PREDICTING INFLUENCERS FROM TWEETS DATA

NOREZZATI BINTI MD. NOR HAYATI

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

© Universiti Teknikal Malaysia Melaka

BORANG PENGESAHAN STATUS TESIS

JUDUL: PREDICTING INFLUENCERS FROM TWEETS DATA

SESI PENGAJIAN: 2014/2015

Saya NOREZZATI BINTI MD. NOR HAYATI
(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 hak milik 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) (TANDATANGAN PENYELIA)

Alamat tetap: 24 Jalan TU38, Taman Tasik Utama, 75450, Ayer Keroh, Melaka.

PM. DR. CHOO YUN HUOY
Nama Penyelia

Tarikh : _____ Tarikh : _____

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

© Universiti Teknikal Malaysia Melaka

PREDICTING INFLUENCERS FROM TWEETS DATA

NOREZZATI BINTI MD. NOR HAYATI

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

© Universiti Teknikal Malaysia Melaka

DECLARATION

I hereby declare that this project report entitled

PREDICTING INFLUENCERS FROM TWEETS DATA

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

STUDENT : _____ Date : _____
(NOREZZATI BINTI MD. NOR HAYATI)

SUPERVISOR: _____ Date : _____
(PM. DR CHOO YUN HUOY)

© Universiti Teknikal Malaysia Melaka

DEDICATION

To my beloved parents..

© Universiti Teknikal Malaysia Melaka

ACKNOWLEDGEMENT

First of all, I would like to thank my parents for being very supportive. Their supports give me strength to complete my project successfully.

I also would like to thank to my beloved supervisor, PM. Dr. Choo Yun Huoy. Without her support, I will never get through this project successfully. And I appreciate her time for every consultations and progress.

© Universiti Teknikal Malaysia Melaka

ABSTRACT

Twitter generates such overwhelming information every day and allows developers to get the public stream for analytic purposes. Influencer analysis is thus important to provide insight on market segmentation. However, there are not many researches done on identifying influencer because there exist no objective quantification to determine influencing factor. This project aims to find the influential factors and to cluster into the influencer and non-influencer clusters. The experimental was collected from Twitter on #NepalQuake topic. The attribute proposed for influencer prediction includes retweets, favorites, followers and friends. Simple scoring algorithm and Fuzzy C-Means cluster were use in the project. Data crawling was done through Twitter application in Python. Comparison was made to compare the performance of the simple scoring algorithm versus the Fuzzy C-Means cluster in generating influencer list. Both proposed algorithm is able to list and produce the most influential user. The result of influencer was obtained after the weightage is set to each attribute to represent the importance of attributes. The scoring algorithm will list the influential user with their rank. There is only one user for the influencer cluster. This is because only one user has obvious number of retweets, followers, favorites and friends. In future work, the algorithm can be improved to be commercialized. The measuring on how important the attribute need to be made and also the algorithm will be apply to the unstructured data.

© Universiti Teknikal Malaysia Melaka

ABSTRAK

Twitter menghasilkan maklumat yang banyak setiap hari dan membolehkan pemaju untuk mendapatkan maklumat awam untuk tujuan analisis. Walau bagaimanapun, tidak banyak kajian yang dilakukan mengenai pengenalpastian pengaruh dalam *Twitter* kerana tidak wujud kuantifikasi bagi menentukan faktor pengaruh. Projek ini bertujuan untuk mencari faktor-faktor pengaruh dan untuk menentukan pengaruh tersebut kepada kumpulan yang berpengaruh dan bukan pengaruh. Kajian ini telah dikumpulkan daripada *Twitter* menerusi topik *#NepalQuake*. Antara perkara yang dicadangkan untuk meramal pengaruh adalah *'retweet'*, *'favorite'*, *'followers'*, dan *'friends'*. Algoritma pemarkahan dan penggunaan *Fuzzy C-Means* untuk kelompok adalah teknik yang digunakan dalam kajian ini. *Python* digunakan untuk mendapatkan data melalui *Twitter*. Perbandingan telah dibuat untuk membandingkan prestasi algoritma yang dicadangkan berbanding teknik *Fuzzy C-Means* dalam menghasilkan senarai pengaruh. Kedua-dua algoritma yang dicadangkan dapat menyenaraikan pengguna yang paling berpengaruh. Hasil pengaruh telah diperolehi selepas pemberat yang ditetapkan kepada setiap atribut untuk mewakili kepentingan setiap sifat-sifat. Algoritma pemarkahan dapat menyenaraikan pengguna yang berpengaruh dengan kedudukan mereka. Hanya ada satu pengguna bagi kelompok yang berpengaruh. Ini kerana hanya seorang pengguna mempunyai bilangan yang ketara. Bagi kerja-kerja masa hadapan, algoritma boleh diperbaiki untuk dikomersialkan. Algoritma yang dicadangkan juga perlu dikaji supaya dapat bekerjasama denga data yang tidak berstruktur.

