CHAPTER 1

INTRODUCTION

1.1 Introduction

Twitter is a kind of social media platform which allows to post of short messages to followers and the public. Twitter was founded in 2006 and had grown in popularity in recently years. By the year 2013, there were over 64,575,000 registered Twitter users and over 200 million tweets being generated per day (Statistic, 2014).

On Twitter, users can share information such as advices, news, moods, facts and rumors with their friends and public, the interface enables users to post short text message (up to 140 characters) that can be read by any other Twitter users, called tweets. (See Figure 1.1)

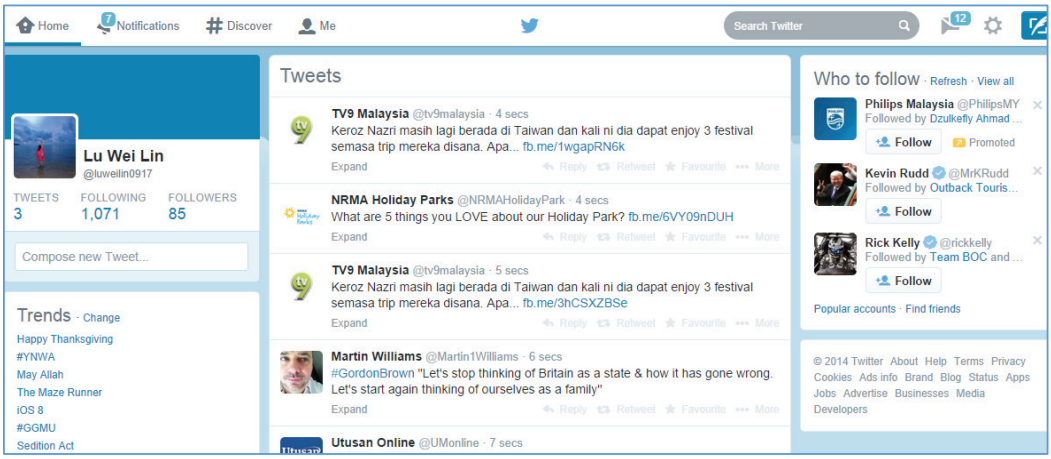

Figure 1.1 Tweets in Twitter

In general, tweet contents include short text, URL and symbols like “#”, “RT”, “@”. Hash tag is used for labeling a tweet by adding “#” tag before a word or phrase. “@” sign (followed by a user name) is used for alerting recipient about the tweets. “RT” means re-tweet information from another person. As for the texts, they are normally written in the form of short sentences.

1.2 Motivation

1.2.1 Trending topic

On Twitter, some topics are mentioned more than usual. This result in trending topics like the one in the “Trends” column of home page on Twitter (see Figure 1.1). When a hot topic is discussed on Twitter and the tweet is tagged with #, trending topic will be created. Among all the tweets within a trending topic, not all the contents are interested by a user. For example, the contents of tweet for trending topic #MH370, “*You are correct, there is NO WAY you can assume control from outside the aircraft.* There’s *NO BUAP #MH370*” in Figure 1. 2 may not be interesting for a particular user. Therefore, within a trending topic, each individual has his own preference in the tweets that he likes to see.

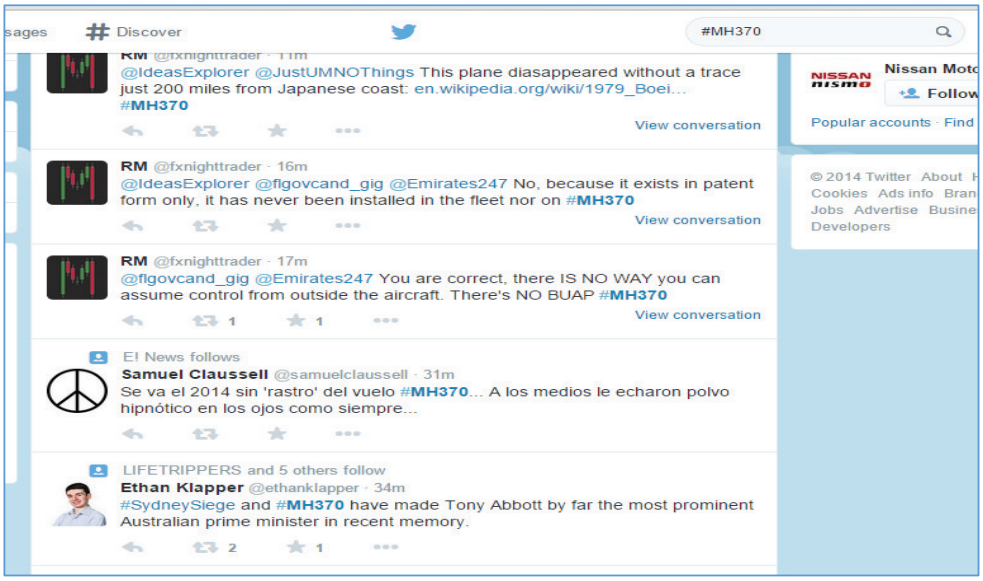

Figure 1.2 Trending tweets for MH370

For example, when many tweets related with MH370 are tagged with "#mh370", trending topic "#mh370" will be created. However, when viewing a trending topic, many tweets are generated every minute. This causes information overload to a user. Therefore, how to personalize these tweets so that the tweets can be more relevant to an individual user has sparked the motivation of this research.

