A COMPARATIVE STUDY OF STOCHASTIC CLUSTERING TECHNIQUES IN HARVESTING EMERGING TRENDS FROM SOCIAL DATA

MOHD. SAFAR A/L SHARIFF

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS*

JUDUL: A COMPARATIVE STUDY OF STOCHASTIC CLUSTERING TECHNIQUES IN HARVESTING EMERGING TRENDS FROM SOCIAL DATA

SESI PENGAJIAN: 2014/2015

Saya MOHAMMAD SAFAR A/L SHARIFF MOHAMMED .

(HURUF BESAR)

mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

____	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
____	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
/	TIDAK TERHAD	

(TANDATANGAN PENULIS)

Alamat tetap: No C-3-4, Pangsapuri Taman Tasik Utama, Jalan Tasik Utama 61, Ayer Keroh, 75450 Melaka, Melaka.

Tarikh: ____________________

(TANDATANGAN PENYELIA)

Prof Madya Dr Azah Kamilah Binti Draman @ Muda

Nama Penyelia

Tarikh: ____________________

CATATAN: * Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)
** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

A COMPARATIVE STUDY OF STOCHASTIC CLUSTERING TECHNIQUES IN HARVESTING EMERGING TRENDS FROM SOCIAL DATA

MOHD. SAFAR A/L SHARIFF

This report is submitted in partial fulfillment of the requirements for the

Bachelor of Computer Science (Artificial Intelligence)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2015

DECLARATION

I hereby declare that this project report entitled

A COMPARATIVE STUDY OF STOHASTIC CLUSTERING TECHNIQUES IN HARVESTING EMERGING TRENDS FROM SOCIAL DATA

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : Date: 21/8/2105

(MOHD SAFAR A/L SHARIFF)

SUPERVISOR : Date: 21/8/2015

(PROF MADYA DR AZAH KAMILAH BINTI DRAMAN @ MUDA)

DEDICATION

I dedicate the work and effort poured into this project to my family and friends. A special thanks goes to my mother whom has single-handedly raised and encouraged me, Devaki A/P Gengatharan.

I also dedicate this current project to my supervisor, Prof Madya Dr Azah Kamilah Binti Draman @ Muda for her guidance, knowledge and advice.

ACKNOWLEDGEMENTS

Apart from the efforts contributed by myself, the success of any project relies highly on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project. I would like to thank my parents for providing me with the opportunity to be where I am. Without them, none of this would be possible. I would also like to thank my friends for their encouragement, input and constructive criticism which are priceless and also for their moral support.

I would like to express my sincere gratitude to my supervisor, Prof Madya Dr Azah Kamilah Binti Draman @ Muda for the continuous support and guidance in completing this project and research. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for completion of my final year project.

Also, I would like to thank all my educators in bachelor degree, for their perceptiveness, understanding, and their skillful teaching. And above all, many thanks to Allah, who helps me get all required resources for completing this research.

ABSTRACT

Social media is an emerging area of interest and attracts many researchers as the latter provide tremendous amounts of data readily available that could be exploited for various reasons such as social networking, decision making and marketing. An emerging field in the area of social media data mining is known as topics detection from social data. A topic is harvested from social data by using an unsupervised machine learning task also known as clustering to group similar social data and recognize the importance of the grouped social data to provide a general distinction which will be known as topic. In this work, two clustering techniques known as the hierarchical clustering technique and density based clustering technique which shares stochastic capability is compared using dataset retrieved from Twitter which corresponds to real world events. The classes present in the dataset is restricted to four main topics and the dataset is then used to test the performance of the clustering algorithms. The performance evaluation being used to evaluate the clustering performance of the clustering algorithms are the V-measure which is the harmonic mean of homogeneity and completeness score of the clustering performance of a clustering algorithm. The results shows that the hierarchical clustering technique outperforms the density based clustering technique in determining the correct number of clusters and assigning the data to their respective clusters reliably. Apart from the comparative studies discussed in this project, an analysis tool based on social data is developed to address the problems related to the social data analysis.

