A Framework for Text Mining in Scientometric Study: A Case Study in Biomedicine Publications

V M M Silalahi, R Hardiyati, I M Nadhroh, T Handayani, R Rahmaida, and M Amelia
Centre for S & T Development, Indonesian Institute of Sciences, Jl.Jend. Gatot Subroto 10, Jakarta, Indonesia

E-mail: mesnans@yahoo.com

Abstract. The data of Indonesians research publications in the domain of biomedicine has been collected to be text mined for the purpose of a scientometric study. The goal is to build a predictive model that provides a classification of research publications on the potency for downstreaming. The model is based on the drug development processes adapted from the literatures. An effort is described to build the conceptual model and the development of a corpus on the research publications in the domain of Indonesian biomedicine. Then an investigation is conducted relating to the problems associated with building a corpus and validating the model. Based on our experience, a framework is proposed to manage the scientometric study based on text mining. Our method shows the effectiveness of conducting a scientometric study based on text mining in order to get a valid classification model. This valid model is mainly supported by the iterative and close interactions with the domain experts starting from identifying the issues, building a conceptual model, to the labelling, validation and results interpretation.

1. Introduction
The data of Indonesians scholarly research publications in the domain of biomedicine has grown in significant quantities. This amount of data can serve as a resource for an analysis to get a picture of the development of science and technology in the domain of biomedicine in Indonesia. The efforts to analyse biomedicine scientific publications in Indonesia have been conducted in the form of bibliometric studies as well as co-authorship network analysis. However, the rich information stored in the full text form has so far not been explored. For this purpose, a scientometric study was conducted through text mining the research publications. One of the goal in the scientometric study is to get a view of how much of the research output that lead to commercialization. To this end a predictive model is built that provides a classification of documents as a proxy to the potency for downstreaming of the research outputs in the direction to commercialization. In addition a framework is proposed which is suited in conducting a scientometric study based on text mining. An investigation is conducted relating to the problems associated with building a corpus and the effort to produce a valid training data through labelling of the document. Most of the existing frameworks in text mining focused on the technicality and management whereby in our framework we put the focus on the close interaction with the domain expert. In general, the issues in biomedical text processing is the problems with large volume, high dimension, and heterogeneous data sources. Moreover there is also issue
related with inconsistent and incomplete data, as well as with data which is rich in context. This is a challenge in the field of biomedical informatics given the data existence such as unbalanced data sets, structured or weakly structured data, noisy and ambiguous labelling [6]. For the purpose mentioned above, in this paper, an effort is described to build a special corpus of research data in the domain of Indonesian biomedicine, with the primary objective of supporting research in the field of scientometrics in Indonesia. The corpus is constructed with a good design in mind by investigating some common conditions such as: the representation of the collected sample, the homogeneity according to the selection criteria, and the balance of the class, as well as the careful documents labelling in order to avoid ambiguous results.

2. Literature Review

According to [10] scientometrics can be defined as the application of quantitative methods related to the analysis of science in the context of information dissemination. Scientometrics can also be defined as a quantitative study of science, communication in science, and science policy [5]. The definition of scientometrics research is also given by [14] as a quantitative study of science and technology. Scientometrics is associated with bibliometrics and has partially the same area as bibliometrics [12]. As far as the authors know, there is no text mining based scientometric research has been found on biomedicine which focuses on the condition of Indonesia as a mega biodiversity country. Meanwhile, some researchers, among others, [8] used a co-word approach to map the results of research produced by Biology Research Centre, Indonesian Institute of Sciences. Ref. [9] conducted a scientometric study using social network analysis methods to investigate central actors, institutions involved, and popular topics. Meanwhile Ref. [2] conducted a foresight study using text mining methods to investigate the many tasks involved in the process. The process involves a qualitative method later in the process after the quantitative method. The case study for the foresight study was conducted in the domain of nanotechnology in Brazil.

3. Framework

The classification model is built within a five step processes starting from conceptual modelling, data searching and selection, data cleaning and pre-processing, models building and selection, and finally evaluation and results interpretation. Fig 1. Shows the text data analytics process for the purpose of a scientometric study.

