TEXT MINING ANALYTICS OF SOCIAL MEDIA STREAM AND SEARCH ENGINE FOR CANCER RELATED DISEASE

BY

NURLIYANA BINTE SALEH

A dissertation submitted in fulfilment of the requirements for the degree of Master of Information Technology

Kulliyyah of Information and Communication Technology
International Islamic University Malaysia

DECEMBER 2015

ABSTRACT

With the rapid development of information and communication technology, seeking information over the internet has been time efficient and easily available regardless of time and place. The vast availability of variety of information on the internet can become a daily routine for the people in this digital age to search for information. The channels that provide information through internet have improvised and grown broader. Previously, people search information through search engine but nowadays, social media has been the upcoming trend for people to search for information too. The ubiquity of internet and availability of information has caused people to acquire information about health easily for their respective purposes. In this study, examination of information seeking behaviour between search engine and social media for cancer related disease will be conducted. Also, to analyse if there is any difference between developing and developed countries information seeking behaviour on cancer related disease. Text mining and content analysis methods are used to analyse the information seeking behaviour about cancer related disease. The findings show that the development of the country influences the information seeking behaviour of the users. The country's infrastructure, economy and level of education plays an important role in the information which is sought by the users about the cancer related disease.

ملخّص البحث

مع التطور السريع لتكنولوجيا المعلومات والاتصالات، أصبح البحث عن المعلومات من خلال شبكة الإنترنت والعثور عليها تتم بسهولة وسرعة بصرف النظر عن الوقت والمكان. ولتَوفر المعلومات المتنوعة بشكل واسع وهائل في هذا العصر الرقمي حفزت الناس وجعلت عملية البحث عن المعلومات من خلال شبكة الإنترنت عادة يومية، وأصبحت قنوات توفير المعلومات أكثر تطورا وشيوعا. في السابق، كان الناس يبحثون عن المعلومات باستخدام محركات البحث، حالياً أصبحت وسائل التواصل الاجتماعي تقليداً شائعاً لدى الناس من أجل البحث عن المعلومات. إنّ وجود الإنترنت في كل مكان، وتَوفر المعلومات مكنت الناس بسهولة على الحصول على المعلومات الصحيحة والدقيقة لأسباب خاصة بهم. تتناول هذه الدراسة المقارنة بين سلوك البحث عن معلومات المتعلقة بمرض السرطان عند استخدام محركات البحث وبين سلوك البحث عند استخدام وسائل التواصل الاجتماعي. كذلك تقوم بتحليل ما إذا كان هناك اختلاف لسلوك البحث عن المعلومات المرتبطة بمرض السرطان بين الدول النامية وبين الدول المتقدمة. تمَّ استخدام طرق تحليل النصوص"Text Mining" وتحليل المحتوى "Content Analysis" لتحليل معلومات عن سلوك البحث المتصلة بمرض السرطان. أظهرت النتائج أن تقدم الدولة له تأثير في سلوك البحث عن المعلومات من قبل المستخدمين. تؤدي البنية التحتية للدولة، واقتصادها ومستوى التعليم دوراً مهما في البحث من قبل المستخدمين عن المعلومات المتعلقة بمرض السرطان.

APPROVAL PAGE

I certify that I have supervised and read this study and that in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Master of Information Technology

.....
Mira Kartiwi
Supervisor

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Master of Information Technology

.....
Madihah Bt. S. Abd. Aziz
Examiner

This dissertation was submitted to the Department of Information Systems and is accepted as a fulfilment of the requirement for the degree of Master of Information Technology

.....
Asadullah Shah
Head, Department of Information Systems

This dissertation was submitted to the Kulliyyah of Information and Communication Technology and is accepted as a fulfilment of the requirement for the degree of Master of Information Technology

.....
Abdul Wahab Bin Abdul Rahman
Dean, Kulliyyah of Information and Communication Technology

DECLARATION

I hereby declare that this dissertation is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Nurliyana Binte Saleh

Signature …… Date ……

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHTAND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

TEXT MINING ANALYTICS OF SOCIAL MEDIA STREAM AND SEARCH ENGINE FOR CANCER-RELATED DISEASE

I declare that the copyright holder of this dissertation is jointly owned by the student and IIUM.

Affirmed by Nurliyana Binte Saleh

…….. ……

Signature Date

Dedication to:

My beloved parents, brothers, sister-in-law, sister, relatives and friends

Thank you for your prayers, endless supports, and having faith in me

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and the Most Merciful, along with Salawat and Salam to our role model, the Prophet Muhammad SAW.