1.2.2 Personalization in Social Media

Personalization means that users only get information according to their preferences. In social media, these preferences can be indicated in some ways such as basing on their profile, relevance feedback etc. Therefore, personalization is an important way to filter the information, especially, in the social media where information overload is a common issue.

For example, in Facebook, personalization allows users to see their favorite posts by clicking the "Like" button on a particular post. This feedback allows users to get specific contents more often than others, whereas if a user dislikes a post, the user can click the "I don't want to see this" button on a post. This feedback allows users to decide that what they want to see fewer posts from that particular source. (see Figure 1.3)

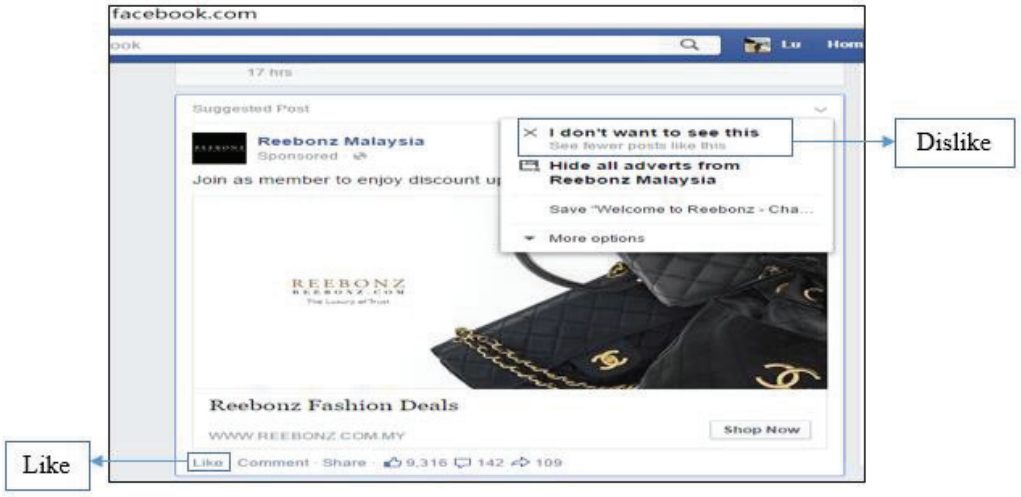

Figure 1.3 "Like" and "Dislike" on Facebook

Similarly, we can filter the information in a similar manner on Twitter, by using "Like" and "Dislike" approaches for tweet filtering. When users are reading trending tweets, they can indicate whether they like or dislike a tweet. For example, a user who prefers topic-based information rather than feeling related contents may annotate the tweets as in Figure 1.4. The first and second tweets are annotated with "Like" because they contain the topic-based information. The third tweet is annotated with "Dislike" because it contains feeling oriented contents. These annotated tweets can be used as feedback to suggest relevant tweets to the same user.

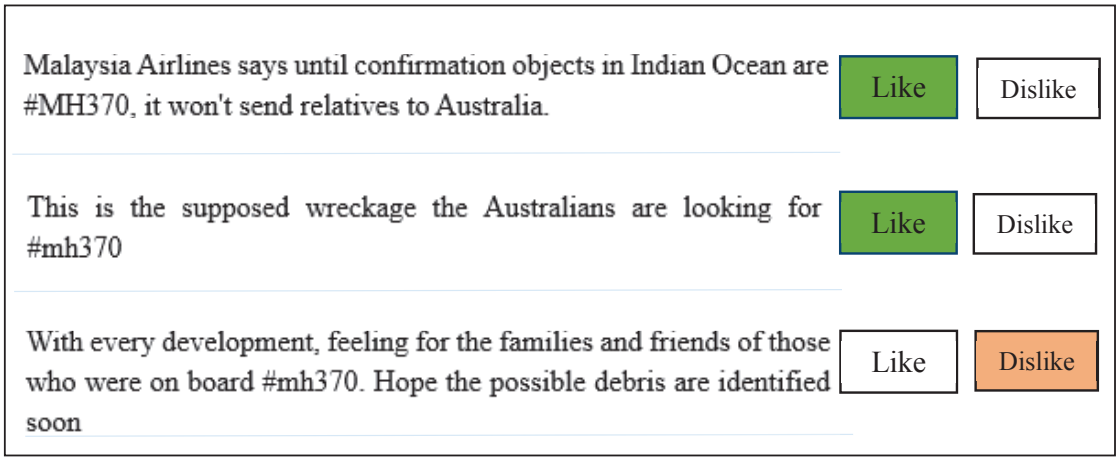

Figure 1.4 Annotating tweets with "Like" or "Dislike".