ABSTRAK

Media sosial adalah salah satu bidang yang semakin menarik perhatian ramai penyelidik memandangkan media sosial dapat menyediakan bilangan data sosial dalam skala besar yang boleh dieksploitasi untuk berbagai sebab contohnya jaringan sosial, proses membuat keputusan dan pemasaran. Salah satu bidang membangun yang berkait rapat dengan bidang perlombongan data media sosial adalah pengecaman topik melalui data sosial. Sesebuah topik dikesan dalam data sosial dengan menggunakan kaedah pembelajaran mesin tanpa pengawasan juga dikenali sebagai pengelompokan untuk mengumpul data sosial yang serupa dan mengenalpasti kepentingan data sosial yang telah dikumpul untuk menyediakan perbezaan umum antara kelompok data sosial yang akan dikenali sebagai topik. Dalam hasil kerja ini, dua teknik pengelompokan iaitu teknik pengelompokan hierarki dan teknik pengelompokan berdasarkan kepadatan yang mempunyai ciri stokastik akan dibandingkan dengan menggunakan set data yang diperolehi daripada Twitter yang berkaitan dengan acara masa kini. Kelas yang wujud dalam set data dibataskan kepada hanya empat jenis peristiwa dan set data tersebut akan digunakan untuk menguji prestasi teknik pengelompokan yang digunakan. Kaedah penilaian yang digunakan untuk menilai prestasi pengelompokan teknik pengelompokan adalah ukuran-V yang juga merupakan min harmonik di antara skor kehomogenan dan kesempurnaan teknik pengelompokan yang digunakan. Hasil kerja menunjukkan bahawa teknik pengelompokan hierarki mengatasi teknik pengelompokan berdasarkan kepadatan dalam menentukan jumlah kelompok yang betul dan memberi kedudukan yang betul kepada data sosial dengan kelompok yang bersesuaian. Selain daripada kajian perbandingan yang dibincangkan dalam projek ini, sebuah alat analisa berdasarkan data sosial dibangunkan untuk menangani masalah-masalah yang berkaitan dengan analisa data sosial.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF GRAPHS

LIST OF FIGURES

LIST OF ABBREVIATIONS

API	**-**	**Application Programming Interface**
GST	**-**	**Good and Service Tax**
IDF	**-**	**Inverse Document Frequency**
SEO	**-**	**Search Engine Optimization**
TF	**-**	**Term Frequency**
TF-IDF	**-**	**Term Frequency-Inverse Document Frequency**

CHAPTER I

INTRODUCTION

1.1. Project Background

Social media has become an emerging area of extensive and intensive research as it provides tremendous data readily available that could be exploited for various reasons such as for social network, decision making and social marketing. Increase of social media sites usage such as Twitter, Youtube and Facebook due to their popularity provides a huge amount of user-generated data available. “Social media has been broadly defined to refer to 'the many relatively inexpensive and widely accessible electronic tools that enable anyone to publish and access information, collaborate on a common effort, or build relationships (Murthy, 2013)”. This opens up new opportunity in harvesting data to contribute in certain fields that are focused on people.

One of the emerging field that exploits data from social media is social media mining. Social media mining is a process to represent, analyze, and extract significant patterns from data retrieved from social media. Social media mining showcases basic concepts and algorithms for investigation of massive data taken from social media. It encompasses useful and suitable tools to formally and directly representing, measuring, modeling, and

mining meaningful patterns from big-scale data taken from social media (Zafarani, et al., 2014). A part of social media mining subfield which uses data from social media is topic detection which generates topics from social data that is currently trending. In this current era of Big Data, data is generated rapidly, in a massive amount with various attributes. This additional attributes were regarded as trivial or insignificant to reduce the volume of data, increase the velocity of accessing data and reduce the variety of data until now. There are many kinds of applications emerging that are tackling the problems that large-scale data introduced.

The additional information provided by users is also as important in order to make good use of the opinion analysis such as time zone and unique id. Twitter is such an application that provides such additional information. Although all the data extracted is crucial in decision support system or knowledge acquisition, the opinion or 'post' given by the user is the main focus in this work. Topic detection is performed on the sentence level. The analysis performed is done on each sentence to determine the similarity between each of the sentence. The sentences is then grouped into clusters based on the core topics of the corresponding sentence. The rise in machine learning tasks has enabled a vast improvements in certain area of interest such as intelligent system and decision support system.

The main focus of this work is to introduce a stochastic unsupervised classification that is capable to mine useful and significant pattern from the data retrieved from social media without any initialization. The technique of interest used in this area of research are hierarchical clustering (Agarwal & Zhai, 2013). The benefits of using agglomerative clustering is the ability of the algorithm to generate its own number of cluster which is crucial in the development of topic detection as the number of clusters is unknown. Another technique being introduced that shares the stochastic nature of hierarchical

clustering algorithm which is density-based clustering (Agarwal & Zhai, 2013). Density based clustering algorithm is used as a comparison technique for the proposed technique.

Besides that, an application utilizing the potential of the research phase is important in order to determine the advantages of the application being developed. Hence, a decision support application is developed as well to provide an overview of emerging trends by mining social data. The application should be able to detect emerging trends and fully utilize the various attributes corresponding to the data mined from social media.

The application being developed provides classification and visualization tools that provides analysis that can be used in various fields such as decision making, social marketing and issues tracking.

1.2. Problem Statements

- There are many kinds of research being done on opinion mining particularly in detecting topics of interest. However, the work which uses unsupervised learning is always considered trivial and is now emerging as competitive as supervised classification.

- There are many kind of unsupervised learning techniques which supports clustering, however the stochastic nature of unsupervised learning only exists in a handful of techniques.
- Topic detection by unsupervised learning is a field still in its infancy which requires a proper implementation and development. It requires more extensive and intensive research to prosper in this current era of Big Data

- There are several commercial implementation on topic detection which is available on the open market. However, there could be some applications which is costly and not fully reliable.
- There is restricted information regarding topic detection due to the fact that user opinion or consensus is considered a part of user privacy which deters the full development and utilization of sentiment analysis implementation.