Alhamdulillah, praise to Allah for his blessings and guidance of His grace, and also for giving me strength, idea, ability and patience. Finally, I could complete this dissertation.

I would like to take this opportunity to express my appreciation to my supervisor, Assistant Professor Dr. Mira Kartiwi for the guidance, advice, supervision and idea that she has provided throughout the research. I can never ask for a better supervisor who is very supportive, kind and understanding like her. May Allah reward her for her kindness.

Enormous thanks to my beloved parents, my dearest siblings, sister-in-law and relatives for their endless support. Thank you for always believing that I can make it till the end. Also, for their endless support and words of encouragement which have been meaningful throughout the research. I am blessed to have them in my life.

Finally, I would like to extend my thanks to my lovely friends who have given me motivational support and strength during the completion of this dissertation.

Alhamdulillah

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Social media has been becoming a popular channel for people to share and acquire information especially among young people where the information reaches a large number of audiences worldwide at a real-time basis. With the ubiquity of internet and social media trends, information seeking has changed (Kadli & Kumbar, 2013). This includes health information seeking.

Healthcare is one of the important sectors to a country where it influences the community and monetary state of a country. Also, it affects the life expectancy and population of a country. Medical organizations are aware that with the power of internet and social media, it helps to improve their reach to patients to improvement the treatments and services (Ventola, 2014). Individuals' medical data is crucial in the healthcare sector to ensure that the correct treatment is given to the patient for their specific illness. For this reason, each data is important and meaningful where it allows practitioners to make knowledgeable conclusion on the sickness of the patient and provide proper treatment to them.

Cancer is known as one of the chronic diseases that has become a worldwide killer and it has created attention to health associations that people whom suffer from cancer disease is higher (Anand et al., 2008). Apart from genetic, one of the other reasons that can cause a person to get cancer is due to their lifestyle and dietary habits (Anand et al., 2008). Persaud (2014) mentioned that social media has been used by many health related associations to create awareness about cancer disease due the

increasing number of users who prefer to browse their social media instead of watching the TV. The news or other information that the users received usually are what their friends or news agencies posted on social media, shared by their friends or news agencies that they followed through on the social media. In addition, the social media has played an important role to the youth with cancer as a motivational support (Crane, 2014). Thus, social media has provided the users more than just information pertaining to the cancer related disease. In addition, the social media can help cancer patient to connect to other cancer patients to provide them with motivational support to overcome their emotional difficulties (Crane, 2014).

1.2 PROBLEM STATEMENT

Social media has been becoming a popular reference for medical information seeking where the information shared through social media can be a faster medium to reach a large group of people within a short period of time (Chou, Hunt, Beckjord, Moser, & Hesse, 2009). Such platform has been used by patients to gain health-related information so as to improve their condition and care. Also, it has been used for interaction between medical organizations with the online communities to inform about new treatments and care pathways (Chirp & Keckley, 2010). With the vast variety of information provided over the internet on the medical information and the benefits of internet usage, internet can be useful and helpful as a source for the cancer disease information seeking. The internet can be the channel where the family of the cancer patients and the cancer patients themselves to understand about cancer disease and the treatments. With the information and knowledge obtained, it can assist them in decision-making, if required.

It is beneficial to identify the information that they sought on cancer related disease. Also, understand the online users' search habit and possible factors that can affect their information seeking behavior about cancer related disease. By understanding the information that they sought, their habits and the factors affecting the information seeking, it can be advantageous to the implementation of Big Data for healthcare industry. Twitter will be used as an example of the social media to illustrate on how information is shared with the limited number of input characters.

1.3 RESEARCH QUESTIONS

The research questions of this research are as followed:

1. What kind of information are searched on cancer disease through online e.g. search engine, social networks etc.?
2. How the information seeking differ between search engine and social media?
3. How the information seeking on cancer related disease is different between developed and developing countries?

1.4 RESEARCH OBJECTIVES

The objectives of this research are posited as followed:

1. To identify the types of information searched through search engine and social media on cancer related term.
2. To evaluate if there is any difference in information seeking between developed and developing countries on cancer related disease.
3. To assess the information seeking differences between search engine and social media.

1.5 RESEARCH SCOPES

The scopes and limitations of this research are as followed:

1. The information that are sought through Twitter social media and Google search engine on cancer disease.
2. Text mining will be performed on Twitter social media data while content analysis will be performed on Google search engine.
3. Then, compare result of the text mining result with content analysis result.

