Using Summarization to Optimize Text Classification

K E Dewi* and R E Sagala

Informatics Engineering Department, Universitas Komputer Indonesia, Jl. Dipatiukur 112-116 Bandung, Indonesia

*kania.evita.dewi@email.unikom.ac.id

Abstract. This study demonstrates the impact of the summary on document classification. The document used is 100 background thesis problems, formatted .txt. Each document will be through preprocessing (case folding, splitting sentence, splitting word, filtering, stop word removal and TF-IDF). Then the document is summarized using the method of extraction. Then all the documents are classified using the FKNNC method. The results of this study were obtained by compiling the classification process to be faster ie for 5 minutes, this is due to the reduced extraction features used. Therefore, the summary can be used to reduce the features in the classification.

1. Introduction

Text classification is the grouping of text based on the similarity of the characteristics of a document with another document [1]. The simplest method of classifying text is K- nearest neighbour or often called KNN [2,3]. The use of word appearance as an extraction feature in classification is often used [4,5]. But the number of words used in the classification would be a problem when the number of documents used is quite large [4]. The number of words used as a feature for a text will be directly proportional to the number of documents used in the classification process [6]. Summarization is a process to derive the essence of a paragraph [7]. With the summary, the sentence used to describe the topic of a paragraph will be reduced, so it is expected to reduce the number of words.

Research on text classification has been done with various methods. One of them in Korde research [8] can be seen that KNN and SVM are better in classification than Naïve Bayes. The KNN method has improved a lot, one of which is in Praetyo research [9], FK-NNC is an improvement of KNN method, it is shown in his research that FK-NNC is better than KNN. In their research for document extraction feature is the word appearance. TF-IDF is the weighting often used in document processing, such as sentiment analysis [10], summary [11], and document classification [12]. As in Tstenjak's research, it is said that with the large amount of data, the time needed to process the classification will increase [5]. Someone has tried with the selection feature to reduce the feature [13], but not using the summary. Summarization is the process of choosing the core sentence [7]. In Indriani [6] research, it is shown that summarization can help reduce extraction features. Only in this study, using SVM method.

Therefore, in this study we will examine the impact of summary extraction by using TF-IDF weighting on text classification using FK-NNC method. Testing is done by calculating the accuracy and timing of classification using the FK-NNC method to the document performed by summary. From the examination, the classification of the document after being summarized becomes faster, because the used extract feature has been reduced.
2. Methods

Document to be used in this research is 100 background problem where 80 data is used as trainer data, the rest 20 data is used as test data. The text format used is .txt. Before entering the classification process, each document will enter into the summary process in order to reduce the word used as an extracted document, without losing the essence of the document.

In general, the process of classifying the text in this research will begin with the pre-process, here every document whether the training data or test data will be done through case folding stage, splitting sentence, splitting word, filtering, stop word removal. After each document through the pre-process stage, each document will be counted the weight of words per sentence in a single document. The weighting used is the Term frequency-inverse document frequency (TF-IDF). This weighting is often used [14] and is considered better than Local-Global-Normalization in the summary process [15]. If in a d-document there is the word $t_i$ as much as TF, then the calculation of the weight of the word $t_i$ in the d document is $W_i$ as in formula 1

$$W_i = TF(t_i, d) \times IDF(t_i)$$

Where $IDF(t_i)$ is Inverse Document Frequency of the i-th word $t_i$. If in a study using D documents and there are as many DF documents containing the word $t_i$, then the $IDF(t_i)$ can be calculated using the formula 2

$$IDF(t_i) = \log \left( \frac{D}{TF(t_i)} \right)$$

Each weighting result is summed up by a sentence, then selected as much as 50% as a sentence. The summary results will be entered into the classification method which in this research will use FK-NNC method. The FK-NNC calculations are based on Prasetyo’s research results [9]. Each test data will be calculated first closer to K nearest neighbor within each class. For the calculation of proximity distance $\bar{x}_i$ with $\bar{x}_j$, where $\bar{x}_i, \bar{x}_j \in R^N$ will be used euclidean distance, as in the following formula 3:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{N} |x_{il} - x_{jl}|^2}$$

The results of calculating the distance of the test data with the training data in each group, then sorted. Sorting starts from the smallest to the largest in each class. Then distance test data to all neighboring K of each class-j summed using the formula 4.

$$S_{ij} = \sum_{r=1}^{K} d(x_i, x_r)$$

The sum of the distance of each class then combined into one, by the formula 5.

$$D_i = \sum_{l=1}^{K} (S_{ij})$$

Where m is the weight rank of the Fuzzy K-Nearest Neighbor, m > 1, in this research is used m = 3. From the result of the distance of each class $S_{ij}$ divided by the sum of the distance of each class $D_i$, in each class-j, the calculation can be seen in the formula 6

$$u_{ij} = \frac{S_{ij}}{D_i}$$

So to get a prediction class for test data, done by selecting the largest membership value. Class prediction can be done with the formula 7.
\[ y' = \arg \max_j \left( u_{ij} \right) \]  

(7)

3. Results and Discussion

In this study, the impact of the summary on the classification of documents viewed from two sides, namely time and accuracy. Classification using a document that has been summarized compared with unbranched documents can be done more quickly. The average difference of speed required up to 5 minutes. The number of words used before being summarized is 4593, after summarizing the word used in the classification of 3908. The results of this study are in accordance with the study [5], which states the more data used, the longer the processing.

