Innovations in Computing Technology and Applications
Volume 2 | Year: 2017 | ISBN: 978-967-0194-95-0

Text Classification of Gout Documents Using K-Nearest Neighbors Method

Nurul Nadiah Ramli and Aryati Bakri

Department of Information Systems, Faculty of Computing, Universiti Teknologi Malaysia, (UTM), 81310, Johor Bahru, Johor, Malaysia

[1] nurulnadiahramli22@gmail.com [2] aryati@utm.my

Abstract. There are variety of medical research from the anciet times till today which leads to existence of a huge number of documents in the repositories. The growth rate of documents nowadays arise the needs of the experts to conduct the classification process towards those documents in order to ensure it could be understands well. This research helps the researchers to overcome the problems of rapid growth of the documents in the databases thus the result of the classification will assist the researchers in organizing the documents. To be more specific, this research tends to do and apply the text classification technique for identifying the related features of the research by using k-Nearest Neighbors algorithm. This algorithm is applied in the classification phase after the process of text pre-processing is done towards the dataset used. As the result, the classification technique that has been applied in this research returns average accuracy of 74.74%, average precision of 83.56 % and average recall of 85.75% which is produced by the ten iteration of the model. It is shown that the technique of k-Nearest Neighbors is one of the suitable algorithms to be applied into the text classification area since the result is high.

Keywords: Text Classification, k-Nearest Neighbor, Gout

1. Introduction

In this globalization era, the growth of technology which has been improved in a rapid phase contributes a lot towards the growth of documents especially in the web-based environment. The documents could be in variety of category which requires a technology to automatically classify all those documents. This is the main reason why this research is being conducted.

In addition, the growth of technology also contributes to the increasing sources of unstructured information such as World Wide Web, biological databases, emails, blog repositories and emails. Thus, this situation causes the increasing needs of some experts to manage a huge amount of document that are available (Stavrianou et al., 2007).

The swift growth of information in textual documents creates great need for techniques for knowledge discovery from text collections (Mete et al., 2010).This situation makes the text classification is in great needs in the various industries because it is important to develop methods so that the discover knowledge embedded in these document repositories can be efficiently done.

Therefore, in order to solve the problem mentioned previously, this research tends to identify the related features for Gout disease. Due to that reason, the technique of text classification is used so that the related features could be recognized. The dataset is originally downloaded from Pubmed which is in the form of PDF documents. Throughout the process of classifying the text documents, this project utilizes Rstudio for classification purpose.

2. Related Work

There are some of the previous works or researches which are related to the Text Classification techniques that were collected and recorded in this chapter. These research papers were collected to review on the comparison of the result of the previous work which uses text Classification techniques in their researches. Table 2.2 discussed on the previous work by the researchers on the text classification techniques.

Based on the knowledge from those research papers, there are numerous types of algorithms that can be applied in text classification. In fact, the performance of the algorithm could be varied depends on some circumstances such as the size of training data set (Manne et al., 2012) and the type of classifier that the algorithm relies on (binary or multiple). The size of the dataset could give impacts towards the classification process of the classifier used in the research. This is stated by (Trstenjak et al., 2014) which mentions that there will be increasing in the time computation when the data is increase.

In the Text Classification area, there are various algorithms that could be used such as KNN, Naïve Bayes, and SVM. In the context of performance of the algorithm used, it could produce distinct result based on the few factors such as dataset used. For example, research done by Rajeswari et al. (2017) which compares on the performance of KNN and Naïve Bayes while doing text classification towards student dataset. Naïve Bayes is applied in the research which focuses on the probabilities of label values of the dataset while KNN focuses on searching of pattern space for the sample within the dataset used. Trivedi et al. (2015) in their research suggest that SVM is the best algorithm that could be used in text classification by applying their research on two datasets which are Diabetes dataset and Calories dataset. In this paper, SVM tends to classify the dataset into positive or negative classes. Although some paper states that SVM could perform better, but in some situations, the other classifier such as Naïve Bayes and KNN could produce better results than SVM (Gandhi and Prajapati, 2012). Since the majority of the researches favors that KNN is the best algorithm among the other, this research tends to choose KNN as the classifier towards the dataset of Gout texts.

. Apart from that, most of the researches that have been discussed for previous works tend to do text classification towards the dataset which is related to normal dataset, like student dataset, emails, technical documents and also the online documents which are not focusing on the specific category such as health, sports or financial documents. However, the research done by Trivedi et al. (2015) covers the dataset on the diabetes and calories which is on the health-related documents. This research triggers the interest of this research to conduct a study of text classification technique towards health-related topic. This is why this research is conducted, since there is not yet available research on text classification technique towards biomedical texts which focuses on the textual or journal documents related to Gout disease.

In the context of domain, the research which done by Trstenjak et al. (2013) is doing text classification on few categories of documents on the dataset such as sports, daily news, politics and finance. It motivates this research to do separation of the dataset into two categories which are distinct by the keywords used in the processing of collecting the dataset. Based on the previous work, the identification of the domain is important. This is due to the reason that the quality of classification is influenced by the category of the document (Trstenjak et al., 2013).