1.6 SIGNIFICANCE OF THE STUDY

The findings of this study would provide understandings on the information that are sought through online pertaining to the cancer related disease. Considering the variety availability of information and the channels where information can be acquired with the rapid development of technology, the Big Data analytics implementation can be at advantage for the health information seeking.

The findings are developed from the text mining for Twitter social media result and content analysis for Google search engine result. It will conclude the potential benefits that big data can provide in social media that can contribute to the cancer patients. Also, the vast available information in the internet can use the big data benefits to assist the cancer patients in decision making with the analytics feature.

1.7 ORGANIZATION OF THE DISSERTATION

Chapter One: This chapter elaborated on the introduction of the study, background of the research, problem statement, research objectives, research scope and significance of the study.

Chapter Two: This chapter describes the literature review of introduction to big data, big data influence to healthcare, social media impact to healthcare and the information seeking for cancer patients.

Chapter Three: This chapter describes the research methodology that has been implemented in this study. It followed by the research framework, which clarifies the process taken to carry out the research. In the last section, the tools used for this research is elaborated.

Chapter Four: This chapter represents the finding and result from data analysis after implementing the research methodology of this study.

Chapter Five: This chapter provides the conclusions, limitations and recommendations for further study.

1.8 CHAPTER SUMMARY

This chapter introduces the background of the research and the reason for the research to be conducted. It will be followed by the problem statement, the scope, the objectives and the questions of the research for discussion. Lastly, the significance of the study and the organization of this report have been described.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

Variety is one of the characteristics of Big Data. With the variety of information, availability health information online will be helpful to many people to seek information through various channels.

In this digital era, data has become an important part of life where data is procured and extracted as part of a person's daily life and activities in an organization. Data are stored, extracted, processed/integrated and analysed for various reasons depending on the purpose of the organizations to produce knowledge for decision making, if any. At present, there are various tools available for organisations to use the data to complete or drive any required activities.

Data is one of the components that are required in most of organization's daily activities. Regardless whether it is stored physically or digitally, data weighs equally important. The difference between physical and digital storage is the efficiency of the data retrieval for usage and knowledge sharing. However, both storage methods have advantages and disadvantages. Data is stored not just for storage purpose but to articulate or generate information at the later stage.

As mentioned by Mehok (2014), if the data is absent, any industry which are technology dependent will not be able to operate. Thus, regardless of the industry, this shows how important the data in the operation of the daily activities. Also, it has created data-driven environment when data is required in most of the activities. Data-

driven is where the activities performed have dependent on the data (Dictionary.com, 2015) .

In healthcare, data can be critical where patients' information and medical history are stored. Data can make a lot of difference in diagnosing a patient's condition.

2.2 BIG DATA

Neuralytix (2014) has defined Big Data as "a set of technologies that creates strategic organization value by leveraging contextualized complete data sets". Big Data is a technology which is quite popular in these recent years. It is defined as "large volumes of high velocity, complex and variable data that require advanced techniques to enable the capture, storage, distribution, management and analysis of the information" by Raghupathi & Raghupathi (2014). This can cause a big investment for implementation. It is possible that big data can bring revolution to the healthcare sector with the ability to carry out analytics against high-volume data in motion and across all specialities which consists diversity (Raghupathi & Raghupathi, 2014).

2.2.1 Characteristics of Big Data

In the portrayal of Big Data, Adolph has identified the characteristics of Big Data in four (4) Vs which are: Volume, Velocity, Variety and Veracity. Volume is associated with the amounts of the data that are stored, processed, retrieved and analyse while the velocity is the speed to flow the sets of data to present as knowledge of information. Variety is the variation of information format that it can be in structured and unstructured form of data. Lastly, veracity is the credibility and quality of the data.

The following illustration below will describe further on the characteristics of four(4) Vs (Adolph, 2014):

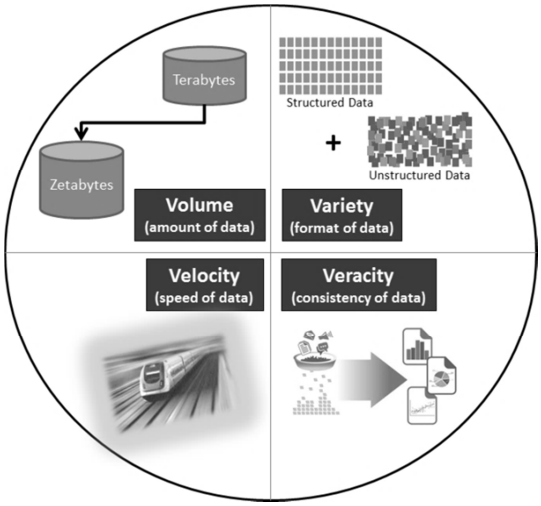

Figure 2.1 Illustration of Big Data Characteristics (Adolph, 2014)

1. *Volume*

   The word 'Big Data' itself is able to describe the quantity of the data. Data has grown with the evolution of technology. Data has been an important part of daily activities where it has to be accessible at anytime, anywhere and by anyone.