In this research used 100 background documents of the problem, which is divided into 5 groups of scholars. The five scientific groups are Information Systems (A), Software Engineering (C), Multimedia (C), Mobile Technology (D), and Computer Science (E). Each group was taken 20 randomly titles and from 20 titles from each group were taken 4 as test data. In the calculation of accuracy used confusion matrix. Table 1 shows the results of classification using KNN method with K = 43 while in table 2 it is shown the confusion matrix using FK-NNC by using K = 7. Can be seen in table 1, good test data from various groups, test results show that 75% of test data are classified as group B, while in table 2, test data 75% test data classified as group E. This is caused, in every field of science has many incidents. This is the contrary to the assumption of classifiers used in the classification process that each group is independent, so that each object has exactly a group [1]. (See Table 1,2).

Table 1. Result of Confusion Matrix of KNN Method with K = 43.

| Predicted class | A | B | C | D | E | Accuracy |
|-----------------|---|---|---|---|---|----------|
| A               | 1 | 3 |   |   |   | 25%      |
| B               | 3 | 4 |   |   |   | 100%     |

Actual Class

| C | 4 | 0% |
| D | 4 | 0% |
| E | 4 | 0% |

Table 2. Result of Confusion Matrix of FK-NNC Method with K = 7.

| Predicted Class | A | B | C | D | E | Accuracy |
|-----------------|---|---|---|---|---|----------|
| A               |   |   | 4 | 0%|   |          |
| B               |   |   |   | 4 | 0%|          |

Actual Class

| C | 1 | 3 | 0% |
| D | 1 | 2 | 25%|
| E | 4 | 100%| 100%|

The results of the total accuracy calculations of this study are still far from expected, unlike in Indriani's research [6] who do the same thing only using SVM method gets 77% and Prasetyo [9] using FK-NNC for classification gets 97% greatest accuracy. However, in this study it is sufficient to show the effect of the summary on the classification. Can be seen in table 3 and table 4, that the accuracy of classification using both KNN and FK-NNC better than without summary.
Table 3. Accuracy of KNN method.

| K=   | By summary (%) | Without Summary (%) |
|------|---------------|---------------------|
| 21   | 25            | 20                  |
| 22   | 20            | 20                  |
| 23   | 20            | 20                  |
| 31   | 20            | 20                  |
| 32   | 20            | 20                  |
| 33   | 20            | 20                  |
| 41   | 25            | 20                  |
| 42   | 30            | 20                  |
| 43   | 25            | 20                  |
| 51   | 20            | 20                  |
| 52   | 20            | 20                  |
| 53   | 20            | 20                  |

Table 4. Accuracy of FK-NN method.

| K=   | By summary (%) | Without Summary (%) |
|------|---------------|---------------------|
| 2    | 20            | 15                  |
| 3    | 25            | 15                  |
| 4    | 20            | 20                  |
| 5    | 25            | 30                  |
| 6    | 25            | 25                  |
| 7    | 25            | 20                  |
| 8    | 25            | 25                  |
| 9    | 30            | 20                  |
| 10   | 25            | 20                  |

4. Conclusions

The conclusion of this study is that the summary can be used to optimize the classification, this can be seen by the decreasing time for the classification of the text. However, for data such as thesis preliminary as in this study, cannot use term frequent feature. This can be seen from the results of small accuracy, this happens because with only the use of term frequent, each group will be intercepted, which is contrary to the assumption of the classification. Therefore, the data as used in this study requires features other than term frequent to be classified.

References

[1] Bishop C M 2006 Pattern Recognition and Machine Learning New York: Springer
[2] Budiharto W 2016 Machine Learning & Computational Intelligence Yogyakarta: Andi
[3] Wang J, Li X 2010 Proc of the International Conference on Information, Networking and Automation (ICINA) (Kunming) Vol. 2 pp. V2-436
[4] Guo G, et al 2003 Proc of the OTM Confederated International Conferences: On the Move to Meaningful Internet Systems (Berlin) pp 986-996
[5] Trstenjak B, Mikac S, Donko D 2014 KNN with TF-IDF based Framework for Text Categorization J. of Procedia Engineering 69 pp. 1356-1364
[6] Indriani N et al 2017 Peringkasan dan Multiclass SVM pada Klasifikasi Dokumen J. of Infotel 9 pp. 416-421
[7] Manuel J, Moreno T 2014 Automatic Text Summarization Hoboken: John Wiley & Sons
[8] Korde V 2012 Text Classification And Classifier: A Survey J. of International Journal of Artificial Intelligence & Applications (IJAIA) 3 pp. 85-99
[9] Prasetyo E 2012 Proc of the Seminar Nasional Teknik Informatika (SANTIKA 2012) (Surabaya) pp. 57-70
[10] Li Y, Shen B 2017 Proc. Of the Computer and Communications (ICCC) (Chengdu) 2017 pp 2584-2588
[11] Irfan M, Zulfikar W B 2017 Proc. Of the Informatics and Computing (ICIC), 2017 Second International Conference (Jayapura) pp 1-5
[12] Li M, et al 2017 Proc. Of the Computers and Communications (ISCC) (Heraklion) pp 924-931
[13] Soucy P, Mineau G W 2001 Proc. Of the IEEE International Conference (San Jose) pp 647-648
[14] Suliantoro D W 2012 Integrasi Pembobotan TF IDF Pada Metode K Means Untuk Clustering Dokumen Teks [Theses] (Yogyakarta: Universitas Atma Jaya)
[15] Dewi K E, Widiastuti N I, Rainarli E 2017 Proc of the Seminar Nasional Informatika dan Aplikasinya (SNIA) (Cimahi) pp A-8