3.0 Methodology

This section explaines on the flow of this research in doing text classification technique. Figure 1 shows the step-by-step process which were done throughout this research which includes the collection of dataset, pre-processing phase, feature selection, the classification phase and also the performance measurement phase analysis.

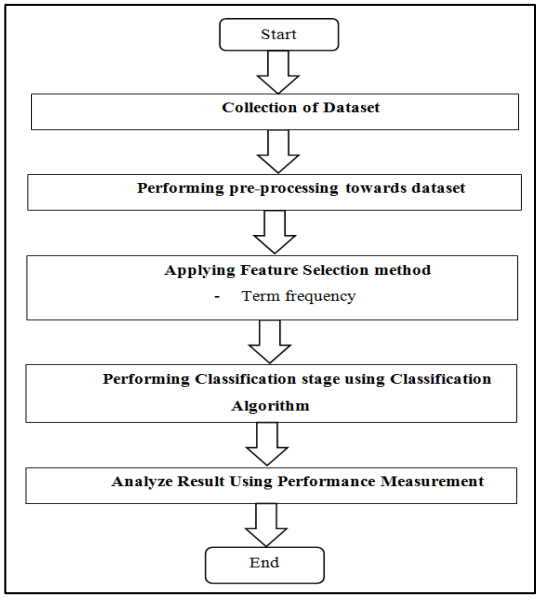

Figure 1 : Research Operational Framework

i. Collection of Dataset

The datasets for the gout is obtained through the databases named Pubmed. In this research, the data set used is divided into two categories, both of the categories are related to gout but it differs in the context of keyword used in the process of collecting data. The first category contains 14 journals and the keyword used in the searching process is “gout and seafood”. The second category of the dataset consists of 50 journals which are also retrieved from Pubmed. The keyword used in data collection process for this particular category is “Risk factor for Gout”.

ii. Text Pre-processing

Pre-processing method plays a crucial role in the application of text classification. This is where the cleaning process of the document or data is done. The goal of doing pre-processing is to identify the relevancy between word and document and also the relevancy between word and category (Gaigole et al., 2013). There are various steps in text pre-processing, but for the purpose of this research, the pre-processing steps focuses on stop words removal and stemming. These two are known as the important steps in doing pre-processing for text classification (Ramasubramanian and Ramya, 2013).

iii. Feature Selection

Feature selection is one of the important tasks to do before the tested text is ready to be mined to reduce the dimension of the original data. Since original data contains a lot of noise and stop words, feature selection steps need to be done. In this research, the steps that were done for feature selection is term frequency. Term frequency of the corpus is needed to be observed to ensure that the tested corpus is using the right documents according to the category desired based on this research. Term frequency(tf) is one of the technique in feature selection that can be used to know how many times a word appears in a corpus. The features item could be in various forms including a word, phrase or even a short language (Patra and Singh, 2013).

iv. Classification

After the pre-processing part is done, the data set is ready for the next phase which is Classification. For classification purpose of this research, the classifier used is KNN classifier. For this research, 70% of the dataset is being chosen randomly as the training dataset which contributes to 45 documents and 30% for testing dataset which equals to 19 documents,Then, in order for the classification phase to be conducted, KNN model is being set up to classify the texts in the 'goutdocs' corpus. Pseudocode of KNN model in next section defines how the KNN classifier works to classify texts.

v. Performance Measurement

In order to measure on the performance of any particular algorithm or technique used, the thing that needs to be done is the performance measurement on the chosen method, which is KNN, for this research. The classifier performance is being measured based on three properties which are *accuracy*, *precision* and *recall*. The model is being run for several iterations to compare on the performance of KNN classifier.

The accuracy is calculated by using the following formula,

$$Accuracy = \frac{(tp + tn)}{(tp + tn + fn + fp)} \quad ...(3.1)$$

Where *tp* is true positive, *tn* is true negative, *fn* is false negative, *fp* is false positive.

The precision is calculated by using the following formula,

$$Precision = \frac{tp}{tp + fp} \quad ...(3.2)$$

Where tp is true positive, fp is false positive.

The recall is calculated by using the following formula,

$$Recall = \frac{tp}{tp + fn} \quad ...(3.3)$$

Where tp is true positive, fn is false negative.

4. Experimental Result

Throughout this research, the operation of all the steps mentioned in the operational framework is done towards the dataset which is obtained through pubmed. This dataset comprises of two categories of Gout documents which differs in the keyword used in the collection of dataset phase. As for the main objectives of this research, the classification of the textual documents in this research is done by using k-Nearest Neighbor (KNN) algorithm. Then, the performane of the classifier is measured through the standard measure as accuracy, recall and precision. Accuracy represents the instances that are accurately classified, whereas precision or also known as positive predictive value is the portion of which of the retrieved instances that are relevant. Besides, recall is the proportion of real positive cases which are correctly predicted as positive.

