UNIVERSITAS KATOLIK DE LA SALLE MANADO
FAKULTAS TEKNIK
Alamat : Kombos II (Belakang Wenang Permai II) - Manado
Telp. (0431) 871971, 877442. Fax (0431) 871972

SURAT TUGAS

No: 221a/ST/A/D.FT/II/2014

Sehubungan dengan pemenuhan Tri Dharma di bidang Penelitian Dosen Universitas Katolik De La Salle Manado Semester Genap tahun akademik 2013/2014, maka bersama ini kami menugaskan:

Dosen	***NIDN***	***Fakultas/Program Studi***	***Judul Penelitian***
Liza Wikarsa, BCS.,M.Comp	0912047902	Teknik/Teknik Informatika	Steps in Text Mining Using Classification Techniques

Demikianlah penyampaian kami, atas perhatian diucapkan terima kasih.

Manado, 19 Februari 2014

Debby Paseru, ST., MMSI., MEd
Dekan Fakultas Teknik

Steps in Text Mining Using Classification Methods

Liza Wikarsa, BCS, MComp
Program Studi Teknik Informatika
Universitas Katolik De La Salle Manado

Abstract
Text mining is useful for extracting logical patterns and rules from the unstructured data. However, the process can be complicated due to the complexity and amount of data itself. There are various classification methods to choose from, depending on the goals at hand. Hence, there is a need to have steps in text mining using classification methods that is customary to provide valuable knowledge for better decision makings by identifying issues pertinent to data collection, cleaning, and pre-processing that can help to avoid overfitting.

Abstrak
Text mining sangat berguna untuk mengekstrasi pola-pola logika dan aturan-aturan dari data yang tidak terstruktur. Akan tetapi, prosesnya dapat menjadi rumit karena kompleskitas dan jumlah datanya. Ada beberapa metode klasifikasi yang dapat dipilih, tergantung dari tujuan yang ingin dicapai. Oleh karena itu, diperlukan langkah-langkah dalam pengerjaan *text mining* dengan menggunakan metode-metode klasifikasi yang dapat dikustomisasi untuk menyediakan pengetahuan berharga dalam pengambilan keputusan melalui identifikasi isu-isu yang berhubungan pengumpulan data, pembersihan, dan pra-proses sehingga mampu menghindari *overfitting*.

Keywords: Steps, Text mining, Classification, and Methods

1. Introduction

Text mining is an extension of data mining that uses knowledge discovery techniques to extract logical patterns and rules from unstructured data (Pritam and Huan 2012, and Shmueli *et al.* 2010). It involves methods at the intersection of information retrieval, text-analysis, natural language processing, and information classification based on large amount of data (Ghosh *et al.* 2012, Hotho *et al.* 2005, and Mehmed 2011). Text mining is used in a wide variety of fields and applications such as in military, medical research, business, and more.

The most compelling factor for the rapid growth of text mining is the growth of data that "is driven not simply by an expanding economy and knowledge base but by the decreasing cost and increasing availability of automatic data capture mechanism" (Shumeli *et al.* 2010: 5). The Internet, for an example, has captured a huge quantity of information that enables companies to shift their focuses from products and services to the needs of their customers based on individual transactions. Unfortunately, the operational database supporting the routine business activity can only handle

simple queries but is not able to perform more complete and aggregate data analysis by extracting useful information from large data sets. Hence, this is considered as a serious problem for companies because data exploration and analysis are highly required to discover meaningful patterns and rules.

The text mining techniques are more complicated and complex than data mining due to unstructured and fuzzy nature of data. In this regard, there are various text mining techniques and approaches that can be used such as classification (supervised) and clustering (unsupervised) (Mehmed 2011, Ringel *et al.* 2010, and Tekiner *et al.* 2009). This research particularly defines the classification method as "the process of learning a set of rules from a set of examples in a training set. Text classification is a mining method that classifies each text to a certain category" (Irfan *et al.* 2004: 5). There are two categories for this classification that are machine learning based text classification (MLTC) and ontology based text classification.