© Universiti Teknikal Malaysia Melaka

TABLE OF CONTENTS

© Universiti Teknikal Malaysia Melaka

© Universiti Teknikal Malaysia Melaka

© Universiti Teknikal Malaysia Melaka

LIST OF TABLES

© Universiti Teknikal Malaysia Melaka

LIST OF FIGURES

© Universiti Teknikal Malaysia Melaka

© Universiti Teknikal Malaysia Melaka

LIST OF ABBREVIATIONS

CSV	**-**	**Comma separated values**
XLS	**-**	**Excel file format**
API	**-**	**Application programming interface**
URL	**-**	**Uniform resource locator**
FCM	**-**	**Fuzzy C-Means**

© Universiti Teknikal Malaysia Melaka

LIST OF APPENDICES

© Universiti Teknikal Malaysia Melaka

CHAPTER I

INTRODUCTION

1.1 Project Background

Social media represent how important information being produced and consumed. User generates content in form of comments, tweets and blog posts establish a connection between producers and the consumers of information. The social media users generate such overwhelming volume of information. This kind of information will be useful to the various marketers around the world. In this project, the information is collected specifically from Twitter data. By searching and predicting the influencers in social media data, it will help us to improve the lifestyle, business and also the authorities.

Influencers are the person or authorities that have a power to be a trendsetter, leader, or someone who sets the rule. These people have a big number of followers, and also can affect others to change what they think. Therefore finding influencers are important nowadays because the market, news media, fashion brand or any company always wanted to know their market and how to approach customers. Especially by searching related topics using a hashtags that Twitter had provided.

Twitter is an online social network also known as a microblogging service that allows user to post 140 characters of short messages. It is called 'tweets'. This microblogging service started in 2006. Twitter users will also allow to read, post, comments, and 'retweet' or also known as share the post they interested in. The

© Universiti Teknikal Malaysia Melaka

information is now expanded in large volume as well as the users. The advantage of using Twitter in this project is that Twitter provides an application programming interface (API) to crawl the public information. Twitter data is accessible for public and useful for data analytic purposes.

In this project, few sizes of data were crawled from Twitter. One topic is selected to make a crawl on the information of users. There are four attributes selected to make a crawl on, which is the number of retweets, number of favorites, number of follower, and number of friends. Details about the functionalities will discuss in the next chapter. The approach of this project is explained as in Figure 1.1.

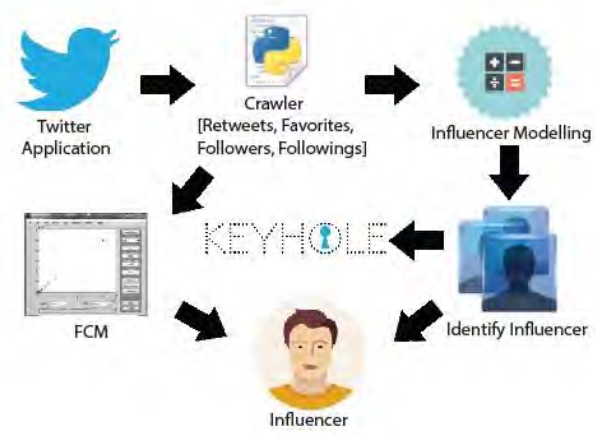

Figure 1.1 : Overall approach for the project

An algorithm is proposed to find the scoring value of the influencers. In addition to that, the influential percentage is made. The scoring is assigned to every data collected and comparison is made. There are several existing tools that can find and predict influential user, therefore these tools will be set as a benchmark for the scoring value and used in comparison phase. More details about the algorithm and implementation will be explained in the following chapter.

Furthermore, the clustering method will be applied in these dataset collected and will compare the result and analysis topic. In this project, fuzzy c-means

© Universiti Teknikal Malaysia Melaka

clustering method is used as the clustering method. The method should be able to cluster the dataset into two clusters which are the influencer group and non-influencer group. This report will explain all steps and technique includes preliminary studies.

1.2 Problem Statement

The problem statement for this project is some authorities or company might want to find the most influential user or a group of influencers in certain topics that relate to their research. There are no objective quantifications for measuring the influencers in Twitter. The measurement of influencer is important and need to find the correct measurement in order to get the most logical and accurate influencer. There are many ways to measure the influencer, but it will return different kind of influencers. For example, the followers count only will return the famous user in Twitter. Thus in this project, the weightage is proposed to each factors of influencers.

1.3 Objectives

The purpose of this research is to create a model that can predict influencer from a big dataset. The objectives of this research are as follow:

a) To identify significant influential factor

   The dataset need to be crawled and. From the information collected, the influential factors need to be identified and justify from the previous studies.

b) To propose a scoring algorithm to quantify influential value of Twitterers

© Universiti Teknikal Malaysia Melaka

- The algorithm should be able to return the value of influencer's score. This algorithm need to add the weightage in order to be fair to all attributes and able to sort the influencer and non-influencer

c) To propose a clustering method to cluster influencers
   - Define one clustering method and cluster the influencer and non-influencers.

