling and then the result is preprocessed for feature selection. Classification is later done using selected classifier to generate output in category.

2.1 Web Crawling

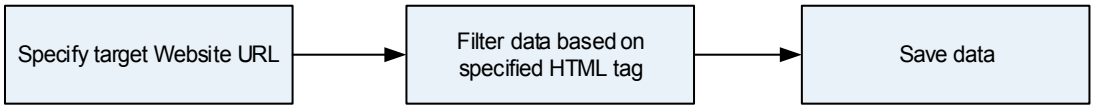

Fig. 2. Web Crawling

First method that used in this research is web crawling, which is used to obtain the dataset. The process is described in Figure 2 which begins by specifying URL from selected website. Then crawler will automatically find all the link that is found in HTML page and will arrange them. From the arranged URL, crawler will then download content of the page [12]. In this research, news articles which are publicized in www.cnnindonesia.com are crawled with the total number of 5,000 documents. These documents consist of 1,000 documents for each category of: Economy, Health, Sports, Politic, and Technology. The documents are randomly split with the ratio of 80:20 for training and testing purposes.

2.2 Text Preprocessing

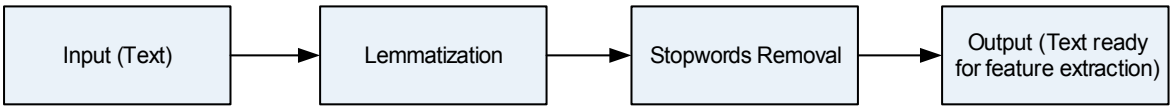

Fig. 3. Text Preprocessing

The next step is to do text processing. As can be seen in Figure 3, the first step begins with lemmatization, which aims to get the base form of each words. Then, stopwords removal is also applied to remove less meaningful words, such as prepositions and conjunctions (ex: apa, yang, dimana, ketika, di, saat, etc) [13]. The last step is to determine which feature selection and classifier that will be used.

2.3 Feature Selection

The feature selection in this research are done by using TF-IDF and SVD. The two methods are chosen based on the previous works.

2.3.1 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is the most popular term weighting scheme in information retrieval [14]. TF-IDF is combination of TF and IDF, first calculate TF of each feature with this formula:

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} \tag{1}$$

f_{ij} is the frequency of term i in document j, while $max_k f_{kj}$ is the frequency of the most common term(term with the highest frequency) in document j(symbolized as term k).

Then calculate the Inverse Document Frequency (IDF) value. Value of IDF for a term i can be obtained with the following formula:

$$IDF_i = \log_2\left(\frac{N+1}{n_i+1}\right) + 1 \tag{2}$$

N is the number of training document used and n_i is the number of training document that contains the term i [15].

In this case, value of 1 is added to avoid *Zero Divisions*. After the values of TF and IDF are obtained, TFIDF can be obtained with this formula:

$$TFIDF_{ij} = TF_{ij}.IDF_i \quad (3)$$

2.3.2 Singular Value Decomposition (SVD)

Calculation with SVD is done after TF-IDF. Vector that obtained after TF-IDF calculation, will be used to calculate SVD. Given a term-document matrix A=[a_1,a_2,…,a_n], SVD can be calculated using:

$$A_{mn} = U_{mm}\ x\ S_{mn}\ x\ V_{nn}^T \quad (4)$$

The result from the feature selection will be processed with the classifier.

2.4 Classification

Result from feature selection is used for classification. The classifier used in the experiments of this research is Naïve Bayes and Support Vector Machine.

2.4.1 Naïve Bayes Classifier

Naïve Bayes calculation is based on Bayes' theorem with assumption that there is no word related to each other (every word is independent). [16] First step is to calculate the probability of each class, with this formula:

$$\hat{P}(c) = \frac{f_c}{f_d} \quad (5)$$

, where f_c is number of training document(s) which is labelled with c class and f_d is the number of all training documents. Then, probability of document to each class is calculated with this formula:

$$\hat{P}(c|d) = P(c) \prod_{i=1}^{n} P(x_i \in d|c) \quad (6)$$

From this formula, a class with the highest probability result will be assigned as the class of document d [16]. There are some variations of Naïve Bayes Classifier used in this research, which are: Multinomial Naïve Bayes Classifier, Multivariate Naïve Bayes Classifier, and Gaussian Naïve Bayes Classifier.