ALGORITHM KNN MODEL

```
1  BEGIN
2  Input training dataset //goutdocs corpus
3  Classify (x,y,z) //x : training data, y: class label of x, z :unknown sample
4  for i = 1 to n do
5      Compute distance d(x_i, z)
6  end for
7  Compute set A containing indices for the k with smallest distance d(x_i, z)
8  return majority label for { y_i, where i € A}
9  END
```

Figure 2 : Pseudocode for KNN model

For classification purpose, KNN model in Figure 2 is applied towards the dataset named ‘goutdocs’. The model was iterated 10 times towards the dataset which gives the values of the Accuracy, Precision and Recall, for each of the iteration, respectively. The

model was iterated ten times so that the performance of the classifier in classifying the texts into the respective categories could be evaluated. The performance could be seen in the distribution of the result in the value produced within the confusion matrix yielded each time the model iterates.

5. Discussion

After doing all the steps mentioned above, the result which is in the recall and precision value is yielded. These values portray on the performance of the chosen algorithm**.**

	Average Precision (%)	**Average Recall (%)**
Dataset	83.56	85.75

Table 1 : The Overall Performance Measure Values

Table 1 shows the average value for precision and recall for both of the datasets. The values of these two measures are contributed by the value within the confusion matrix when the evaluation of classifier is conducted. For the calculation of recall, it involves the value of true positive and false negative. This average value of 85.75% shows that the most of the tested documents are precisely classified. This is because out of 19 tested documents, the value for true positive for all of the iteration shows more than half of the number of the tested documents, which means they are precisely classified.

For precision part, the calculation involves the value of true positive and false positive. The average value of 85.36% shows that most of the tested documents are correctly classified. This is due to the reason that as previously mentioned, out of 19 documents, the value for true positive for all of the iteration shows more than half of the number of the tested documents, which means they are correctly classified. Apart from that, the value for false positive is low which shows that the possibility of the classification of the document is wrong, is relatively low.

From the findings shown by the Table 1, it could be concluded that KNN classifier is suitable for doing text classification. The recall and precision value could portrays that the textual documents could be precise and correctly classified by using KNN since both of the measure even achieved maximum percentage of recall and precision value.

6. Conclusion

While conducting this research, there are a few problems that had been faced throughout the research process. The main problem is on the dataset itself. In the process of collecting data, there might be online database or repositories which not provide for downloadable or free accessible full text. The researchers need to purchase the document, or kindly access the repositories website through institution internet connection, since there might be the institution that the researcher studied or worked on, subscribe to the desired

repositories. Apart from that, there are also limitation on the features of the Rstudio itself which sometimes not compatible with the laptop that we used.

İf there is any person who wish to do further research in text classification area, it is advisable to use another technique such as Support Vector Machine since it can overcome the limitations that possessed by KNN algorithm. In addition, it is better to apply various methods so that the performance of the classifier could be compare. The implementation of various techniques can help researchers to understand on the dataset much better. This is due to the reason that the researcher could see how the result changes between several techniques.

References

Gaigole, P. C., Patil, L. H. and Chaudhari, P. M. (2013). Preprocessing Techniques in Text Categorization. National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2013). Proceedings published by International Journal of Computer Applications® (IJCA). 1-3.

Gandhi, V. C. and Prajapati, J. A. (2012). Review on Comparison between Text Classification Algorithms. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS). Vol. 1, Issue 3, September – October 2012. ISSN : 2278-6856 . 75-78.

Manne, S., Kotha, S. K. and Fatima, S.S. (2012). Text Categorization with K-Nearest Neighbor Approach. Proceedings of the InConINDIA 2012. 413-420.

Mete, M., Yuruk, N. Xu, X. and Berleant, D. (2010). Knowledge Discovery in Textual Databases: A Concept-Association Mining Approach. Data Engineering, International Series in Operations Research and Management Science. DOI : 10.1007/978-1-4419-0176-7_1. 225-243.

Patra, A. and Singh, D. (2013). A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms. International Journal of Computer Applications. Vol. 75– No.7, August 2013. 14-18.

Rajeswari, R.P., Juliet, K. and Aradhana (2017). Text Classification for Student Data Set Using Naive Bayes Classifier and KNN Classifier. International Journal of Computer Trends and Technology. Vol. 43 Number 1 – January 2017 (IJCTT). ISSN: 2231-2803. 8-12.

Ramasubramanian, C and Ramya, R. (2013). Effective Pre-Processing Activities in Text Mining Using Improved Porter's Stemming Algorithm. International Journal of Advanced Research in Computer and Communication Engineering. Vol. 2, Issue 12, December 2013. 4536-4538.

Stavrianou, A, Andritsos, P. and Nicoloyannis, N. (2007). Overview and Semantic Issues of Text Mining. SIGMOD Record. September 2007 (Vol. 36, No. 3. 23-34.

Trivedi, M., Sharma, S., Soni, N. and Nair, S. (2015). Comparison of Text Classification Algorithms. International Journal of Engineering Research & Technology (IJERT). Vol. 4 Issue 02, February-2015. ISSN: 2278-0181.334-336.

Trstenjak, B., Mikac, S. and Donko, D. (2014). KNN with TF-IDF Based Framework for Text Categorization. Procedia Engineering. 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013. doi : 10.1016/j.proeng.2014.03.129. 1356 – 1364.

,