Comparison of feature selection techniques in classifying stroke documents

Nur Syaza Izzati Mohd Rafei[1], Rohayanti Hassan[2], RD Rohmat Saedudin[3], Anis Farihan Mat Raffei[4], Zalmiyah Zakaria[5], Shahreen Kasim[6]
[1,2,5]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Malaysia
[3]School of Industrial Engineering, Telkom University, Indonesia
[4]Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Malaysia
[6]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info

Article history:

Received Dec 16, 2018
Revised Feb 14, 2019
Accepted Feb 28, 2019

Keywords:

Feature selection
Information gain
Pearson correlation
Support vector machine
Text classification

ABSTRACT

The amount of digital biomedical literature grows that make most of the researchers facing the difficulties to manage and retrieve the required information from the Internet because this task is very challenging. The application of text classification on biomedical literature is one of the solutions in order to solve problem that have been faced by researchers but managing the high dimensionality of data being a common issue on text classification. Therefore, the aim of this research is to compare the techniques that could be used to select the relevant features for classifying biomedical text abstracts. This research focus on Pearson's Correlation and Information Gain as feature selection techniques for reducing the high dimensionality of data. Towards this effort, we conduct and evaluate several experiments using 100 abstract of stroke documents that retrieved from PubMed database as datasets. This dataset underwent the text pre-processing that is crucial before proceed to feature selection phase. Features selection phase is involving Information Gain and Pearson Correlation technique. Support Vector Machine classifier is used in order to evaluate and compare the effectiveness of two feature selection techniques. For this dataset, Information Gain has outperformed Pearson's Correlation by 3.3%. This research tends to extract the meaningful features from a subset of stroke documents that can be used for various application especially in diagnose the stroke disease.

Corresponding Author:

Rohayanti Hassan,
School of Computing, Faculty of Engineering,
Universiti Teknologi Malaysia,
81310 Johor Bharu, Johor, Malaysia.
Email: rohayanti@utm.my

1. INTRODUCTION

In the century, the vast amount of available text documents that related to biomedical produces new challenges for the researchers in collecting specific information concerning any particular diseases such as stroke or about any specific interest in any field. The text document can be from many sources of like World Wide Web, governmental electronic repositories, biological databases, and news articles which all of this are in the form of unstructured information [1]. This issues and situation are growing fast that we need some experts to manage this huge amount of document that are available in many repository that have been mentioned in [2].

Recently, there are several approaches that have been proposed by many researchers to identify terms in biomedical literatures due to difficulties for users to find the effectively and efficiently ways for organizing data and retrieving relevant information from the text such as by performing text classification or known as text mining technique. Text classification is a process of categorizing documents automatically into their predefined classes based on their contents [3]. Mete et al. [4] also defined text classification as a process of discovering textual information that must have a fixed class for each text. However, big issue of classifying is a high dimensionality of data that consist of redundant or irrelevant data [5]. In order to reduce this high dimensionality, feature selection is the best solution to use [6]. As mentioned by [7], by employing the feature selection in classification, there are benefits can be gained such as reducing in time and storage. Other than that, the feature selection method can improve the performance by removing the redundant or irrelevant data based on their weight itself.

Among many feature selection techniques or also called as filter approaches, Wu et al. [8] describes that the proposed probabilities approach that called SVM-based probability feature selection can avoid the problem bias towards data that able to outperformed Information Gain and Chi-Square. In other hand, Sharaff et al. [9] have compared filter approaches which are Chi-Square and Information Gain with Support Vector Machines (SVM), Naïve Bayesian and J48 classifier in classifying the spam emails. A similar study also has been conducted by [10] and [9], where they used filter and wrapper approaches namely Information Gain, Gain Ratio, Chi-Square, Correlation Feature Selection, Linear discriminant analysis and Random Forest that have been applied also in classifying the spam emails. The result shows filter approaches enable the classifier achieves the improvement on classification accuracy by reducing the number of unnecessary attribute while wrapper approaches has potential highly desirable reduce the number of features but it will not affecting to accuracy of classifier. Meanwhile, [11] has introduced an approach of combination feature selection based on the average weight of features to classify Arabic corpus. In another work, [12] claimed that Distinguishing Feature Selector is better than Gini Index in selecting the features of OHSUMED dataset. In classifying the SMS Spam collection, [13] has claimed that Pearson Correlation performed with the highest accuracy compared to Symmetric Uncertainty, Chi-Square and Mutual Information feature selection techniques. This is due to Pearson Correlation is more simple and reduce computional time in building the text classification model.