1.2.3 Feature Selection for Classification

One common approach used to suggest related tweets is by classifying the new tweets based on the training tweets. For enabling personalization, training tweets are the tweets provided by users via indication whether they like or dislike some seen tweets. In this thesis, a "Like" and "Dislike" category model will be constructed for learning these features based on the tweets indicated by users. The category model will be used for classifying new tweets based on the learnt features.

As such, feature selection is crucial to choose the representative keywords. Since tweet content is a very short compared to the normal document, our feature selection will focus on feature selection method that is suitable for very short texts.

1.3 Problem Statement

This study focuses on the personalization of tweets for trending topic. The main goal is to categorize tweets based on users' "Like" or "Dislike". However, there are some

problems during the categorization. Firstly, since the content of discussion for trending topic may change over time, such as during the incident of MH370, and one month after the event. This will affect the tweets that a user may like or dislike. Therefore, training of tweets need to adapted the change on contents based on short term rather than long term trending tweets. So we need to find out a reasonable small number of training tweets that can be used to build "Like" and "Dislike" category model. The main reason for us to target small number of training tweets is that the category model can be trained in fast manner to accommodate to topic change for trending topic. Secondly, tweets is short, thus limit the training by using approach like term frequency. As such, we need to find an approach that does not require frequencies especially in feature selection.

As the above mentioned, some problems related to personalization of tweets for trending topic that can be summarized as follows:

1. What is the numbers of tweets that can be used to build "Like" and "Dislike" category model with reasonable accuracy?

2. How can we select relevant features from the training tweets to construct the Like and Dislike category model?

1.4 Objective

According to the problem statement, we can list the main objectives as follows:

1. To determine the number of training tweets to build the category model with reasonable accuracy.

2. To suggested correct “Like” and “Dislike” tweets based on the constructed category model.

3. To verify the stability of category models across different topics by selecting suitable features.

1.5 Scope of Research

This thesis focuses on the personalization of trending tweets for five trending topics, which are #MH370, #MH17, #South Korea Ferry, #Thailand Election and #Lee Chong Wei. These five trending topics are the news-based topic. For each trending topic, around 1000 tweets are collected.

Those five trending topic tweets are annotated as “Like” or “Dislike” category by five users whose are experienced Twitter users. Each user has around six to eight years’ experience for various social media applications. The users are Zhang Ling, Zhang Qingxia, Zhang Taining, Kong Lingxi and Ji Ru. The annotated tweets are used in the framework evaluation.

In order to suggest tweets based on user's preference. Two kinds of category model will be built, which are "Like" and "Dislike" category model. We assume that a user would like to see his or her preference topic more often.

1.6 Contribution

In this thesis, the main contribution is that a framework for personalizing tweets for trending topic is proposed and evaluated. Tweets are classified according to users' preference into "Like" and "Dislike" category that is different with other tweets classification in previous works, such as emotion (happy, anger, sad, disappoint, etc), incident (pre-incident, during-incident and post-incident), and so on. A framework of tweets personalization using training approach is designed. A "Like" and "Dislike" category model is constructed from the training tweets and then new incoming tweets are classified based on the user's preferred contents captured in the category model.

In addition, there are some sub contributions arisen from this framework. First, aligned with our objective in finding out whether small number of tweets is sufficient in a quick training, we found that at least eight tweets are required in building "Like" and "Dislike" category model so that a reasonable accuracy of 81.00% can be achieved.

Overall, 65.04% average accuracy is achieved for the five news-based topics. Out of these five topics, higher accuracy is achieved for trending topic #Lee Chong Wei, which is 81.00%. It is because the tweets posted in this category has similar tweets content, for example, few tweets express the congratulation to Lee Chong Wei, proud

of Lee Chong Wei or stand with Lee Chong Wei etc. This can be seen from the relatively low features (141 features) compared to topic like #mh370 (548 features) captured from data collection with the same number of tweets. (see the Table 5.3) The result shows that the proposed framework is promising in suggesting preferred tweets for news-based topics.

Lastly, since our scope focuses on news-based topics, we extend our framework to include news-based contents to reinforce the content relevancy. We assume that a user who is viewing a trending topic would want to view topic-related contents for that topic, and thus it is important to filter out any tweets that talk about other matters. Therefore, external source in the form of news articles are used as additional training dataset for the "Like" category of our category model. By combining the features from tweets training tweets with features from the external source training texts as topic-based category model, we have successfully enriched the category model by news related keywords. The classification accuracy in average has achieved incremental of around 9.23% accuracy for each trending topic when news-based contents are used in building the category model.

1.7 Organization of Thesis

The reminder of this thesis is organized as follows. Chapter 2 describes some of the related works. Chapter 3 present details of methodology for construction of category model. Chapter 4 describes the proposed work that is topic-based feature selection.

The work is evaluated in chapter 5. The conclusion and discussion are presented in Chapter 6. Finally, some future directions are described in Chapter 6