Indonesian Journal of Electrical Engineering and Computer Science
Vol. 8, No. 3, December 2017, pp. 610 ~ 615
DOI: 10.11591/ijeecs.v8.i3.pp610-615

Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model

M. Ali Fauzi*[1], Agus Zainal Arifin[2], Sonny Christiano Gosaria[3], Isnan Suryo Prabowo[4]
[1]Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia
[2,3,4]Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
*Corresponding author, e-mail: moch.ali.fauzi@ub.ac.id

Abstract

Since the rise of WWW, information available online is growing rapidly. One of the example is Indonesian online news. Therefore, automatic text classification became very important task for information filtering. One of the major issue in text classification is its high dimensionality of feature space. Most of the features are irrelevant, noisy, and redundant, which may decline the accuracy of the system. Hence, feature selection is needed. Maximal Marginal Relevance for Feature Selection (MMR-FS) has been proven to be a good feature selection for text with many redundant features, but it has high computational complexity. In this paper, we propose a two-phased feature selection method. In the first phase, to lower the complexity of MMR-FS we utilize Information Gain first to reduce features. This reduced feature will be selected using MMR-FS in the second phase. The experiment result showed that our new method can reach the best accuracy by 86%. This new method could lower the complexity of MMR-FS but still retain its accuracy.

Keywords: *News Classification, Information Gain, Feature Selection, Maximal Marginal Relevance, Naïve Bayes*

1. Introduction

Since the rise of WWW, document available online is growing rapidly. One of the example is online news. In Indonesia, there are so many websites providing news form various categories. Users can access online news easily on the internet but it is difficult and time consuming to find what is really needed. Therefore, automatic news classification is needed to obtain relevant information quickly.

There are several works about news classification through statistical and supervised machine learning techniques including Naïve Bayes [1, 2], Maximum Entropy [3], Neural Network [4, 5], and Support Vector Machine (SVM) [6-9]. Specifically for Indonesian news classification, there several methods have been applied including Naïve Bayes [10], SVM [11], and single pass clustering [12]. The vast majority of the works use bag-of-word with TF-IDF term weighting as the features. All of the works focus on classification by topic and showed a promising result. However, there are still many challenges in this field.

Like many text classification problem, news classification problems major characteristics are high dimensional feature space and high level of feature redundancy. Even for a moderate-size document collection, we can have thousands of unique terms as feature using bag-of-word models. This high-dimensional features clearly contribute to the high computational complexity. It is desirable to reduce the dimension of features and select only the best features while maintaining classification accuracy, and it is also desirable to do that job automatically. Hence, automatic feature selection is needed.

Feature selection for text classification is a well-studied problem; its goals are improving classification effectiveness, computational efficiency, or both. Feature selection is a process of selecting a subset of the features or terms available for describing document before applying a learning algorithm [13]. Feature selection is necessary to reduce computational complexity with a little loss in classification quality. Even for many cases, feature selection can improve classification accuracy [14].

One of the most used feature selection methods in document classification is Information Gain (IG) [15]. IG has been used in text classification frequently and proven giving

Received August 29, 2017; Revised October 5, 2017; Accepted November 23, 2017

good results [14, 16]. Information Gain measures the decrease in entropy when the feature is given versus absent. IG is very good in cases where all features are not redundant. But in cases where many features are highly redundant with each other, we have to employ other technique such as more complex dependence models.

There some studies focused on feature redundancies. Lee et al [17] proposed a feature selection method called Maximal Marginal Relevance for Feature Selection (MMR-FS) to tackle this problem. This method was created to eliminate IG's weakness against redundancy between features in text classification by combining Maximal Marginal Relevance with Information Gain. The experiments results show that MMR-FS has better performance compared to IG for text classification. However, MMR-FS's complexity is high, increased quadratic time with respect to the number of features, since it has to compute pair wise IG (IGpair). We can tackle this problem by reducing some features before employ MMR-FS in two phased feature selection. In this study, we propose a novel two-phase feature selection method for Indonesian news classification. In the first phase, we utilize IG to reduce some features. In the next phase, MMR-FS is employed for feature selection. Then, we use Naïve Bayes to classify the news since this method is simple yet powerful for text classification.