In this paper, this study aims: (i) to identify the related features on risk factors of stroke, (ii) to perform feature selection removing irrelevant features on document of risk factors of stroke, and (iii) to evaluate Pearson Correlation and Information Gain techniques in classifying stroke documents. The strong related stroke documents were identified at the end of the classification process. This paper is organized as follows: In Section 1, we present the introduction of this paper. Then, in Section 2, the material and method are discussed in detail. While Section 3 presents the result and discussion of the experimental results. Finally, Section 4 provides the conclusion of this research.

2. MATERIAL AND METHOD

Figure 1 illustrates the proposed research framework of this study that consists of six phases. In phase 1, the process begun by identifying the issues and risk factors that related stroke disease. Mostly the risk factors were found and extracted from American Stroke Association, National Heart, Lung, and Blood Institute (NIH) and Stroke Foundation sources. For stroke document datasets, among the available databases, the PubMed database was referred because it is free databases that stores publicly accessible full-text of articles. For every document, title and abstract parts have been scanned whether the document belongs to risky or non-risky stroke document. Furthermore, these two parts were selected because they potray the whole content in documents. Table 1 shows keywords that have been used to search the stroke documents.

Based on Table 1, 100 documents have been selected as a dataset in this study which later have been divided into two categories risk factor and non-risk factor. As shown in Table 2, the first category is "risk factor" that contains 60 journals while the second category is "non-risk factor" contains 40 journals. Even though the keyword "factors of stroke" or "risks of factors of stroke" are used for searching the documents, the document that not related to keyword also exists in the query so that the documents that not related to risk factors will put in non-risk factors category.

Table 1. Keywords used in Searching Stroke Documents

Database	Keyword	Result (Documents)
PubMed	Stroke	284575
	Factors of Stroke	99141
	Risk factors of stroke	60948

Pre-processing method plays an important role in the application of text classification which this phase involved the cleaning and preparing of text to proceed to next step. The aim of this pre-processing is to select the relevant words that carry the meaning and remove the words that not contribute to differentiating between the documents [14]. In this study, the pre-processing step focused on stop words removal and stemming. These two were known as the important steps in doing pre-processing for text classification [14]. The stop words removal aims to reduce the dimensionality of term space while the stemming discovers the root word or base word for any particular term. The document term matrix (DTM) is output from this phase that contains the documents within the corpus for its rows while the columns represent the count for each of the features that appear within the corpus in the csv format. R language was used in pre-processing step.

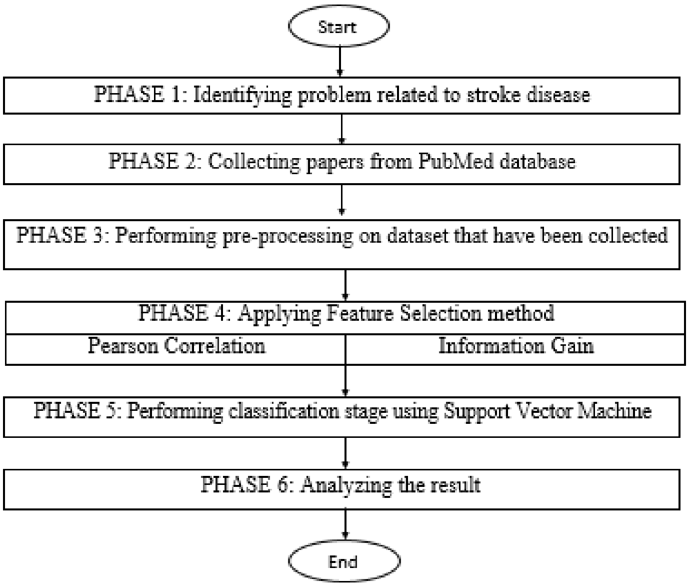
Start
PHASE 1: Identifying problem related to stroke disease
PHASE 2: Collecting papers from PubMed database
PHASE 3: Performing pre-processing on dataset that have been collected
PHASE 4: Applying Feature Selection method
Pearson Correlation
Information Gain
PHASE 5: Performing classification stage using Support Vector Machine
PHASE 6: Analyzing the result
End

Figure 1. Research framework

Table 2. Stroke Documents

Class	Document Category	Number of Documents
Yes	Risk Factor	60
No	Non-Risk Factor	40

After the text pre-processing is done, the Pearson’s Correlation and Information Gain feature selection are applied to filter the datasets using Waikato Environment for Knowledge Analysis (WEKA). Feature selection can solve the high dimensional of data that always occur in text classification by reducing the irrelevant, noise and redundant features which are burden on challenging tasks [15]. Based on the different strategies of searching, feature selection can be classified into three approaches which are filter approach, wrapper approach and embedded approach [16]. This study only focus on Pearson’s Correlation and Information Gain which are under filter technique.