2.4.2 Support Vector Machine

For this research, the Non-Linear Support Vector Machine using the RBF kernel is used, with multi-class SVM one against one strategy. One against one strategy is a method for multi-class classification in support vector machine which pairs every single possible class with each other, resulting in more than 1 classifiers [17]. One against one strategy is also known as pairwise coupling or round robin, which consists in constructing one SVM for each pair of classes [18].

After creating the necessary number of classifiers, each classifier will then create a hyperplane using the following formula of solving the Dual Lagrange problem by finding the value of α:

$$Maximize\ Ld\ (\alpha) = \sum_{i=1}^{N} \alpha i - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j x_i{}^T x_j \quad (7)$$

$$Subject\ to \begin{cases} 0 \leq \alpha_i \leq C \\ \sum_{i-1}^{N} \alpha_i y_i = 0 \end{cases} \quad (8)$$

y = -1 or +1(-1 is for the class c, and +1 is for the other class)

x = vector of document i(or j)

N = Training Document for each class

We will also require to find the value of b to assign a class to incoming input documents.

$$b = y_i - \sum_{i-1}^{SV} \alpha_i y_i K(x_i, x_j) \quad (9)$$

The RBF Kernel that is used has the following formula:

$$K(x_i, x_j) = e^{(-\gamma ||x_i - x_j||^2)}, \gamma > 0 \quad (10)$$

γ= user specified value.

After all the above values are computed, the hyperplane is created and each time an input document is submitted, the following formula will be used to decide which class the document belongs to:

$$d(x) = sign\left(\sum_{i-1}^{SV} \alpha_i y_i K(x_i, x_j) + b\right) \quad (11)$$

SV = number *of* support vectors (training documents that lies in the margin of the hyperplane)

The result is either +1 or -1, which classifies the document to a corresponding class.

Since this research uses a one-against-one approach for the multi-class SVM variant, every new document will go through all available classifiers, and then, the class with the most voted amount will be the designated class for the input document.

Then, the experiment that is conducted follows the following procedure:

- In this research, each of the document is previously labelled with their respective class, and each of the documents belong to one of the 5 classes (Economy, Health, Technology, Sports, and Politics)
- Text preprocessing: Each of the documents used as training or testing will be lemmatized and stopwords removal is applied on them, resulting in a new document that will be processed.
- Feature Selection: TF-IDF or TF-IDF + SVD will be applied on each of the training and testing documents.

Classifiers: Each of the explained classifiers above will be tested in this phase using the data which is previously processed by the Feature selection algorithm.

3. Result and Discussion

After conducting the research, the following results are obtained:

Table 1. Comparison of Accuracy Each Algorithm Combination.

Combination Used	Precision	Recall	Time (Seconds)
TFIDF + GNB	-	-	-
TFIDF + BNB	0.9822558	0.9820000	0.7015419
TFIDF + MNB	**0.9841519**	**0.9840000**	**0.7020838**
TFIDF + SVM	0.9794023	0.9790000	74.9765017
TFIDF + SVD + GNB	0.3591179	0.3460000	11.4571054
TFIDF + SVD + BNB	0.3925421	0.3050000	11.8058102
TFIDF + SVD + MNB	-	-	-
TFIDF + SVD + SVM	0.4360882	0.3910000	1.3520746

As can be seen in Table 1, the combination of TFIDF + Multinomial Naïve Bayes (MNB) provides the highest value of precision and recall, which is around 98.4% followed by the combination of TFIDF and Bernoulli Multivariate Naïve Bayes (BNB) which is around 98.2%. These results are supported by previous work which stated that MNB tends to outperform BNB if the vocabulary size is relatively large [19]. As for the 2 combinations TFIDF + Gaussian Naïve Bayes (GNB) and TFIDF+SVD+MNB, no results are obtained due to the requirements needed (MNB cannot accept vectors that have negative values and the process using the combination of TFIDF and GNB results in an error). The combination of TFIDF+SVD+GNB give the worst result of all as the data in the experiments are not continuous. GNB relies on the assumption that continuous values associated with each class are distributed according to Gaussian Distribution.

The use of SVD in the research is not for improving accuracy, but as a trade-off of time consumed for processing by keeping the performance. However, as can be seen in Table 1, SVD does not perform well in the experiments despite having recommended by other previous works. From the experimental results, we obtained that features reduced by SVD seems to lose the unique value which represents the document, hence, reducing accuracy. SVD is supposed to make information extraction more computationally efficient, however it is not always performing well [20].

Other than the resulting precision and recall values obtained, the processing time also varied between the combinations, with also TFIDF+MNB giving the fastest time.

