ISSN: 2477-5126
e-ISSN: 2548-9356

A Comparison of Text Classification Methods k-NN, Naïve Bayes, and Support Vector Machine for News Classification

Fanny[1*)], Yohan Muliono[2], Fidelson Tanzil[3]
[1,2,3]School of Computer Science, Bina Nusantara University, Jakarta
[1,2,3]Jl. Kh. Syahdan No.9, RT.6/RW.12, Palmerah, Jakarta Barat, DKI Jakarta 11480, Indonesia
email: [1]fanny.sa@binus.edu, [2]ymuliono@binus.edu, [3]fitanzil@binus.edu

Received: 9 April 2018; Revised: 7 Mei 2018; Accepted: 13 Mei 2018

***Abstract* – In this era, a rapid thriving Internet occasionally complicates users to retrieve news category furthermore if there are plentiful of news to be categorized. News categorization is a technique can be used to retrieve a category of news which gives easiness for users. Internet has vast amounts of information especially at news. Therefore, accurate and speedy access is becoming ever more difficult. This paper compares a news categorization using *k*-Nearest Neighbor, Naive Bayes and Support Vector Machine. Using vary of variables and through a several steps of preprocessing which proving k-Nearest Neighbor is producing a capable accuracy competes with Support Vector Machine whereas Naive Bayes producing just an average result, not as good as *k*-Nearest Neighbor and Support Vector Machine yet as bad as *k*-Nearest Neighbor and Support Vector Machine ever reach. As the results, *k*-Nearest Neighbor using correlation measurement type produces the best result of this experiment.**

***Abstrak* – Pada zaman sekarang, perkembangan internet yang begitu pesat kadang-kadang mempersulit pengkategorian berita apalagi jika ada banyak berita yang harus dikategorikan. Kategorisasi berita adalah teknik yang bisa digunakan untuk mengambil kategori berita yang memberi kemudahan bagi pembaca berita. Internet memiliki sejumlah besar informasi terutama di berita. Oleh karena itu akses yang akurat dan cepat menjadi semakin sulit. Makalah ini mengulas kategorisasi berita dengan menggunakan *k-Nearest Neighbor, Naive Bayes* dan *Support Vector Machine*. Penggunaan berbagai variabel dan melalui beberapa tahap *preprocessing* yang membuktikan *k-Nearest Neighbor* menghasilkan akurasi yang mampu bersaing dengan *Support Vector Machine* sedangkan *Naive Bayes* hanya menghasilkan hasil rata-rata, tidak sebagus *k-Nearest Neighbor* dan *Support Vector Machine* namun juga tidak memberikan hasil yang seburuk yang dihasilkan dengan teknik *k-Nearest Neighbor* dan *Support Vector Machine*. Sebagai hasil yang didapatkan, *k-Nearest Neighbor* dengan tipe pengukuran *correlation* memberikan hasil yang terbaik dan konsisten selama eksperimen dengan parameter dan validasi data yang berbeda-beda.**

***Kata Kunci* – news classification, news, information retrieval, text classification.**

*) **Corresponding author:** Fanny
Email: fanny.sa@binus.edu

I. INTRODUCTION

Along with the currently issue about big data and the rapid development of internet, information retrieval and text mining has become a popular research field in the world. Especially in the field of data or text classification [1].

Study on text classification abroad dated back to the late 1957, [2] done some research works and propose a text classification using word frequency method.

The first text classification research has been held in 1960 [3], then the other expansion of text classification has been done in many area like text information retrieval, electronic meetings, and text filtering [1].

This research conducts an experiment to compares three most popular methods of text classification: *k*-Nearest Neighbor, Naïve Bayes, and Support Vector Machine in News Categorization to classify the category of news in English using several pre-processing texts that will be explained in II. A several experiments will be conducts in this research involving 6 news categories: Entertainment, Health, Sports, Economy, Politics and Technology, with 90 samples of each category cited from Fox news, New York Times, ABC news, and BBC. The main purpose of this research is to find a best method to categorize a news so that the text can be categorized without thoroughly read a full text.