Basically, both feature selection technique used the same input that refer to document term matrix (DTM) which it is the result of pre-processing stage and the WEKA software capable to load CSV format of DTM and convert to ARFF format to proceed to feature selection phase. Besides, both of this technique also applied feature ranking that providing a rating of the features that orderly by their score to the evaluator and generally it performs the rank which features should be obtain high or low rank according to the selected features in the given datasets [17].

The concept of Pearson Correlation and Information Gain technique in selecting the subset of relevant features from the extracted features of the stroke documents are explained as above:

a) Pearson’s Correlation

The way Pearson correlation coefficient ρ dealings the strength of the relationship between two features to find the similarity between of them, is based on value which the giving a value between +1 and –1, where 1 indicates positive, 0 indicates no correlation and -1 is negative correlation [18].

b) Information Gain

Information Gain measures the amount of information in bits obtained for prediction of a class by determining the presence of a feature in a dataset. It determines the change in entropy when the feature is present vs. when the feature is absent. Entropy is a measure of uncertainty or unpredictability in a system. It is the basis for Information Gain attributes ranking methods [18].

The effectiveness of those two feature selection techniques is evaluate in classification phase as mentioned earlier. The experiments conduct using three sets of features which involved subset of dataset before and after feature selection using IG and Pearson's Correlation. The classification process also performs by WEKA tools which SVM classifier is used. In this research, 70% of dataset is being chosen randomly as the training which contribute 70 documents and 30% for testing dataset that equal to 30 documents.

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 SVM, for this research. The classifier performance is being measured based on three properties which are accuracy, precision and recall. The model is being run on subsets of stroke documents without and with feature selection. The accuracy is calculated by using the following formula stated by [19],

$$\text{Accuracy} = \frac{(tp+tn)}{(tp+tn+fn+fp)} \tag{1}$$

The precision is calculated by using the following formula stated by [19],

$$\text{Precision} = \frac{tp}{(tp+fp)} \tag{2}$$

Where *tp* is true positive, *fp* is false positive.
The recall is calculated by using the following formula stated by [20],

$$\text{Recall} = \frac{tp}{(tp+fn)} \tag{3}$$

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

3. RESULTS AND DISCUSSION

In this section, analysis on text preprocessing and also analysis on feature selection will be discussed in details.

3.1. Analysis on Text Preprocessing

Figure 2 shows the most frequent features within the documents from “strokedocs” corpus using Pearson Correlation and Information Gain evaluation. The classification process aim to identify the strong related stroke documents. As a result, after pre-processing is done towards the corpus of text documents, the most frequent features that have been extract from documents within corpus using Graph Bar as visualization method. The features demonstrate that “Stroke” present the highest number of frequency. “Stroke” show the highest number of frequency because original datasets mostly about the text or document regarding to Stroke disease which is searched by using certain keywords and those keywords always include word ‘stroke’ in the query. In addition, the documents in corpus mostly review about “Stroke”, for that reason “Stroke” present the highest frequency compared to other terms.

Apart from that, from the extracted documents there are five most frequent risk factors of stroke that have been mined as tabulated in Table 3. Based on Table 3, there are a few risk factors appear which show that risk factors like hypertension, age, smoking, diabetes were always being issues on the documents. Hypertension also known as high blood pressure is the common risk factors of stroke which it put a strain on all the blood vessels throughout our body including the brain that the lead one then our heart has to work much harder to keep the blood circulation going but this strain can damage our blood vessels which causing them to become harder and narrower, a condition called atherosclerosis then it makes a blockage more likely to occur, which could cause a stroke or transient ischaemic attack (Stoke Association, 2012). Even though, the rare case, this extra strain may cause a blood vessel to weaken and burst inside the brain that will causing bleeding into surrounding tissue that called haemorrhagic stroke (Stoke Association, 2012). According to State of the nation. (2018), stroke can attack to anyone of any age including babies and children and usually the causes of stroke in children are very different from those in adult. Besides, State of nation. (2018) also stated that the rate of first time strokes in people aged 45 and over is expected to increase by 59% in the next

