Classification of Biomedical Literature in Hypertension and Diabetes

Nur Aniq Syafiq Rodzuan[a,1], Shahreen Kasim[a,2,*], Mohanavali Sithambranathan[a], Muhammad Zaki Hassan[a]

[a] Faculty of Computer Science and Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, Malaysia.
[1] AI170189@siswa.uthm.edu.my; [2]shahreen@uthm.edu.my
* corresponding author

ARTICLE INFO

Article history
Received April 25, 2020
Revised April 30, 2020
Accepted August 11, 2020

Keywords:
classification
biomedical literature
hypertension
diabetes

ABSTRACT

Textual information gives us more clear information as it is presented using words and characters, which is easy for humans to understand. To extract this kind of information, text mining was introduced as new technology. Text mining is the process of extracting non-trivial patterns or knowledge from text documents or from textual databases. The purpose of this research paper is to perform and compare keyword extraction using statistical and linguistic extraction tools for 120 text documents related to hypertension and diabetes disease. In order to draw this comparison, RStudio, a statistical-based tool and TerMine, a linguistic-based tool have been used to demonstrate the process of extracting the specified keyword from the biomedical literature. Thus, classification evaluation using Naïve Bayes classifier is carried out in order to evaluate and compare the performance of the statistical and linguistic approaches using these tools. Experimental results show the result of the comparison and the difference between both tools in executing extraction keywords.

1. Introduction

Diabetes or medically termed, ‘Diabetes Mellitus’ is categorized as a high blood glucose level that results in the deficiency of insulin produced in the body, or the body’s resistance to the effect of insulin [1]. Frequent urination, increased thirst and increased hunger are signs of high blood sugar. Diabetes mellitus has two types of categories. Type 1 diabetes is insulin-dependent diabetes (IDDM). This type occurs when there is no longer insulin or very little insulin produced by the pancreas. The other type of diabetes is non-insulin produced by pancreas or the insulin produced is not absorbed effectively by the cell in the body [4].

Over 246 million people suffer from diabetes worldwide with a majority of them being women. According to the World Health Organization (WHO) report, diabetes is ranked fifth as the fatal disease with no treatment yet to be reported and the amount of individual diagnosed from this disease is predicted to increase to over 380 million by 2025[6].

Hypertension also known as high blood pressure can be affected by many factors, such as physical inactivity, tobacco and alcohol use. There are almost one billion people who have been affected with hypertension or high blood pressure, in which two-thirds are in developing countries

according to World Heart Federation. According to Centers for Disease Control and Prevention, in 2014, more than 410000 Americans had lost their lives due to the hypertension that includes 1100 deaths per day.

This is alarming because it will make the heart work harder to pump blood out to the body and it will lead to the hardening of arteries, also to stroke, kidney disease and the development of heart failure [5]. Textual information is presented by using words and characters and this will provide a lot of fine information for users. Therefore, the purpose of this research is to perform and compare keyword extraction using statistical and linguistic extraction tools from text documents related to the treatment of hypertension and diabetes.

2. Method

2.1. Dataset Collection

The framework of the research will be presented based on the research flow. There are six phases established before the end result can be achieved. The following Figure 1 shows the flow of this research framework.