![Figure 1. Text mining process for the scientometric study.](image)

There are frameworks proposed for research study based on text mining. However most of the proposed frameworks have a focus on the technicality in a text mining process. The CRISP-DM is a generic model for the data mining based project. It can be used as a generic check-lists useful for planning the project [15]. Some are very specific in purpose such as in [11]. This paper proposes a framework in doing a text mining in a scientometric study which has a focus on the role of the domain experts. Fig. 2 shows a framework built to perform text-based scientometric study. For this purpose a conceptual model of document classification is constructed in accordance with the objective of the study. The objective is to classify document according to drug design process as a proxy to show the potency to commercialization. The basis used is the theoretical approach or best practices contained in official documents that can be used as a reference. For the purpose of building this conceptual model, domain experts in biomedicine are involved. In addition to this, the roles of a domain expert range from validating the conceptual model, labelling the documents, and interpreting results. The conceptual model building activity is supported by a literature study in which key features or predictive terms identification is conducted to support the conceptual model. The corpus development includes the searching, selection and the pre-processing of the document. Furthermore, in the proposed
framework there are also activities carried out such as named entity mining, relation mining and event mining for broader purposes in scientometric study of. A named entity (NE) is a phrase or combination of phrases in a document that indicates a particular object or group of objects, such as genes, proteins, cells, drugs, chemicals, and diseases in the biomedical literature. An attempt to obtain data on this NE will provide links into descriptions to semantic layers e.g. by connecting annotations from NE in the text with ontologies to gain structured knowledge.

**Figure 2.** A framework for text mining based scientometric study.

### 3.1. Conceptual Modeling

Conceptual model constructs are built with reference to the existing official model such as by the FDA [3] and then through the synthesis of some of the models present in the literature such as in [7]. In an effort to develop a conceptual model suitable for the research question in this scientometric study, there will be several in-depth interviews with domain experts and resource persons from research centres, industry as well as from academics (universities). In-depth interviews are also conducted to obtain facts in the practice of promoting biomedicine research results to the industry for its downstreaming purposes, as well as to get an overview of the issues, problems faced and how the solutions are taken to go downstream. From the data analysis perspective, the conceptual model is built by taking into account the amount of publication data that can be collected, the estimated size of the distribution of documents or publications according to the classification model, and to a more detailed level with regard to mentioned predictory terms in the text. The predictory terms obtained were analysed to get input in constructing a conceptual model and then grouped in a list of terms that became predictors for the purpose of the initial classification. The list of terms obtained is then used as the material in the discussion with the domain expert in an in-depth interview. Table 1 shows a sample from the results of the features identification for the classification model.

**Table 1.** Sample from the result of feature selection for the classification model.

| Stage in new medicine process discovery | Stage in new medicine process discovery |
|----------------------------------------|----------------------------------------|
| 1 Basic Research                       | potential compound; inhibition; activation |
| 2 Drug Discovery                       | inhibition; activation; therapeutic effect; promising compound; lead compound; sequence homology studies; DNA microarray; structural genomics; high-throughput screening |
3 Pre-Clinical studies antioxidant; assay; anti-proliferation effect; cytotoxic effect; cultured cell; efficacy, bioactive protein fraction; antithrombotic; thrombolytic agents; toxicity study; dose; mice; male; female; fasted; observation period; major organs; rat; male; female; adult; body weight; pregnant; prenatal developmental study; efficacy; safety; absorption; distribution; metabolism; excretion; bioavailability; bioequivalence; dosage level

4 Clinical Trials effectiveness; tumor necrosis factor-α (TNF-α) levels; dengue hemorrhagic fever (DHF); double-blind; randomized; placebo-controlled trial; patients; human, treatment group; placebo group; significant; volunteers; safety; efficacy

The initial conceptual model is then used to label the collected documents which will then be validated by the expert. The results of this analysis form the basis for building a computational model by text mining method. After processing the data analysis, the next step is to confirm the results that has been obtained through Focus Group Discussion (FGD). In the FGD, domain experts are expected to build consensus of the results and possibly refining the labelling. The results of these discussions are expected to support in better research policy recommendations in the future.