2. Research Method

In general, as seen in Figure 1, the Indonesian news classification system in this study consists of three main stages, preprocessing, features selection and classification. The first stage involves several steps including tokenization, filtering, stemming and term weighting. This system will use bag-of-word model, the unique terms from preprocessing stage would be the original features of each documents. Through the feature selection stage, some of the best features were selected from the original feature set. In this study, we propose a novel two-phase feature selection method for Indonesian news classification. In the first phase, we utilize IG to reduce some features. In the next phase, MMR-FS is employed for feature selection. Finally, in the last stage, the news is classified using Naïve Bayes classifier.

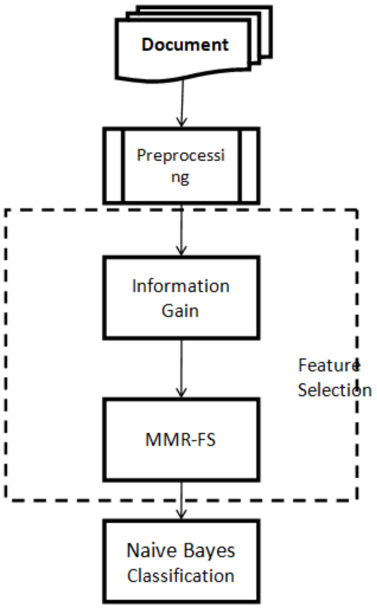

Figure 1. System Main Flowchart

2.1. Preprocessing

Preprocessing consists of tokenization, filtering, stemming and term weighting. In the tokenization process, conducted some process including case folding, changing all capital letters into lowercase, remove punctuation, numbers, and characters other than the alphabet

[18-21]. Filtering or stopwords removal is removing uninformative words based on the existing stoplist. In this case, we use stoplist by Tala [22]. Stemming is a process to convert every words to its root by removing affixes such as prefix, infix and suffix. In this case, we use Arifin-Setiono Stemmer [12].

2.2. Two-Phase Feature Selection Model

Feature selection is a process of selecting a subset of the features or terms available for describing document before applying a learning algorithm [13]. Feature selection is necessary to reduce computational complexity with a little loss in classification quality. Even for many cases, feature selection can improve classification accuracy [14]. Feature selection in this study is consisted of two phases. First phase is IG and the second phase is MMR-FS. The first phase is used to reduce the least correlated feature, so MMR-FS will be implemented to calculate fewer features. After the feature selection stage, documents will be classified using Naive Bayes classifier.

We used a different variation of IG in the first phase before the feature selection by MMR-FS. This IG is different from IG in MMR-FS where the importance of terms *w* only paired against every class *C*. The IG will calculate the importance of terms *w* to each existed document *d* and defined as follow:

$$IG(w_i; d) = P(w_i)\sum_k P(d_k|w_i)\log P(d_k|w_i) + P(\overline{w_\iota})\sum_k P(d_k|\overline{w_\iota})\log P(d|\overline{w_\iota}). \quad (1)$$

where *p(w_i)*

where $p(w_i)$ is the probability that word w_i appear, $\overline{w_\iota}$ means that word w_i does not occur, $p(d_k)$ is the probability of the *k*th document value, $p(d_k|w_{i,j})$ is the conditional probability of the document *k* value given that w_i appear, $p(w_{i,j})$ is the probability that w_i and w_j appear together, and $\bar{w}_{\mathrm{i,j}}$ means that w_i and w_j do not appear together but w_i or w_j can appear.

In information retrieval, marginal relevance is the linear combination of relevance and novelty. A document has high marginal relevance if it is both relevant to the query and contains minimal similarity to the previously selected documents. In document retrieval and summarization, marginal relevance should be maximized, hence the method is labeled as Maximal marginal relevance (MMR) [23].

$$MMR = arg \max_{D_i\in R\setminus S}\left[\lambda * sim_1(D_i;Q) - (1-\lambda)\max_{w_j\in S} sim_2(D_i|D_j)\right] \quad (2)$$

where C = {D1, . . . ,Di, . . .} is a document collection (or document stream); Q is a query or user profile; R = IR(C,Q,h), i.e., the ranked list of documents retrieved by an IR system, given C and Q and a relevance threshold h, S is the subset of documents in R which is already selected; R\S is the set difference, Sim1 is the similarity metric used in document retrieval and a relevance ranking between documents (passages) and a query; and Sim2 can be the same as Sim1 or a different metric. Given the above definition, MMR computes incrementally the standard relevance-ranked list when the parameter k = 1, and computes a maximal diversity ranking among the documents in *R* when *k* = 0. Intermediate values of *k* in the interval [0, 1], a linear combination of both criteria should be optimized. MMR-FS is defined as follows:

$$MMR_{FS} = arg\ \max_{D_i\in R\setminus S}\left[\lambda * IG(w_i;C) - (1-\lambda)\max_{w_j\in S} IGpair(w_i;w_j|C)\right], \quad (3)$$

where *C* is the set of class labels, *R* is the set of candidate features, *S* is the subset of features in *R* which was already selected, *R\S* is the set difference, i.e. the set of as yet unselected features in *R*, *IG* is the information gain scores, and *IGpair* is the information gain scores of co-occurrence of the word (feature) pairs. IG and *IGpair* were defined as follow:

$$IG(w_i; C) = P(w_i)\sum_k P(C_k|w_i)\log P(C_k|w_i) + P(\overline{w_\iota})\sum_k P(C_k|\overline{w_\iota})\log P(C_k|\overline{w_\iota}) \quad (4)$$

$$\begin{aligned} IGpair(w_i; w_j|C) \\ &= P(w_{i,j}) \sum_k P(C_k|w_{i,j}) \log P(C_k|w_{i,j}) \\ &+ p(\overline{w_{\iota,\jmath}}) \sum_k P(C_k|\overline{w_{\iota,\jmath}}) \log P(C_k|\overline{w_{\iota,\jmath}}), \end{aligned} \quad (5)$$

where *p(w$_i$)* is the probability that word w_i appear, $\overline{w_\iota}$ means that word w_i does not occur, *p(C$_k$)* is the probability of the class *k* value, $p(C_k|w_{i,j})$ is the conditional probability of the class *k* value given that w_i appear, *p(w$_{i,j}$)* is the probability that w_i and w_j appear together, and $\overline{w}_{\mathrm{i,j}}$ me means that w_i and w_j do not appear together but w_i or w_j can appear [15].

MMR-FS has complexity of $\frac{1}{2}n(n-1)$, with *n* is the number of features. In order to reduce the computation cost, we decided to apply IG before implementing MMR-FS to reduce the features. IG will show the relation between word and document with some of the most correlated will be used in MMR-FS. MMR-FS will only processed *m* features with *m<n.* Total complexity of this method is complexity of IG and MMR-FS combined which is $n + \frac{1}{2}m(m-1)$.

2.3. Naïve Bayes Classifier

Naïve Bayes Classifier is based on Bayesian theorem. This method is simple and efficient, but it can surpass more complex classification methods [24]. A good feature selection must be used because Naïve Bayes Classifier is highly sensitive to feature selection. Good feature selection will improve the precision of this classifier.

In this stage, the classification with Naïve Bayes method was implemented to separate category or topic automatically. In the area of text classification there are two different models of Naïve Bayes classifiers in common use: the Multi-Variant Bernoulli Event Model and the Multinomial Event Model [25]. Both of these two models use the Bayes rule to classify a document. Given a document d_i; the probability of each class C_j is calculated as

$$P(C_j|d_i) = \frac{P(d_i|C_j) * P(C_j)}{P(d_i)} \quad (6)$$

As $P(d_i)$ is the same for all class, then label d_i; the class label of d_i; can be determined by

$$label\ (d_i) = \arg\max\{P(C_j|d_i)\}. \quad (7)$$

In this study, we use multinomial event model. A document is regarded as a bag of words in Multinomial event model. The order of words is not considered but the frequency of each word in the document is counted. In this model, a similar Naive Bayes assumption is made. Denote the number of times word w_k occurs in document d_i as n_{ik}. The probability $P(d_i|c_j)$ can be computed by:

$$P(d_i|C_j) = P(|d_i|)|d_i|! \prod_{k=1}^{|V|} \frac{P(w_k|C_j)^{n_{ik}}}{n_{ik}!}, \quad (8)$$

where $|d_i|$ as the number of document d_i.

3. Results and Analysis

Experiment conducted by using 300 news taken from Kompas.com where every news is in Indonesian language. There are 250 news used as training set and the other news would be used as test set for Naïve Bayes Classifier. The database has 5 categories in this experiment; they are ‘International’, ‘Sport’, ‘Economy and Business’, ‘Travelling’, and ‘Science and Technology’. The same number of documents taken from each category for training set. However a random number from each category used for testing set.