2.2. Research Framework

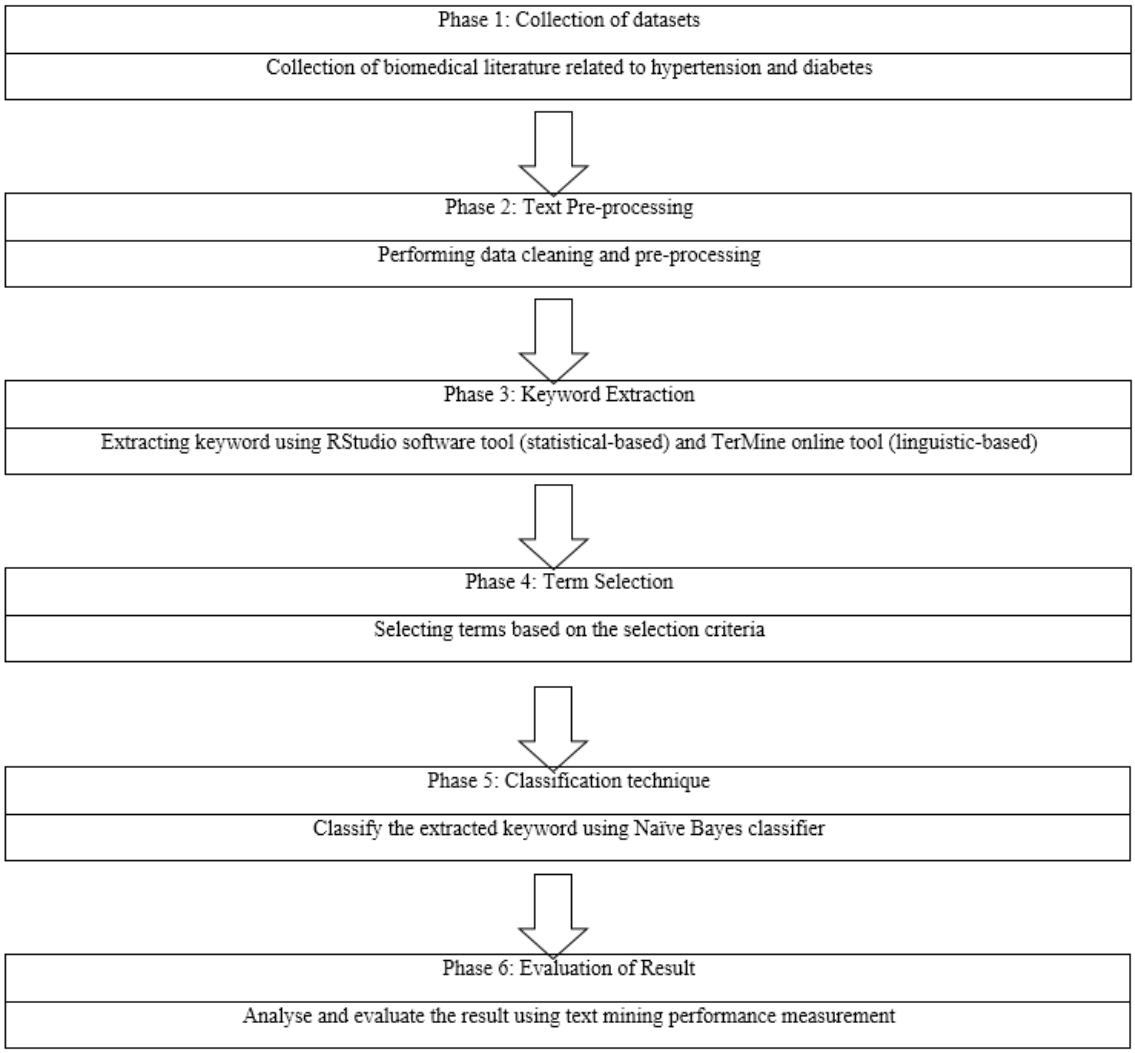

Fig 1. Research workflow

2.2.1. Phase 1: Collection of Datasets

The most important step r to initiate this research is to collect the datasets. The datasets are collected mostly from the Internet as most of the papers related to the research can be downloaded easily. The text documents used in this researchare is 60 hypertension and 60 diabetes journals and research papers collected from PubMed website, https://www.ncbi.nlm.nih.gov/pubmed/. PubMed has been chosen because PubMed comprises more than 30 million citations for biomedical literature from MEDLINE, life science journals and online books (PubMed). The documents are selected based on biomedical texts that consist keywords related to hypertension and diabetes.

The keywords used to search the biomedical literature are “hypertension” and “diabetes”. For the keyword “hypertension”, as the first result after inserting the keyword in the search bar, 516,525 resources were shown. There are many resources but some of the documents were already outdated as they had been published since year 1990 and above. Therefore, after applying the publication dates filter for 5 years, there were 107,907 resources left. For the keyword “diabetes”, as the first result after inserting the keyword, 717,782 resources were shown. The publication dates filter for 5 years was also applied to this keyword and the results show that there were 2017,116 resources shown within the range of 5 years.

2.2.2. Phase 2: Text Pre-processing

During this phase, text pre-processing is done in order to make sure that the important keywords are included in the text. There are two types of text pre-processing involved in this research, which are statistical and linguistic pre-processing. There are many types of phases in text pre-processing, but in this research, the phases involved are data cleaning, stop word removal, stemming and part-of-speech (POS) tagging.

a. **Statistical Pre-processing**

The statistical pre-processing was carried out using the RStudio tool. The tool can be downloaded from the website https://www.rstudio.com/products/rstudio/download. After installing the RStudio, the R tool must be installed in the computer in order to make sure that RStudio is functioning.; the tool can be downloaded from the website https://cran.r-project.org/bin/windows/base/.