3.2. Data Searching and Selection

The data used in this study is a collection of international scientific publications from Scopus, Science Direct, DBLP, PubMed Central, Google Scholar as well as other accessible online data sources. The process of collecting data is done by using search keywords that are built in accordance with the research goals. The data of international scientific publications collected are data derived from authors with Indonesian affiliates up to 2017. Due to limited amount of data collected from qualified publication sources, time limitation and other resources, then downloading text from websites and Google Scholar proved to be a viable solution. The search also investigates the research sites and academic institution sites in Indonesia. Some university websites found to be very helpful, because it contains a list of Indonesian biomedical journal. The biomedical documents collected for the corpus are full-text papers in the form of journals, workshop reports, or conference proceedings. Most of them are text in pdf format which is then converted into plain text format encoded in UTF8 format. Data collection is done in several stages. The first step is to determine the search terms used in collecting data on the basis of the research questions. The search terms used to capture biomedicine-related articles are such as biomedicine, drugs, and herbal medicine, medical biology combined with the word ”Indonesia”. The result of the searches through the Scopus database is 209 articles which is still quite small. Therefore, the addition of data related to Indonesia biomedicine articles is captured through Google Scholar, university repository, and the like. This addition generated a database of 662 articles. The steps used in the data selection are as follows: the types of documents are only in the form of articles in English, articles are produced by Indonesian authors and articles are related to Indonesian biomedicine, which uses Indonesian resources. Data selection starts from the title, abstract, and then the full-text of an article. In this step, all data is aggregated to be checked for duplication. Based on the results of data integration, we obtained data of Indonesian international scientific publications in biomedicine as much as 543 articles.

3.3. Data Cleaning and Data Preprocessing

Text conversion is largely based on the conversion from pdf format to plain text format. Documents are collected either in pdf image format or pdf text but mostly in pdf text format. To meet the needs with common format documents in the annotation and analysis process, the format will be a plain text. Therefore, two types of document converter tools are used: pdf-to-txt converter and OCR tool which is available online. In preparing our corpus we emphasize the importance to anticipate the process of text annotation. To this end we plan to build a specialized purpose web application to ensure the
effectiveness and efficiency of the process. The category of semantic annotations with an identifier of concepts and semantic categories derived from ontology fall outside this study. In building the corpus our objective is to consider that the corpus should be useful for the broader research community and other scientometric related research through the development of practical and useful text mining resources. For the first purpose, four characteristics are considered in the corpus i.e. to make them easy to use through a good documentation, a balance in representation, and the availability of data on inter-annotator agreements [1]. Two major issues arise as common problems are the large feature dimensions and the small sample sizes. To overcome this problem pre-processing is carried out which includes tokenization, word frequency filtering, stop-words removal and stemming. Then a feature selection is designed that includes the use of the TF/IDF method as well as knowledge engineering using domain experts. The reduced number of features will bring benefits in speeding up the algorithms, improving learning accuracy, and a better model understanding. Our goal is to have a highly correlated feature subset with the predictive classes, but not correlated with each other.

3.4. Models Building and Selection

Models building and selection were performed on the basis of lexical analysis based on the bag of words method to quantify a term using TF/IDF measurement. For the text classification some models of the supervised machine learning methods are investigated such as k-NN (Nearest Neighbours) which is a simple model suitable for small dimensional data to complex models such as SVM (Support Vector Machines). In this study, three classifiers are used: Naive Bayes classifier which is often used as a baseline study, a k-NN classifier, and a SVM classifier. The Bayesian approach was used as a basis in our study because of the simplicity and the speed of training. Bayesian classification performance is in fact the best in our overall metric evaluation. On the other hand, SVM produces the least accurate classification in our investigation. Previous research has shown that no single learning approach is clearly superior to all cases, and in fact, different learning algorithms often produce similar results [4] especially after careful optimization process. One factor that can have a major impact on the success of learning algorithms is the nature of the data used to characterize the tasks to be learned. If data fails to show the regularity of the exploited statistics of computer learning algorithms, then learning will fail.

3.5. Results Evaluation and Interpretation

The output from the document classification model is a useful resource for a scientometric study. Combining the results with the bibliometric analysis and a visual presentation tool such as VOSviewer will give many perspectives in the analysis. The results obtained from the development of the document classification model will be validated qualitatively using domain expert in a FGD session. For that purpose the data details of the features which is a variable in a classification is provided. Domain experts can provide interpretation of feature-based relationships and presentation-based classification results. The domain experts are put in the position of interpreting the results of a document classification model, whereby they should certainly trust the evidence they have learned about the texts and the features. At the same time they should cultivate open-mindedness about how the results of the model might challenge conventional thinking. Researcher’s domain knowledge is also invaluable for interpreting and contextualizing results in the first place. For the quantitative evaluation of the model, a confusion matrix is used to evaluate a classifier’s performance. Two popular performance measures are precision, which measures if a document is correctly classified, and recall, which measures how well the classifier detects documents in a given category. For the overall evaluation purpose we used the aggregated F-score as a single metric whereby the score achieved in the Naive Bayes model is 98% in classifying BR (Basic Research), 97% in classifying DD (Drug Design), 97.8% in classifying PC (Pre-Clinical), and 100% in classifying CT (Clinical Trial). Because of the limitation of our research, the reported results are only from text mining the abstract of the publications. Some studies show that articles in full-text form give different results to those in abstract form whereby large amounts of information or data are not found in abstracts, but only in the full-text
of f, or even in tables and captions [13]. For a broader purpose various other methods such as clustering and topic modelling can be applied to support a broader perspective in scientometrics and as a tool for evaluating and constructing interpretations of the results. A document clustering model is evaluated by looking at how the cluster model explains variation in the document-term matrix compared to treating the documents as one cluster. Re-clustering is possible as a results of not optimal labelling by the domain expert.