Due to the complexity and paramount of data to deal with, there is a need to have a better understanding of the problem before getting into the details of algorithms to be used. It is strongly urged to employ a list of steps in text mining, particularly in classification methods, that is customary to provide valuable knowledge for decision support. Having said that, it is also urged to identify issues pertinent to data collection, cleaning, and pre-processing that can help to avoid overfitting.

2. Statement of Purpose

The purpose of this research is to enlist steps in text mining using classification methods that is customary to provide valuable knowledge for better decision makings by identifying issues pertinent to data collection, cleaning, and pre-processing that can help to avoid overfitting.

3. Research Objectives

1. Enlisting steps in text mining using classification methods that are customary to provide valuable knowledge for better decision makings.
2. Discussing issues related to data collection, cleaning, and pre-processing that can help avoid to overfitting.

4. Research Question

What are steps in text mining using classification methods to provide valuable knowledge for better decision makings by identifying issues pertinent to data collection, cleaning, and pre-processing that can help to avoid overfitting?

5. Literature Review

5.1 Definition of Text mining

It is arguable that there is no universally agreed definition of text mining partly because it is being used by different communities for different purposes. Nevertheless, Hotho *et al.* (2005) explains that text mining is generally the process of extracting interesting information and knowledge from unstructured text. They also point out that text mining is interdisciplinary methods used to draw "on information retrieval, machine learning, statistics, computational linguistics and especially data mining" (2005: 2).

In support of this, Clark (2013) further adds that text mining is the process of extracting meaning from text in the form of concepts, the relationships between the concepts or the actions performed on them and presents them as facts or assertions. It is believed that text mining essentially corresponds to the extraction of information that is clearly and explicitly stated in the text (Bhushan *et al.* 2014). This study therefore defines text mining as the extraction of logical patterns and rules from unstructured data to gain valuable knowledge for different purposes.

5.2 Classification Methods

As mentioned above, classification methods or supervised learning are widely used to classify each text to a certain category based on the learning process done on a set of rules from a set of examples in a training set. Korde and Mahender define classification as "a set of logical rules that convert expert knowledge on how to classify texts under the given set of categories" (2010: 85).

5.3 Classification Algorithms

Classification can be further divided into two categories that are machine learning based text classification and ontology based text classification (Irfan *et al.* 2014). Each classification category leads to different algorithms that can result in varied performances. Algorithm consists of specific procedures used to implement a particular technique such as classification tree, discriminant analysis, and the like.

According to Ghosh *et al.*, classification algorithms are useful for discovering a model for the class in terms of the remaining attributes. Having said that, we need "the training data set to build a model of the class label based on the other attributes such that the model can be used to classify new data not from the training data set" (2012: 223). The following figure depicts different algorithms to choose from.

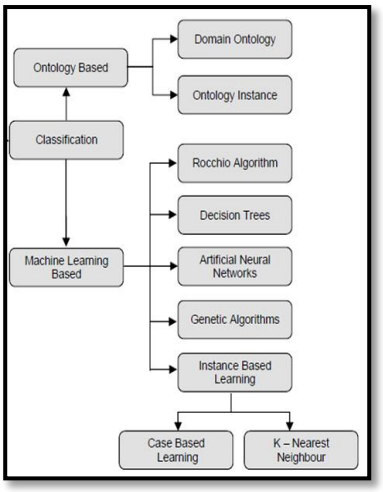

Figure 1: Classification Algorithms in Text Mining (Irfan *et al.* 2014, Hotho *et al.* 2005, Shmueli *et al*. 2010)

Regardless what classification algorithms used, it is important to have clear goal definitions that lead to suitable model deployment. Shumeli *et al.* (2010) explain that model is an algorithm as applied to a dataset, complete with its settings (many of the algorithms have parameters that the user can adjust). In order to do so, we need to have steps in text mining using classification techniques that are customary for exploring data and building models. We can thus select the one that is most useful for better decision makings.