II. LITERATURE STUDIES

A. Classification

In classification there are two general problems of knowledge, the first one is there already an underlying joint distribution complete statistical knowledge of observation x and the true category θ or there is no knowledge of the underlying distribution except that which can be inferred from samples [4]. For the first problem, a standard Bayes analysis can be used to provides an optimal decision, in another problem, Nearest Neighbor has been proven good to solve the problem in several ways [4-6].

One of classification problem that necessary to discuss is text classification. In information retrieval and text mining, classification is an important foundation. The main task is to

assign a document to the predefined categories according its content and the training samples [7]. Text classification has been used significantly, like for email filtering in government departments or companies, spam mail detection [8], knowledge enrichment [9], and in news articles classification [10]. The email filtering can spread out emails to the corresponding departments according to the content, not only for filtering junk mails. Text classification also used in search engine in the website for filtering the content that users concerned or not and interested or not as well [1]. The processes of indexing is start by reads a received document, chooses one or several of the index terms from the "corpus" and then maps the selected terms with the given categories, the assignment of terms to each document is either it applies to the document or categories in question or not [3].

B. Classifiers

In classification there are two general problems of knowledge, the first one is the traditional *k-Nearest Neighbor* text classification. For classification, this method used all the training data sets which make the training process calculation become more complex and can not shows the different results from the different samples [11].

The next method is Naive Bayes classifier, which is known as a simple Bayesian classification algorithm. This method is still used for text classification because it is fast and easy to implement [12-13]. These probabilistic approach makes a tough assumption about the process of generating the data, and which uses a collection of labeled training samples to estimate the parameters of generative models [14].

The third method is Support Vector Machine (SVM), which are based on structural risk minimization [15]. The idea is to find which hypothesis h that has the smallest true error. An upper bound can be used. For connecting the error of a hypothesis h with the error of hypothesis on the training set and the complexity, we can use the upper bound. When learning about text classifiers, it will deal with large feature (more than 10000 dimensions) [16]. SVM is recommended to using for handling large feature of data sets.

III. METHODOLOGY

The methodology in classify the documents (news) divides into two main processes. Those are processing document and validation for classifying. Before validation process, the document has been processed before to ease the classification by removing some unimportant words org characters. There are four steps in process the documents, as follows:

Step 1: Tokenization. In tokenization process, the non-letters characters are removed. The purpose of this process is to remove the character that not necessary for the information retrieval.

Step 2: Filtering stopwords. In filtering stopwords, the stopwords character is removed. It is because the stopwords will not be useful for information retrieval from document.

Step 3: Filtering tokens (by length). In this step, the word that out of the range that has been set before will be remove. In this step, the minimum and maximum length of character is set. The purpose of removing these characters is to reducing processing times and some unimportant words.

Step 4: Stemming. The purpose of stemming is to change affixes words to its root words, For Example 'driving', 'drives' have a same meaning with 'drive' so the affixes '-ing','-s' will be removed and changed according to the words and stemmer rules.

After the document has been proceed, the next process is to produce the contents by transforming a full text documents to a document vectors. This process is prepared to decrease the complexity of text documents and make the handling process easier. The transformation process is by using a Bag of Words representation assuming that each feature is a single token [17]. The bag of words will contain a term and term frequency with its categories. Term Frequency is the number of how many times the x words appear in category y [18]. A word can be described as a term if a word appears many time in 1 classification documents, if a word appears in many categories of documents, the word will not be counted as a term.

The next process is validation. Validation is the main process of this paper objectives, where in validation, news classification is obtained. The classification process covers the process of splitting the data for training and testing, classifier selection, and the result of classification.

IV. EXPERIMENT AND RESULT

For the experiment, this paper uses the documents that collected from some news like ABC news, Fox news, New York Times, and BBC news like has been explained in section I. The documents are divide into 6 categories: entertainment, health, economy/business, technology, sports, and politic with 30 total data (documents) for each category. This paper does three times of experiment. The parameter for process the documents for those three experiments is shown in table I.