2. *Velocity*

   As the technology evolves and the data grows, the accessibility of the data may be challenged as to the speed of access and the quality of data. With the increase of data, the speed of the real-time streaming data might pose a concern. Thus, the response time for any request is crucial.

3. *Variety*

The availability of data may originate from various sources where it can be accessed or displayed in various formats (structured and unstructured). Big Data technology data should be able to cater to process or manage various formats. Also, it is important to remain the authenticity and credibility of the data when stored, processed, retrieved and analysed.

4. *Veracity*

When the available data are from various sources, in various formats and in large quantities; the credibility and the accuracy of the data can be uncertain which can cause data inconsistency, incompleteness, ambiguities and latency. As a result, cost efficiency may arise due to poor data quality. Thus, Big Data technology must be able to retain the data quality regardless of format, source and size of data.

In addition to the above four(4) Vs, as illustrated below, another two(2) Vs has been identified as the characteristics of Big Data by Kettleborough (2014):

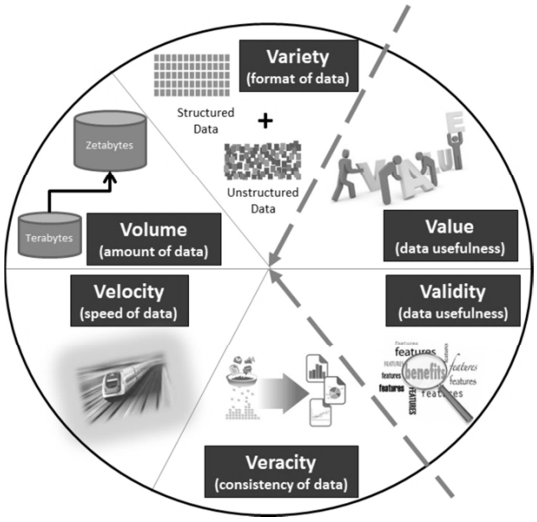

Figure 2.2 Illustration with Additional Big Data Characteristics (Kettleborough, 2014)

5. *Validity*

   Even though information is available from various sources in various formats, the data should be able to aid in decision making. In other words, the data available must be useful in the big data technology so that it is valid for capturing, storage, processing and assessing.

6. *Value*

   In contrast, even if the data is valid to be captured, stored and processed, it may not portray the data's worth. Thus, means that the data should be able to portray its benefits and usefulness where it is more than just for capturing, storage, processing and usage in the big data technology.

2.2.2 Emergence of Big Data

The development of Big Data in the digital world has created changes across all industries. Big Data innovation has created efficiency where it can benefit any organization, provider and consumer. One of the changes is how information is being transmitted, stored, processed and accessed. The fast growing of data over many years has caused an evolution in Big Data development. To add on, there is a need to provide prompt response and better quality of information to the consumer.

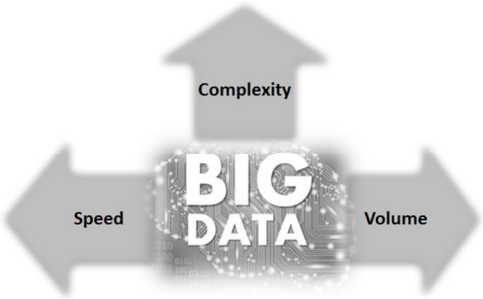

Figure 2.3 Illustration of the key reasons by Treadway & Fuchs

With reference to the illustration above, the infrastructure breaking points of Big Data are caused by three (3) key reasons which are complexity, speed and volume. As stated by Treadway & Fuchs (2011):

1. *Complexity*

   Data access has gone beyond text and numbers where it includes instantaneous access to information and shared infrastructure. Thus, data has become more reliable with various data types. However, as it is more complex, the simple method to search, store and categorize data will no longer is relevant and inefficient.