In the experiment, our method will be tested against IG and MMR-FS. Features will be reduced to 10, 15, 20, 30, 50, 100, 150, 200, 300, 400, 500, and 600 features. The tuning for λ values for MMR-FS and our method is done by using 11 different λ values (i.e. 0, 0.1, 0.2, . . 1). The tuning determines best λ value to be used in experiment which is 0.6.

The experiment conducted to calculate the accuracy of our implemented method. Figure 2 show that all of the three selection feature algorithm can increase the accuarcy of Naive Bayes. In classification with Naïve Bayes without feature selection the accuracy is 44%. All of the three algorithm reach their optimal accuracy on 500 features. Our methods have higher accuracy under 200 features, meanwhile MMR_FS always improved significantly until reach its optimal on 500 features.

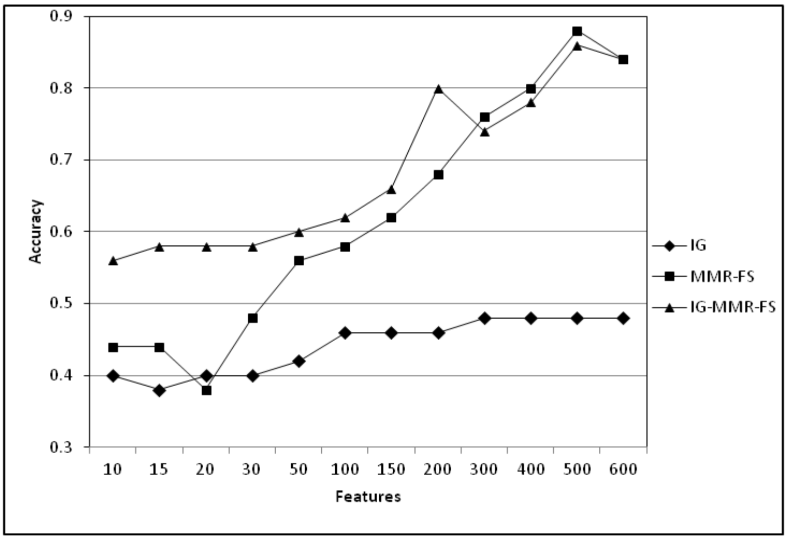

Figure 2. Accuracy comparison between IG, MMR-FS and IG-MMR-FS

In the experiment, each method starting with low accuracy when the features numbers are lows, but their accuracy began to rise as the number of features increasing. Our proposed method managed to achieve higher accuracy than MMR-FS when used in less than 200 features as seen in Fig 2. However its accuracy began to be surpassed by MMR-FS when 300 or more features are used.

MMR-FS is not feasible to implement on large text database because of its computational time and cost. In the conducted experiment, the proposed method for 50 documents and 3500 features finished faster than MMR-FS. This problem occur because the complexity of MMR-FS which is 1/2 n(n-1). The more features it need to compute, the time and cost consumed to compute MMR-FS is growing in quadratic rule.

Our proposed method reduced the computational cost for MMR-FS at the expense of precision. The complexity of our proposed method is n+1/2 m(m-1) where m is the number of features that left behind after feature reduction by IG. The proposed method complexity is quadratic but it will calculate fewer features than MMR-FS.

4. Conclusion

In this study, a two-phase feature selection method is proposed. First phase is IG and the second phase is MMR-FS. The first phase is used to reduce the least correlated feature, so MMR-FS will be implemented to calculate fewer features. After the feature selection stage, documents will be classified using Naive Bayes classifier. Based on the experiment, our proposed method tend to have higher than standalone IG or MMr-FS. Our methods have higher accuracy under 200 features, meanwhile MMR_FS always improved significantly until reach its optimal on 500 features. The experiment result showed that our new method can reach the best accuracy by 86%. The two-phase feature selection method is managed to reduce the computation cost and time, but still manage to retain classification quality.

References

[1] Chy AN, Seddiqui MH, Das S. *Bangla news classification using naive Bayes classifier*. In Computer and Information Technology (ICCIT), 2013 16th International Conference. IEEE. 2014; pp. 366-371.

[2] Li-guo D, Peng D, Ai-ping L. A new naive Bayes text classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*. 2014; 12(2): 947-952.

[3] Sawaf H, Zaplo J, Ney H. *Statistical classification methods for Arabic news articles*. Natural Language Processing in ACL2001, Toulouse, France. 2001 Jul 6.