4. Conclusion

After we conduct the experiments, we conclude the following:

- TFIDF and MNB combination provides the best result for news classification in Indonesian language (98.4%), followed by TFIDF + Bernoulli Naïve Bayes (BNB) (98.2%). These results outperform the previous similar study which achieved 85% of accuracy [11].
- In terms of time consumed to process the data, MNB and BNB both gives the best result despite having very huge amount of data extracted by TF-IDF.
- The use of SVD in reducing features dimension is not always performing well. However, there is still a probability of it not performing well due to the data set used in the research, as the data is not yet pre-trained (the data is not yet validated by an expert, hence, the label given by the source could be inaccurate).

For further research, we plan to use a training data set in Indonesian language which has been tested by a language expert to make sure that the result is accurate. The use of more categories will also provide more convincing result.

References

1. Yang CC, Chen H, Hong K. Decision Support Systems. Visualization of large category map for Internet browsing. 2003;: p. 89-102.
2. Wong AH, Abednego L. Pengelompokan dokumen otomatis dengan menggunakan TFIDf classifier, naive bayes classifier dan KNN. http://library.unpar.ac.id/index.php?p=show_detail&id=204170#. 2015;: p. 1-120.
3. Aggarwal CC, Zhai C. A Survey of Text Classification Algorithms. 2013;: p. 169-170.
4. Zhang W, Yoshida T, Tang X. A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications. 2011 March; 38(3): p. 2758-2765.
5. Wang W, Carreira-Perpinan MA. The Role of Dimensionality Reduction in Classification. In Conference on Artificial Intelligence; 2015; California: AAAI. p. 2128-2134.
6. Zareapoor M, K.R. S. Feature Extraction or Feature Selection for Text Classification: A Case Study on Phising Email Detection . New Delhi:; 2015.
7. Trstenjak B, Mikac S, Donko D. KNN with TF-IDF based Framework for Text Categorization. Procedia Engineering. 2014; 69: p. 1356-1364.
8. Trivedi M, Sharma S, Soni N, Nair S. Comparison of Text Classification Algorithms. International Journal of Engineering Research & Technology. 2015 February; 4(2): p. 334-336.
9. Rennie JD, Shih L, Teevan J, Karger DR. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In Proceedings of the 20th International Conference on Machine Learning (ICML-03); 2003. p. 616-623.
10. Ramdass D, Seshasai S. Document Classification for Newspaper Articles. ; 2009.
11. Liliana DY, Hardianto A, Ridok M. Indonesian News Classification using Support. International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:5. 2011;: p. 1015-1018.
12. Sharma S. Web Crawler. International Journal of Advanced Research in Computer Science and Software Engineering. 2014 April; 4(4): p. 1379-1381.

13. Hs. W. Kalimat. In Hs. W. Bahasa Indonesia, Mata Kuliah Pengembangan Kepribadian di Perguruan Tinggi. Jakarta: Grasindo; 2012 . p. 146-147.
14. Plansangket S, Gan JQ. A query suggestion method combining TF-IDF and Jaccard Coefficient for interactive web search. Artificial Intelligence Research, Volume 4 No 2, ISSN 1927-6982. 2015;: p. 119-125.
15. Rajaraman A, Ullman JD, Leskovec J. Mining of Massive Datasets; 2011.
16. Vangelis M, Androutsopoulos I, Paliouras G. Spam Filtering with Naive Bayes – Which Naive Bayes? Athens:; 2006.
17. Davood Mahmoodi ASHKMT. FPGA Simulation of Linear and Nonlinear Support Vector Machine. Journal of Software Engineering and Applications. 2011 May; 4(5): p. 320-328.
18. Milgram J, Cheriet M, Sabourin R. One Against One or One Against All: Which One is Better for Handwriting Recognition with SVMs. In Tenth International Workshop on Frontiers in Handwriting Recognition; 2006; La Baule.
19. Raschka S. Naive Bayes and Text Classification: Introduction and Theory. Cornell University Library. 2014 October: p. 1 -20.
20. Gamallo P, Bordag S. Is singular value decomposition useful for word similarity extraction. Language Resources and Evaluation. 2011 May; 45(2): p. 95-119.
21. Juniawan I. Klasifikasi Dokumen Teks Berbahasa Indonesia Menggunakan Minor Component Analysis. http://repository.ipb.ac.id/bitstream/handle/123456789/13007/G09iju.pdf?sequence=9&isAllowed=y. 2009;: p. 1 -26.
22. Aggarwal CC, Zhai C. A Survey of Text Classification Algorithms. ;: p. 169-170.