1.4 Scope

This project focuses on hashtags topics that exist in Twitter application. Twitter allow public stream extraction for one week duration, thus in this project the chosen topic is '#NepalQuake'. The factor of influencer is selected based on the preliminary studies to do the data crawling and also the proposed technique will add the weightage that is total of 10 is divided to the influential factors. Furthermore, the data will be cluster to two clusters.

1.5 Project Significance

The project significance can be divided into two sub topic which is academic significant and industrial significant.

Academically, the project proposed an algorithm to be used for influencer searching and in future for predicting and decision making purposes.

Industrial significance can be explained for future business decision making. By having this type of model or application, the prediction can be made in various types of data. Otherwise, in future, everybody can predict whether a user is an influencer by only reading a single tweet. This model will help the marketer around

© Universiti Teknikal Malaysia Melaka

the world to improve their business. The company will find their target by finding this influencer and also to check their own market influence.

The authorities will be using this model for future decision making especially if the data is collaborated with the unstructured data. Also the model is useful when the data is getting bigger in future. Furthermore, the model can be useful in terms of disaster issues. Such as nature disaster, this model could give alert to certain rescue to the place needed, or the media company can cover as many news at the same time.

1.6 Expected Output

In the end of this research, the suggested algorithm could be able to describe and determine the influencer from huge set of data collection. This means that the proposed model will capable to work with big size of data or simply called the big data.

On the other hand, the proposed algorithm will display both scoring values and predict the accuracy of the model created. All of the crawled data will be saved into the spreadsheet for monitoring purposes. In addition to that, the clustering method will be able to display a group of influencers and non-influencers. At the end of both techniques, the expected results are that it will produce the same username or same information.

1.7 Summary

As the conclusion to this chapter, the proposed technique will be developed in order to make a prediction. Prediction model will ease the human power to make calculation, observation and intuitions. Therefore this model is useful and after it is

© Universiti Teknikal Malaysia Melaka

specialized in predicting the influencer, it will specifically be more useful to certain authorities to improve their business. The report is organized in the following order, literature review in chapter 2, project methodology and techniques in chapter 3, influencer modelling in chapter 4, results and analysis in chapter 5 and conclusion in chapter 6.

© Universiti Teknikal Malaysia Melaka

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

This chapter will explain the literature review with respect to this project. All the findings and collection of information regarding the topics which include the twitter information will explain in this chapter. Literature review is the phase of finding the previous research and related technical papers that is related to the project. It defines the process of searching related project, collect the information, analyzing the information to related scope, and determine the conclusion. Furthermore, to evaluate other people studies to find the related works and methodology.

In this chapter, the research was made on several related topics. Earlier studies were made on how to measure the influencer or specifically stated as the scoring measurement. The next issue is focused on how to crawl the data from Twitter. More studies were made on the existing tools and comparison of these tools functionalities.

© Universiti Teknikal Malaysia Melaka

2.2 Social Networks

Nowadays, there are too many social networks that allows users to describe their opinions and at the same time share with their friends around the world. The content that all users are sharing is called a social media. Social media is part of the image, blogs, songs, video, or text that user upload in their social networks, (Social Media vs Social Networking, 2013). The example of social networks in this era is Facebook, Twitter, LinkedIn, Instagram, Flickr, Tumblr and many more. The users of these social networks are from students, teenagers, adults, news media, authorities and also celebrities around the world. In this topic, the analysis of the social network is made.

2.2.1 Twitter

This microblogging service is using the alias "*@username*" to mention other user to join the conversation and each user can post 140 characters per time (Boyd, Golder, & Lotan, 2010). On one day, user have the limitations of tweeting 1000 post of messages or also known as messages. On the other hand, twitter also provide a famous hashtags function "*#topic*". This hashtags or trending section will list down the most mentioned topic at time. Every user who uses this hashtag will be assumed to have a homophily relationship or topical similar relationship, (Weng, Lim, & Jiang, 2010). Twitter also having the friends and followers function which is user can decide whether they want to be followed or having friends, and otherwise follow other. The other user who follows the owner of the profile is called followers. While if other user being followed by the profile owner is called following or friends. Each user can follow 2000 users. Another unique function of Twitter is the retweet function. Retweet is also known as sharing. Because of the character limit for posting tweets, Twitter provide retweet function for sharing or embed the post to own post. Favorite is also another function in Twitter, but this function almost being ignored by people. The functionality is same as the well-known 'like' function on Facebook. But differ to Facebook, by favorited a post, the post will be saved in user's profile.

© Universiti Teknikal Malaysia Melaka