[4] Selamat A, Yanagimoto H, Omatu S. *Web news classification using neural networks based on PCA*. In SICE 2002. Proceedings of the 41st SICE Annual Conference. IEEE. 2002 Aug 5; 4: 2389-2394.

[5] Chen Y, Xu L. Based on Weighted Gauss-Newton Neural Network Algorithm for Uneven Forestry Information Text Classification. *Indonesian Journal of Electrical Engineering and Computer Science*. 2014 May 1; 12(5): 4091-4100.

[6] Manevitz LM, Yousef M. One-class SVMs for document classification. *Journal of Machine Learning Research*. 2001; 2(Dec):139-54.

[7] Cui L, Meng F, Shi Y, Li M, Liu A. *A hierarchy method based on LDA and SVM for news classification*. In Data Mining Workshop (ICDMW), 2014 IEEE International Conference on. IEEE. 2014 Dec; 14: 60-64.

[8] Kumar RB, Kumar BS, Prasad CS. Financial news classification using SVM. *International Journal of Scientific and Research Publications*. 2012 Mar;2(3):1-6.

[9] Krishnalal G, Rengarajan SB, Srinivasagan KG. A new text mining approach based on HMM-SVM for web news classification. *International Journal of Computer Applications*. 2010 Feb 25;1(19):98-104.

[10] Asy'arie AD, Pribadi AW. *Automatic news articles classification in indonesian language by using naive bayes classifier method.* In Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services 2009 Dec 14 (pp. 658-662). ACM.

[11] Liliana DY, Hardianto A, Ridok M. Indonesian news classification using support vector machine. *World Academy of Science, Engineering and Technology*. 2011 Sep 21;57:767-70.

[12] Arifin AZ, Setiono AN. *Klasifikasi dokumen berita kejadian berbahasa indonesia dengan algoritma single pass clustering.* In Prosiding Seminar on Intelligent Technology and its Applications (SITIA), Teknik Elektro, Institut Teknologi Sepuluh Nopember Surabaya. 2002.

[13] Dasgupta A, Drineas P, Harb B, Josifovski V, Mahoney MW. *Feature selection methods for text classification*. InProceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining 2007 Aug 12 (pp. 230-239). ACM.

[14] Forman G. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*. 2003;3(Mar):1289-305.

[15] Uğuz H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems*. 2011 Oct 31;24(7):1024-32.

[16] Joachims T. *Text categorization with support vector machines: Learning with many relevant features. Machine learning*: ECML-98. 1998:137-42.

[17] Lee C, Lee GG. Information gain and divergence-based feature selection for machine learning-based text categorization. *Information processing & management.* 2006 Jan 31;42(1):155-65.

[18] Fauzi MA, Arifin AZ, Yuniarti A. Arabic Book Retrieval using Class and Book Index Based Term Weighting. *International Journal of Electrical and Computer Engineering (IJECE)*. 2017 Dec 1;7(6).

[19] Pramukantoro ES, Fauzi MA. *Comparative analysis of string similarity and corpus-based similarity for automatic essay scoring system on e-learning gamification*. InAdvanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on 2016 Oct 15 (pp. 149-155). IEEE.

[20] Fauzi MA, Arifin A, Yuniarti A. Term Weighting Berbasis Indeks Buku dan Kelas untuk Perangkingan Dokumen Berbahasa Arab. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*. 2013;5(2).

[21] Fauzi, M.A., Utomo, D.C., Setiawan, B.D. and Pramukantoro, E.S., 2017, August. *Automatic Essay Scoring System Using N-Gram and Cosine Similarity for Gamification Based E-Learning*. In Proceedings of the International Conference on Advances in Image Processing (pp. 151-155). ACM.

[22] Tala FZ. *A study of stemming effects on information retrieval in Bahasa Indonesia. Institute for Logic, Language and Computation*, Universiteit van Amsterdam, The Netherlands. 2003 Jul.

[23] Carbonell J, Goldstein J. *The use of MMR, diversity-based reranking for reordering documents and producing summaries*. InProceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval 1998 Aug 1 (pp. 335-336). ACM.

[24] Chen J, Huang H, Tian S, Qu Y. Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*. 2009 Apr 30; 36(3):5432-5.

[25] McCallum A, Nigam K. *A comparison of event models for naive bayes text classification*. InAAAI-98 workshop on learning for text categorization 1998 Jul 26 (Vol. 752, pp. 41-48).