6. Discussion

For the past few years there has been a lot of research in the area of text mining, especially the classification techniques, for various purposes (Korde *et al.* 2010, Shumeli *et al.* 2010, Dalal and Zaveri 2011, Ghosh *et al.* 2012, Bhushan *et al.* 2014, and Irfan *et al.* 2014). The text mining techniques are complicated and complex due to unstructured and fuzzy nature of data. Hence, there is a need to have a list of steps in doing text classification that is customary to meet the goal at hand. This study divides the steps into three different phases that consist of numerous steps to serve different purposes accordingly. The three phases are i) pre-processing, ii) model selection, and iii) result validations.

6.1 Phase I: Pre-processing

6.1.1 Objectives

1. To collect data from the database(s).
2. To present data in clear, structured, and consistent format to be used in the analysis.

6.1.2 Outputs

Training dataset, validation dataset, and test dataset.

6.1.2 Steps

Step 1.1: Develop an understanding of the purpose of the text mining project

Step 1.2: Assemble a target data set to be used in the analysis.

Step 1.3: Explore, clean, and preprocess the data

Pre-processing may consume considerable processing time but it can guarantee successful implementation of data exploration and analysis. There are two basic methods of text pre-processing to use such as feature extraction and feature selection. Irfan *et al.* (2014: 3) further categorize feature extraction into 'Morphological Analysis (*MA*), (b) Syntactical Analysis (*SA*), and (c) Semantic Analysis (*SA*)'. MA basically deals with individual words represented in a text and mainly consists of tokenization, removing stop words, and stemming word (Korde and Mahender 2012: 86, Dalal and Zaveri 2011).

- *Tokenization:* A document is treated as a string, and then partitioned into a list of tokens.
- *Removing stop words*: Stop words such as "the", "a", "and", and others are frequently occurring, so the insignificant words need to be removed.
- *Stemming word:* Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form, e.g. implementation to implement, calculating to calculate, working to work

Irfan *et al.* also explains that Syntactical Analysis "provides knowledge about the grammatical structure of a language that is often termed as syntax. For instance, the English language comprises of noun, verb, adverb, punctuation, and other parts of speech" (2014: 3). On the other hand, Semantic Analysis is used to find concepts, events, and relationships between them.

Feature selection is useful for eliminating irrelevant and redundant information from the target text (Dalal and Zaveri 2011). This can be done by scoring the words. Feature selection has three methods to choose from such as frequency based feature selection, latent semantic indexing (LSI), and random mapping.

The following diagram will depict the pre-processing using feature extraction and feature selection.

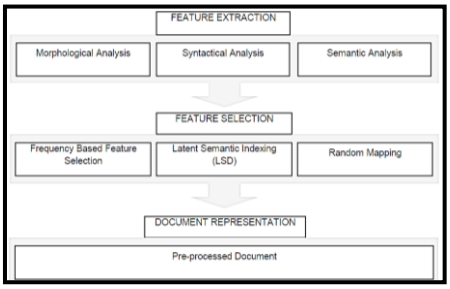

Figure 2: Pre-processing Using Feature Extraction and Feature Selection (Dalal and Zaveri 2011, Korde and Mahender 2012, Irfan *et al.* 2014)

The next thing to do is to analyze the multivariate data sets. Then, do clean the target set by removing the observations containing noise and those with missing data. Review and examine the data to see what messages they hold; records might be aggregated into groups of similar records. Lastly, it is recommended to use graphical analysis by looking at each variable separately as well as looking at relationship between variables.

Step 1.4: Reduce the data, if necessary, and (where supervised training is involved) separate them into training, validation, and test datasets as shown below.

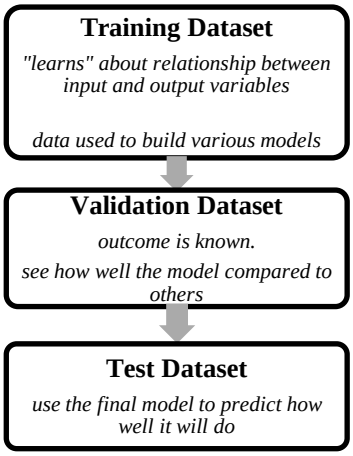

Figure 3: Datasets for Text Mining

6.1.4 Issues and Solutions to Pre-Processing

There are several issues pertinent to pre-processing including:

- *classifying unstructured data*

It is important to pay attention to the types of variables used which can be numerical or text (character). They can be continuous, integer, or categorical. Categorical variables are basically unordered (nominal variables) like North America, Europe, Asia, and others. Meanwhile, ordinal variables are such as high value, low value, and nil value. Having said that, categorical variables must be decomposed into a series of dummy binary variables.

