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.

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
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

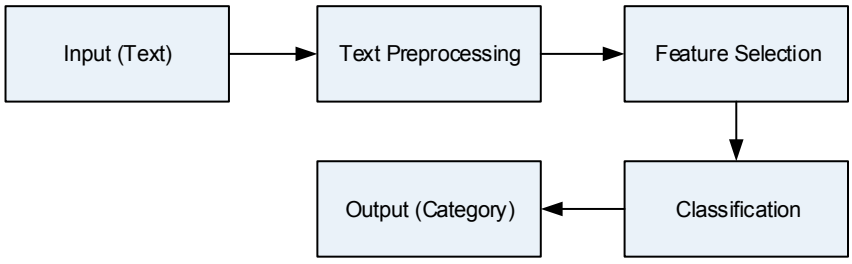

Fig. 1. Methodology

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.