20 years. Actually, risk factor of age is closely related to the way of their lifestyle itself which involve consumption of alcohol, illegal drug and also smoking habit (State of nation, 2018). Besides, there are about 5% stroke occurs in adults around 18 to 44 years old due to the substance abuse like consumption of alcohol, illegal drug and also smoking habit that stated by (Ríos, F et al. 2013)

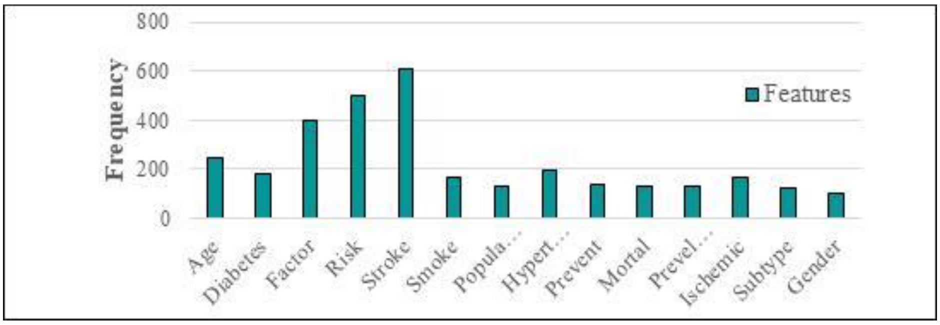

Figure 2. Frequent features of the dataset

Table 3. Risk Factors of Stroke

Features	Frequency
Age	250
Hypertension	197
Diabetes	180
Smoking	167
Gender	100

3.2. Analysis on Feature Selection

Table 4 shows the number of features that have selected after applying the different feature selection. There are 2021 features from the original document after the extraction information from unstructured to structured information that represent by features. After applying Pearson’s Correlation feature selection only 923 feature are selected while when applying Information Gain feature selection only 9 features were selected. When applying different feature selection, the result also differs due to the weighted that apply the feature selection itself. The performance of classification was then tested using SVM with using different feature selction techniques namely Pearson’s Correlation and Information Gain. Figure 3 demonstrates the accuracy of SVM classifier on different feature selection techniques. The highest accuracy was performed when using with Pearson’s Correlation feature selection technique which is 94.12%. Information Gain was then performed with 91.18% accuracy, while the accuracy without using any feature selection technique was only 79.41%. This showed that selection of the relevant features able to boost the accuracy of text classification.

On the other hand, the performance of classification also been tested in term of precision and recall. Precision can be known as positive predictive value which measure the portion that shows the level of relevant of the retrieved instance that will be affected the value of accuracy. Meanwhile, Recall also known as sensitivity which measure the fraction of relevant instances that are retrieved instance that will be affected the value of accuracy. Yet, Pearson’s Correlation has outperformed Information Gain with the highest precision and recall, at 94.10%. The recall and precision value could portray that the textual documents could be precise and correctly classified by using Pearson Correlation and Information Gain (IG) feature selection since both of the measure even achieved maximum percentage of recall and precision value.

Table 4. Result of Number of Features Selected with Different Feature Selection

Feature Selection(FS)	Number of features selected
Without FS	2021
Pearson’s Correlation	923
Information Gain	9

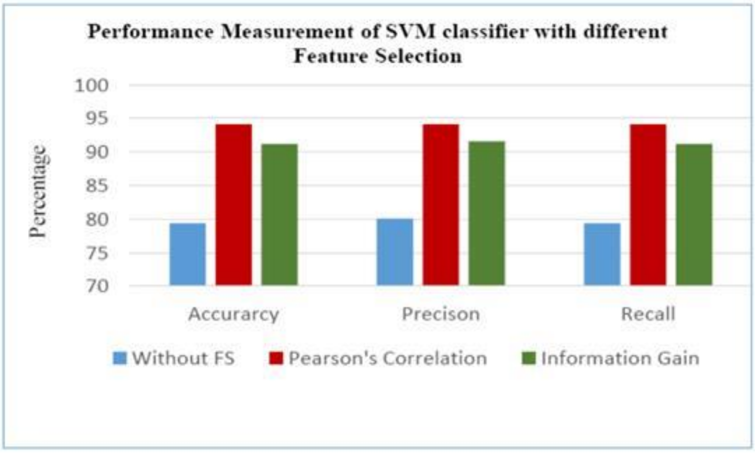

Figure 3. Performance measurement of SVM classifier with different feature selection

4. CONCLUSION

A huge amount of biomedical documents in repository gives difficulties for users to find the effectively and efficiently terms in biomedical literatures. Hence, a strategy to identify features on risk factors of stroke was proposed and the used of feature selection techniques such as Pearson Correlation and Information Gain were successfully filters the irrelevant features on documents. For future works, expert validation could be considered as a part of weighted quatification in selecting the more relevant features.