Examples:
Student – Yes/No;
Unemployed – Yes/No;
Employed – Yes/No;
Retired – Yes/No

Variable selection is useful for finding the relationship between Y and a single predictor variable X. While doing that, we must try to avoid overfitting or overreached because we might somehow mislabel the noise in the data as it were a signal. Other words, we ended up "explaining" some variation in the data that was nothing more than chance variation.

- *handing large number of data*

Another question often asked when doing text mining is "how many variables and how much data?". According to Shmueli *et al.* (2010), for classification is to have at least 6 x m x p records (m= the number of outcome classes; p=

the number of variables). However, these numbers may vary depending on the goals at hand.

Do not forget to look for missing values. If the number of records with missing values is small, those records might be omitted. However, if we have a large number of variables, even a small proportion of missing values can affect a lot of records.

The last thing to do is normalizing (standardizing) data. This can be done by subtracting the mean from each value and divided by the standard deviation of the resulting deviations from the mean.

6.2 Phase II: Model Selection

6.2.1 Objectives

1. To learn the process for the training data set.
2. To build various models that are to be examined.
3. To assess the performance of each model by comparing models and pick the best one.

6.2.2 Outputs

Several models and their performances.

6.2.3 Steps

***Step 2.1*:** Choose the classification algorithms to be used (K-nearest neighbors, Naïve Bayes, classification trees, logistic regression, neural nets, or discriminant analysis).

***Step 2.2*:** Use algorithms to perform the tasks.

This is typically an iterative process – trying multiple variants, and often using multiple variants of the same algorithm (choosing different variables or setting within the algorithm). Where appropriate, feedback from the algorithm's performance on the validation data is used to refine the setting. The following steps can be used when required to assess performance of the chosen model with new data.

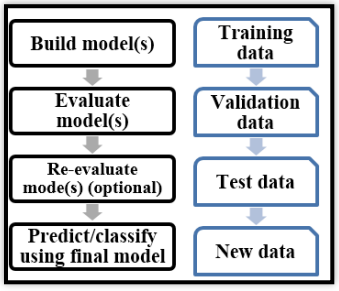

Figure 4: Building and Selecting Model(s)

6.2.4 Issues and Solutions to Model Selection

No single method is found to be superior to all others for all types of classification. Hence, there is a need to experiment with more classification algorithms in order to achieve better classification results.

6.3 Phase III: Result Validations

6.3.1 Objectives

To show validated patterns produced by the selected text mining algorithms in the wider data set.

6.3.2 Outputs

Model deployment.

6.3.3 Steps

***Step 3.1*:** Perform proper statistical hypothesis testing.

***Step 3.2*:** Verify that the patterns produced by the text mining algorithms occur in the wider data set.

Step 3.3: Interpret the results of the algorithms

This involves making a choice as to the best algorithm to deploy, and where possible, testing the final choice on the test data to get an idea as to how well it will perform. (Recall that each algorithm may also be tested on the validation data for tuning purposes; in this way the validation data become a part of the fitting process and are likely to underestimate the error in the deployment of the model that is finally chosen).

***Step 3.4*:** Deploy the model

This involves integrating the model into operational systems and running it on real records to produce decisions or actions.

6.3.4 Issues and Solutions to Result Validations

Using the final model to see how well it do upon the test data. The performance of a classification algorithm is significantly affected by the quality of data source. Irrelevant and redundant features of data not only increase the cost of text mining process, but also reduce the quality of the result in some cases. Each algorithm has its own advantages and disadvantages with their time complexity.