In the statistical pre-processing, the phases involved are data cleaning, stop word removal and stemming. The data cleaning phase includes a few steps such as characters conversion into lowercase, numeric removal, punctuation removal and whitespace removal. Stop word removal is a phase where the number of common words used in the text document is reduced. Next, for stemming, words in the text documents will be classified in terms of their root or stem words. All the three phases were done by running the command in RStudio.

b. **Linguistic Preprocessing**

This process helped in finding named entities and improved the selection of nouns or other important words from a document [7]. All related terms were identified by applying POS tagging, extracting word sequences of adjectives or nouns and stop-list. The linguistic pre-processing was done using the TerMine tool.

2.2.3. Phase 3: Keyword Extraction

After the pre-processing for the biomedical literature was complete in the previous phase, this phase is all about the keyword extraction on the text documents. The tools in this phase are the same tools used in the previous phase which are RStudio and TerMine.

a. **Statistical Approach using RStudio**

RStudio is an integrated development environment (IDE) for R. RStudio includes a console, a syntax-highlighting editor that supports direct code execution, and tools for plotting, history, debugging and workspace management. RStudio provides the most widely used open source and

it is enterprise-ready and it can be run on the desktop or in a browser that was included with RStudio Server or RStudio Server Pro. Term frequency is used as the method of statistical feature in order to measure the keyword. Term frequency indicates how often a word occurs in a text document.

b. **Linguistic Approach using TerMine**

TerMine is an online text mining tool by The National Centre for Text Mining (NaCTeM). The TerMine demonstrator combines the c-value multiword term extraction and AcroMine acronym recognition. The demonstration system will explain the input resources by recognising the multiword term by c-value and acronym recognised by AcroMine.

The method of the linguistic feature used in this research to measure the keyword is c-value. By using c-value, the result from the keyword extraction will be represented with the c-value score. C-value combines the linguistic and statistical analyses as it is a domain-independent method for automatic term recognition (ATR).

2.2.4. Phase 4: Term Selection

During this phase, the selection is completed by selecting through exact selection criteria. For the statistical approach, the term frequency higher than, or equal to 30, is selected, and for the linguistic approach, the c-value score higher than, or equal to, 3 is selected. Term selection is used because the range of the targeted extracted keyword is in the range of 80 to 120. Term selection can help in improving the prediction performance in classifying the data as it removes unnecessary features within the document [3].

2.2.5. Phase 5: Classification Technique

Weka tool accepts only the Arff file format. Hence, ArffViewer in Weka is available to convert other file formats into the Arff file format. Classification is completed by choosing the right classifier in Weka. In this research, the classifier used is Naïve Bayes classifier. The method used to train the data is the K-fold cross-validation.

Naïve Bayes is a simple Bayesian Network that assumes on finding nodes that are restrictively independent of each other [10]. The term represents that there is restrictive independence among the features or attributes. The probability parameters are predicted from the training data where the parameter predicted from the data is completed by using maximum likelihood estimation [9].

K-fold cross-validation is a good evaluator for acquiring the error rate of a learning algorithm [8]. This method is one of the most popular and practical methods because it is simple and it has an obvious universality. It is also used to evaluate the probability of an evaluator [2].

2.2.6. Phase 6: Evaluation of Result

The final phase of this research is to evaluate the result. After producing the result, the result will be analysed and evaluated using performance measurement. The measurement of the performance of the classification task in Weka will be using the precision, recall and F-measure. Precision measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means more false positives. Recall measures the completeness, or sensitivity of a classifier. Higher recall means less false negatives, while lower recall means more false negatives. Precision and recall can be combined to produce a single metric known as F-measure, which is the weighted harmonic mean of precision and recall. The main advantage of using F-measure is that it is able to rate a system with one unique rating.

3. Result and Discussion

True Positive (TP) gives information on how often the data that correctly classified the document is related to the diseases. False Positive (FP) refers to how often the data that classified the document was related to the diseases when there was no relation at all. Next, True Negative (TN) is to see how often the data that correctly classified the document was unrelated to the diseases. Finally, False Negative (FN) deals with how often the data that classified the document was not seemingly related to the diseases but in fact, was related.

Figure 1 shows the result of performance measurement for both statistical and linguistic based tools. For the statistical-based tool, the average precision value is 0.588, the recall value is 0.592 and the F-measure value is 0.590. There is less difference between precision and recall. This is probably because the values of false positive (FP) and false negative (FN) are close to each other.