ACKNOWLEDGEMENTS

This research was supported by FRGS Grant and UTM Grant under Vot No. 4F973 and 16H73 respectively.

REFERENCES

[1] Singh, A. (2013). Text Mining : A Burgeoning technology for knowledge extraction, 1(March), 22–26.

[2] Stavrianou, A., Andritsos, P., and Nicoloyannis, N. (2007). Overview and semantic issues of text mining. ACM SIGMOD Record, 36(3), 23. https://doi.org/10.1145/1324185.1324190

[3] Surkar, M. Y. R., and Mohod, P. S. W. (2014). A Review on Feature Selection and Document Classification using Support Vector Machine, 3(2), 933–937.

[4] 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.

[5] Bali, M., and Gore, D. (2015). A Survey on Text Classification with Different Types of Classification Methods, 4888–4894. https://doi.org/10.15680/ijircce.2015.0305174

[6] Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. International Conference on Machine Learning (ICML), 1–8. https://doi.org/citeulike-article-id:3398512

[7] Morariu, D. I., Ulescu, R. G. C., and Breazu, M. (2013). Feature Selection in Document Classification □ Relevant □ □ □ Retrieved □.

[8] Wu, K., Lu, B.-L., Uchiyama, M., and Isahara, H. (2007). A probabilistic approach to feature selection for multi-class text categorization. Advances in Neural Networks–ISNN 2007, 1310–1317

[9] Sharaff, A., Nagwani, N. K., and Swami, K. (2015). Impact of Feature Selection Technique on Email Classification. International Journal of Knowledge Engineering-IACSIT, 1(1), 59–63. https://doi.org/10.7763/IJKE.2015.V1.10

[10] Parimala, R., and Nallaswamy, R. (2011). A Study of Spam E-mail classification using Feature Selection package. Global Journal of Computer Science and Technology, 11(7), 45–54.

[11] Adel, A., Omar, N., and Al-Shabi, A. (2014). A comparative study of combined feature selection methods for Arabic text classification. Journal of Computer Science, 10(11), 2232–2239. https://doi.org/10.3844/jcssp.2014.2232.2239.

[12] Parlak, B., and Uysal, A. K. (2016). The impact of feature selection on medical document classification. Iberian Conference on Information Systems and Technologies,CISTI,2016–July(1503). https://doi.org/10.1109/CISTI.2016.7521524

[13] DeepaLakshmi, S., and Velmurugan, T. (2016). Empirical study of feature selection methods for high dimensional data. Indian Journal of Science and Technology, 9(39), 1–6.

[14] 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, 2(12), 4536–4538. Retrieved from www.ijarcce.com

[15] Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. Computers and Electrical Engineering, 40(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[16] Miao, J., and Niu, L. (2016). A Survey on Feature Selection. Procedia - Procedia Computer Science, 91(Itqm), 919–926. https://doi.org/10.1016/j.procs.2016.07.111

[17] Dinakaran, S., and Thangaiah, P. R. J. (2013). Role of Attribute Selection in Classification Algorithms. International Journal of Scientific & Engineering Research, 4(6), 67–71. https://doi.org/June 2013

[18] Phyu, T. Z., and Oo, N. N. (2016). Performance Comparison of Feature Selection Methods. MATEC Web of Conferences, 42, 6002.

[19] Sokolova, M. and Lapalme, G. (2009). A Systematic Analysis of Performance Measures for Classification Tasks. Information Processing and Management 45. doi:10.1016/j.ipm.2009.03.002. 427–437.

[20] Pitre, S., Hooshyar, M., Schoenrock, A., Samanfar, B., Jessulat, M., Green, J. R., Dehne, F. and Golshani, A.. (2012). Short Co-Occurring Polypeptide Regions Can Predict Global Protein Interaction Maps. Bioinformatics. DOI: 10.1038/srep00239. 1 -10.