7. Summary

1. There are three phases of text mining using classification methods that are pre-processing, model selection, and result validations.
2. Pre-processing phase is mainly to gather, explore, clean, and pre-process data that are to be used in analysis. Feature extraction and feature selection can be employed for pre-processing. The outputs of this phase include training dataset, validation dataset, and test dataset. Issues to consider in this phase are such as classifying unstructured data and handing large number of data.
3. In the Model Selection phase, there are a few models made based on the learning process from the training set. Each model will be thus examined and compared its performance through iterative process. Based on the feedback gathered from the testings, a model is picked and refined using validation data. No single method is found to be superior to all others for all types of classification. Hence, there is a need to experiment with

more classification algorithms in order to achieve better classification results.

4. In Result Validations phase, the patterns produced by the text mining algorithms occur in the wider data set is carefully verified. The performance feedback is important to ensure the quality of the results while considering the cost of text mining process.

8. Recommendations

There are number recommendations made for this study as follows:

1. To create a more thorough framework for each phase of text mining that can be used by both classification and clustering techniques.
2. To provide working templates and deliverable checklists required for steps designed in each phase of text mining to ensure that deliverable(s) of each step is of high quality, reliable, and manageable.
3. To provide step-by-step guidance in building and selecting models using various algorithms and tools.

9. References

1. Bhushan, J. G., Pushkar, U. W., Shivaji, P. K., and Nikhil, V. K. (April 2014) 'Searching Research Papers Using Clustering and Text Mining.' *International Journal of Emerging Technology and Advanced Engineering* 4, (4) 788-791.
2. Clark, J. (2013) *Text Mining and Scholarly Publishing*. Netherlands: Publishing Research Consortium.
3. Dalal, M.K., and Zaveri, M. A. (August 2011) 'Automatic Text Classification: A Technical Review'. *International Journal of Computer Applications* 28, (2) 37-40.
4. Ghosh, S., Roy, S., and Bandyopadhyay, S.K. (June 2012) 'A tutorial review on Text Mining Algorithms'. *International Journal of Advanced Research in Computer and Communication Engineering* 1, (4) 223-233.
5. Hotho, A., Nurnberger, A., and Paab, G. (2005) 'A Brief Survey of Text Mining' [online]. Available from *<www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf* > [5 August 2014].
6. Irfan, R., King, C. K, Grages, D., Ewen, S., Khan, S.U., Madani, S.A., Kolodziej, J., Wang, L., Chen, D., Rayes, A.R., Tziritas, N., Zu, C.Z., Zomaya, A.Y., Alzahrani, A.S., and Li., H.Z. (July 2014) 'A Survey On Text Mining in Social Network'. *The Knowledge Engineering Review* 0, 1-24.
7. Korde, V., and Mahender, N. C. (March 2010) 'Text Classification and Classifiers: A Survey'. *International Journal of Artificial Intelligence & Applications (IJAIA)* 3, (2) 85-99.
8. Mehmed, K. (2011) 2nd edn. *Data Mining: Concepts, Models, Methods, and Algorithms*. Canada: John Willy & Sons.
9. Pritam, G. and Huan, L. (2012) 'Mining Social Media: A Brief Introduction' *Informs* [online]. Available from <http://dx.doi.org/10.1287/educ.1120.0105>[25 March 2014].
10. Ringel, M. M., Teevan, J., and Panovich, K. (2010) 'What Do People Ask Their Social Networks, and Why: A Survey Study of Status Message Question & Answer Behavior'. *Proceedings of International Conference on Human Factors in Computing Systems (CHI 10)*, Atlanta, GA, USA, 56-62.
11. Shmueli, G., Patel, N. R., and Bruce, P.C. (2010) *Data Mining for Business Intelligene: Concepts, techniques, and Applications in Microsoft Office Excel with XLMiner*. New Jersey: John Wiley & Sons, Inc.
12. Tekiner, F., Aanaiadou, S., Tsuruoka, Y., and Tsuji, J. (2009) 'Highly Scalable Text Mining Parallel Tagging Application'. *Proceedings of IEEE 5th International Conference on Soft Computing, Computing with Words and Perception in System Analysis, Decision and Control (ICSCCW)*, China, 1-4.