TABLE I
PARAMETER FOR DOCUMENT PROCESSING

Parameter	Mode Uses
Tokenization	Non-letters mode
Filtering stopwords	English standard
Filtering tokens (by length)	Minimum 4 characters, maximum 25 characters
Stemming	Lovins, porter, and wordNet

TABLE III
SPLIT VALIDATION FOR EXPERIMENT

Experiment	Total Data per Category	Data for Training	Data for Testing
I	30	70%	30%
II	30	70%	30%
III	30	80%	20%

After the document processing, the next process is validation. The first step is splitting the data for training data and testing data. This division shown in table II. The data distribution for learning and testing will not be random condition. In that case, every time the experiments will produce same results, if the data used for the experiments is identical. First and second experiment use 70% of total data for training and 30% of total data for testing. Same treatment will be done to experiment 2. And for the last experiment, data for training will be increased to 80% to see the convergence of the result after two times doing experiments with using different data.

The last step in validation is classification process. The parameter used for experiment in classification is shown in table III and by using that parameter, the experiment run, and the result is obtained. The first experiment result shown in table IV.

TABLE IIIII
CLASSIFIER FOR CLASSIFICATION

Classifier	Type
k-NN	Euclidean Distance
	Cosine Similarity
	Correlation Similarity
Naïve Bayes	Estimation Mode: Full
SVM	RBF
	Linear
	Sigmoid

TABLE IVV
EXPERIMENT #I RESULT

Classifier		Stemmer		
		Porter	**Lovin**	**Wordnet**
k-NN	Euclidean	17.86%	17.86%	17.86%
	Cosine	75%	78.57%	80.36%
	Correlation	82.14%	82.14%	76.79%
Naïve Bayes		64.29%	66.07%	62.50%
SVM	RBF	64.29%	64.29%	62.50%
	Linear	64.29%	64.29%	62.50%
	Sigmoid	64.29%	64.29%	62.50%

From the result of first experiment, it shows some interesting points. By using a different stemmer, the result of classification will be different. For *k*-NN, the result is surprisingly bad using Euclidean measurement type, but for the cosine and correlation measurement type the result is good. For the Naïve Bayes classifier, the result is in average, not as good as *k*-NN using correlation measurement and not as bad as *k*-NN using euclidean. And for SVM, the result is average like Naïve Bayes. As the result of this experiment, the highest accurate classifier result is *k*-NN using porter and lovin stemmer with correlation measurement type. To make sure the accuracy of result, this research conducts a second experiment using different data but with same treatment. The result of the second experiment shown in table V.

TABLE V
EXPERIMENT #II RESULT

Classifier		Stemmer		
		Porter	**Lovin**	**Wordnet**
k-NN	Euclidean	84.44%	82.22%	88.89%
	Cosine	84.44%	82.22%	88.89%
	Correlation	84.44%	80.00%	93.33%
Naïve Bayes		64.29%	64.29%	71.11%
SVM	RBF	15.56%	15.56%	82.22%
	Linear	84.44%	77.78%	82.22%
	Sigmoid	84.44%	77.78%	82.12%

TABLE VI
EXPERIMENT #III RESULT

Classifier		Stemmer		
		Porter	**Lovin**	**Wordnet**
k-NN	Euclidean	86.11%	83.33%	86.11%
	Cosine	83.33%	83.33%	86.11%
	Correlation	86.11%	83.33%	86.11%
Naïve Bayes		64.29%	64.29%	77.78%
SVM	RBF	5.56%	5.56%	5.56%
	Linear	80.56%	83.33%	86.11%
	Sigmoid	80.56%	83.33%	86.11%

Unlike the first experiment held before, this experiment shows that *k*-NN is surprisingly good regardless of the variables used. Naïve Bayes still doing like before a good result but not as well as *k*-NN. However, SVM in this experiment shows a bad result using RBF regardless of stemmer used, but SVM also produce a good result which can equally good as compared to *k*-NN. And the best result goes to *k*-NN and SVM which *k*-NN using porter stemmer with euclidean, cosine, and correlation measurement type and SVM using porter stemmer in linear and sigmoid kernel type.