1.3. Objective

- To search for the best methodology to be used in trends detection from social data particularly in area of unsupervised classification which has stochastic traits.
- To determine the reliability of stochastic nature in the proposed clustering technique being used in ensuring the clustering is performed optimally.
- To compare unsupervised learning techniques which are capable of stochastically clustering social data in a text domain.
- To prepare an open source commercial application that provides analysis on social data which includes classification of current trends.

1.4. Scope

There are several scopes which is reflected in this area of research. First of all, only unsupervised learning techniques which is clustering techniques is used in the development of this project. Next, the scale of dataset used in this project is restricted to only certain language which is English. The purpose of this project is to determine the relationship between each data retrieved from social media and generate clusters that contains data that are similar to each other. Therefore, it is crucial that preprocessing is done to speed up the processing duration and result accuracy. Other matters such as impact of implementation and cost involved are taken into consideration. The project is developed using open source tools that are suitable to be used in the development of the project and are also robust to ease the development of the project.

1.5. Project Significance

This project could illuminate researchers in the process of decision making using machine learning techniques that are highly dependent on supervised learning.

Supervised learning has been admitted as the most favorable learning techniques to be used in the development of a predictive or classification model. However, this technique requires readily available data to be used in the development of the model. The shortcomings of this technique is the model might not be robust to handle data that are different from the data that is used in the training of a supervised model. Several researchers has overcome this problem by introducing the ability for the model to generalize however, generalizing

might not be as accurate as needed in predicting or classifying the result given the missing data.

Unsupervised learning is the main center of attention in this project as it does not require training and it tries to find structure or pattern that is present in the data provided at a time and is not biased by the previous data. This is highly important as data at different scale of time might not be related and might cause extreme bias in a supervised learning model. Unsupervised learning is also important in terms of detecting noise present in data since it provides an overview of the pattern and unusual patterns or anomaly can be regarded as noises existing in the data.

Besides that commercial implementation from various sources that performs data mining on social data to determine trending issues may provide a good platform to track current trends, however there are several commercial implementation which might involve costs and the methods being deployed by those commercial implementation might not be transparent to the public which does not promise reliability and soundness of information provided. Hence, it is important to deploy a proper framework that encompasses the best in classification and make it available to the public for various purposes such as marketing strategy, issue tracking and decision support system.

1.6. Expected Output

- A benchmark in stochastic clustering of social data that could be used for future works and research.
- A robust and reliable model of clustering social data that produces emerging topics that is considered as trending or critical.

- A decision support system that provides a vast variety of analysis taken from the social data that further enhances certain fields such as decision making, social marketing and issue tracking.

1.7. Conclusion

As the project is concerned with harvesting topics of interest from social data, the development is mainly centered on clustering the data taken from social media and certain attributes are taken from the clusters to generate possible topics of interest. The topics taken from those clusters is considered as trending, critical or informative. The topics assigned to each clusters is user-defined, however the decision of topic naming is helped by the list of topics generated. This will in turn help users in several decision making process in real time.

CHAPTER II

LITERATURE REVIEW

2.1. Introduction

In this research, the information about the theories and techniques that is used in this particular field is obtained through journals, books and the Internet. Journals are provided by various sources such as IEEE Xplore Digital Library (IEEE, 2015) and ScienceDirect (Elsevier, 2015). Books are readily available in UTeM library whereas other useful information or data are extracted from various sources from the Internet.

Throughout the literature review, there has been a lot of useful experience, enlightenment and lessons that is gained technically and educationally. The process of finding facts in literature review is a crucial and tedious process as the resource available is either limited, unavailable or expensive. Fortunately, many progress has been made in this field of research which could help facilitate this research.

2.2.Facts and Findings

Facts and findings is a central part of understanding the flow and directions of this research. There has been various types of research being done in terms of social media, text feature weighting and clustering techniques that can be utilized in this research.

2.2.1. Domain

The domain that is related to current work is in the field of social media. Social media provides a large amount of data that could be used for purpose of research and pave new foundation in knowledge discovery and data mining field. Besides that, the domain of current work is bound to unsupervised learning. Supervised learning has been a well-known field favored by many researchers in the data mining and machine learning task. However, unsupervised learning model provides several benefits over supervised learning model such as possibility to learn larger and more complex data of any kind.

The domain in current work is also related with the text processing and extraction of features from textual data. Text processing is an emerging field in machine learning task which pose several challenges. The text processing domain is strictly in the range of statistical model of textual data since the amount of data in current work is large. Finally, the unsupervised learning model being deployed in current work which is clustering has several techniques which could be applied in determining the groups associated with the social data. In current work, only the hierarchical and density based clustering technique is considered as the promising techniques because of the latter stochastic nature in clustering the social data.