4. Conclusion
This study proposes a framework for doing text mining in a scientometric study. From the experience in developing a document classification model as a proxy to its downstreaming potential, a lot of deviations occurs in the process from the initial planning. This is due to a lack of a guidance that can be used in performing the various tasks required. Therefore a framework is needed to conduct a scientometric study based on text-mining. The experience shows that the development of the corpus requires considerable effort. Based on our experience and the success in developing the document classification model for the purpose of a scientometric study, this framework can guide for the effective conduct. This framework for doing research in scientometrics with text mining relies heavily on the feedback of the domain experts in constructing conceptual models, labelling, validating and interpreting the results. In order to reduce the dependency, large data is required to enable the machine to do the learning, but mostly the amount of data that can be collected is not large so that it becomes a limitation. The framework development will be further enhanced through case study research in the different domain of interests and also through exploring the text mining in the full-text form of document, and the use of semantically richer methods. The analysis used in this study will also provide new knowledge for scientometric research in Indonesia.

References
[1] Cohen B, Fox L, Ogren P and Hunter L 2005 Corpus Design for biomedical natural language processing, ACL-ISMB workshop on Linking Biological Literature Ontologies and Databases.
[2] De Miranda Santo M, Coelho G M, dos Santos D M, and Fellows Filho L 2006 Text mining as a valuable tool in foresight exercises: A study on nanotechnology. Technological Forecasting and Social Change, 73 8 1013-27
[3] FDA 2017 The Drug Development Process http://www.fda.gov/ForPatients/Approvals/Drugs/default.htm. Accesed on 11 January 2017
[4] Hall M A 1999 Correlation-based feature selection for machine learning
[5] Hess D J 1997 Science Studies: An advanced introduction. New York: New York University Press
[6] Huang S, Zhou J, Wang Z, Ling Q and Shen Y 2017 Biomedical informatics with optimization and machine learning
[7] Hughes J P, Rees S S, Kalindjian S B and Philpott K L 2011 Principles of early drug discovery Br. J. Pharmacol. 162 1239–49
[8] Lukman B 2012 Analisis Bibliometrika Berdasarkan Pendekatan Co-words: Pemetaan Laporan Hasil Penelitian Pusat Penelitian Biologi. Bogor: Institut Pertanian Bogor
[9] Nadhiroh I, Amelia M, Handayani T, Hardiyati R and Simamora N 2015 Analisis Jaringan Ko-authorship Publikasi Ilmiah Internasional Indonesia dalam Penguatan Iptek Masa Depan (Bidang Energi, Pangan, dan Kelautan) Jakarta: Pusat Penelitian Perkembangan Iptek
[10] Našimov V V, Mulchenko Z M 1969 Naukometriya. Izuchenie Razvitiya Nauki kak Informatsionnogo Protessa Scientometrics Study of the Development of Science as an Information Process, Nauka, Moscow, (English translation: 1971. Washington, D.C.: Foreign Technology Division. U.S. Air Force Systems Command Wright-Patterson AFB Ohio NTIS Report No.AD 735- 634
[11] Patumcharoenpol P, Doungpan N, Meechai A, Shen B, Chan J H and Vongsangnak W 2016 An integrated text mining framework for metabolic interaction network reconstruction PeerJ. 4 e1811

[12] Saini P K 2014 Application of Scientometric Analysis in Library Net Work: A Comprehensive Study. International Journal of Emerging Research in Management & Technology ISSN: 2278-935

[13] Shatkay H and Feldman R 2003 Mining the biomedical literature in the genomic era: an overview. Journal of computational biology 10 6 821-855

[14] Van Raan A 1997 Scientometrics: State of The Art. Scientometrics 205-218

[15] Wirth R and Hipp J 2000 CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining 29-39