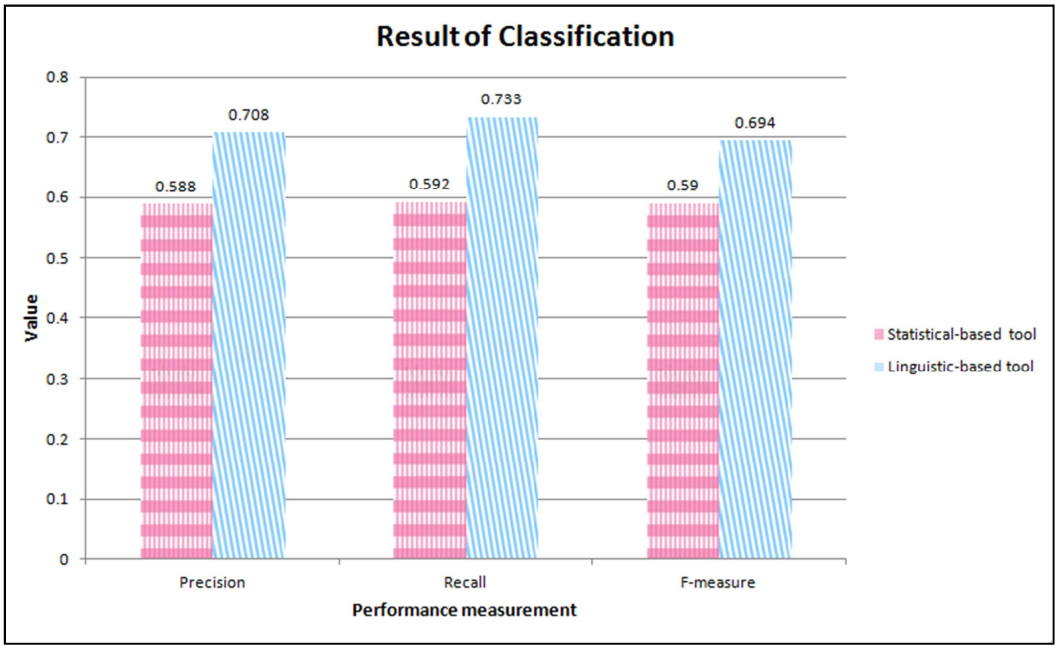

Fig 2. Performance measurement result

Table 1. Result of database hypertension and diabetes using Naïve Bayes

Approaches	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Statistical (RStudio)	60.3	58.8	59.2	59.0
Linguistic (TerMine)	72.4	70.8	73.3	69.4

4. Conclusion

For the linguistic-based tool, the average precision value is 0.708, the recall value is 0.733, and F-measure is 0.694. The result of the linguistic-based tool is obviously higher compared to the result of the statistical-base tool. The higher the F-measure value, the better the predictive classification procedure. A score of 1 means the classification procedure is perfect while the lowest possible F-measure is 0. In this research, the value of 0.694 is near-perfect because the F-measure value is nearest to 1 instead of the value of 0.90. Therefore, the result from the linguistic-based tool is better than the statistical-based tool.

References

[1] Ali, R., Hussain, J., Siddiqi, M. H., Hussain, M., and Lee, S. (2015). H2RM: A Hybrid Rough Set Reasoning Model for Prediction and Management of Diabetes Mellitus, 15921–15951.

[2] Arlot, S., and Celisse, A. (2009). A survey of cross-validation procedures for model selection, 4, 40–79. doi: 10.1214/09-SS054.

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

[4] Gülçin Yıldırım, E., Karahoca, A., and Uçar, T. (2011). Dosage planning for diabetes patients using data mining methods. Procedia Computer Science, 3, 1374–1380.

[5] Holland K., (2017), Everything You Need to Know About High Blood Pressure (Hypertension). Retrieved from http://www.healthline.com/health/high-blood-pressure-hypertension.

[6] Iyer, A., S, J., and Sumbaly, R. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining and Knowledge Management Process, 5(1), 01–14.

[7] Jurafsky, D., and Martin, J. H. (2016). Part-of-Speech Tagging. In Speech and Language Processing. Retrieved from http://en.wikipedia.org/w/index.php?title=Part-of-speech_tagging&oldid=550410494.

[8] Kale, S., Kumar, R., and Vassilvitskii, S. (2011). Cross-Validation and Mean-Square Stability.

[9] Pineda A. L., Yea Y., Visweswarana S., Coopera G. F., Wagnera M. M., and Tsuia F., J Biomed Inform. (2015) December ; 58: 60–69. doi:10.1016/j.jbi.2015.08.019.

[10] Spasić, I., Greenwood, M., Preece, A., Francis, N., and Elwyn, G. (2013). FlexiTerm: a flexible term recognition method. Journal of Biomedical Semantics, 4(1), 27.