For the last experiment, different data will be used as well, and the training rate of data will be increased to 80% and the testing will be 20% to check whether anything different if the testing data increased. The result of this last experiment shown in table VI.

Similar like the first and second experiment, the result of *k*-NN is good regardless of the variables and stemmer. Naïve Bayes still perform a good classification but not as well as *k*-NN. The interesting points here is SVM once again, produce a bad result in RBF kernel type. And the highest accurate classifier still goes to *k*-NN and SVM like the other experiments with porter and wordNet using euclidean and correlation measurement type for *k*-NN and wordNet using linear and sigmoid kernel type for SVM.

V. CONCLUSION

From three experiments has been held, Naïve Bayes shows a stable result, Naïve Bayes Classifier never been the worst nor the best, shows a stable performance, appropriate to a low risk taker for use Naïve Bayes for classifier and a big statistical learning data for naïve bayes, without expecting the best result of it, for k-NN it surprisingly good for a few training data and nearly produce an accurate result for three time. Although k-NN ever get a worst result regardless of that, k-NN is being the best for three times also. Whereas SVM Classifier still performs a random result can produce equally well as k-NN but sometimes the performance drops.

For the future works, a similar experiment can be held, using different variables and classifiers. Keep using k-NN and SVM as a comparison. Naïve Bayes already shows an average performance, different classifiers should be replacing Naïve Bayes. Different categories of news can show a different result also.

REFERENCES

[1] Z. Yong, L. Youwen, and X. Shixiong, "An Improved KNN Text Classification Algorithm Based on Clustering," J. Comput., vol. 4, no. 3, pp. 230–237, 2009.

[2] H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM J. Res. Dev., vol. 1, no. 4, pp. 309–317, 1957.

[3] M. E. Maron and J. L. Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval," J. ACM, vol. 7, no. 3, pp. 216–244, 1960.

[4] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, 1967.

[5] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and a. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 881–892, 2002.

[6] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor-based algorithm for multi-label classification," vol. 2, pp. 718 – 721 Vol. 2, 2005.

[7] X. X. Su Jinshu, Zhang Bofeng, "Advances in Machine Learning Based Text Categorization," J. Chem. Inf. Model., vol. 53, pp. 1689–1699, 2013.

[8] D. Sharma, "Experimental Analysis of KNN with Naive Bayes, SVM and Naive Bayes Algorithms for Spam Mail Detection," vol. 8491, no. 4, pp. 225–228, 2016.

[9] S. Hassan, M. Rafi, and M. S. Shaikh, "Comparing SVM and Naïve Bayes classifiers for text categorization with Wikitology as knowledge enrichment," Proc. 14th IEEE Int. Multitopic Conf. 2011, INMIC 2011, pp. 31–34, 2011.

[10] L. Pradhan, N. A. Taneja, C. Dixit, and M. Suhag, "Comparison of Text Classifiers on News Articles," Int. Res. J. Eng. Technol., vol. 4, no. 3, pp. 2513–2517, 2017.

[11] S. Tan, "An effective refinement strategy for KNN text classifier," Expert Syst. Appl., vol. 30, no. 2, pp. 290–298, 2006.

[12] D. D. Lewis, "Representation and learning in information retrieval," vol. 7, 1992.

[13] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," no. 1973, 2003.

[14] A. Mccallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classi cation," 1997.

[15] V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer, 1995.

[16] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," pp. 2–7.

[17] Y. Zhang, R. Jin, and Z. H. Zhou, "Understanding bag-of-words model: A statistical framework," Int. J. Mach. Learn. Cybern., vol. 1, no. 1–4, pp. 43–52, 2010.

[18] a. Selamat, H. Yanagimoto, and S. Omatu, "Web news classification using neural networks based on PCA," Proc. 41st SICE Annu. Conf. SICE 2002., vol. 4, pp. 2389–2394, 2002.