Perhubungan Berkat Melalui Media Sosial

Effective Communication Through Social Media

Nor Zainah Siau[1], Ak Mohd Syarif PHM Yussof[2],

Universiti Teknologi Brunei

Zainah.siau@utb.edu.bn[1]

Abstract

Communication involving technology, i.e., modern communication, allows instant communication anywhere and anytime. This has changed the way we communicate and it has become a necessity in our everyday life. Web application such as email and web pages, and new media such as social media and Internet, enable information or messages to be sent and retrieved easily and quickly. Before the birth of new media, we are limited to traditional methods such as libraries, telegraph and printed material as our source of information dissemination. This paper explores the contribution of social media in information dissemination and presents the analysis of this information for the benefit of individuals and organisations. A prototype sentiment analysis system using supervised Linear Support Vector Classification was developed to extract and analyse sentiments (opinion or feeling) created from the social media. The sentiment extracted is associated with polarities of positive or negative for a specific subject and the outcome can be used to assist individual or organisation in their decision making.

1.0 Introduction

The purpose of this paper is to present a method to analyse and communicate information effectively, which could benefits individuals and organisations. Communication (verbal or non-verbal), in general, takes at least two ends (sender and receiver) to exchange information, ideas, skills etc. Effective communication, in this context, refers to the extraction of specific information from an information source/generator to create a good relationship between people and organisations.

Technology has revolutionised our communication today. Modern communication uses wireless network, fiber optics, satellites, which enables communication across long distances in real time. Because of these global communication possibilities, organisations or businesses have ever been able to reach their customers worldwide, 24/7 on a daily basis and individuals are able to do transaction online without the need to be physically present at the premise.

Social media such as Twitter, Instagram and Facebook is the new platform for people to express their thought and feelings on aspects of everyday life and this emerging trend has caused the explosive growth of data, called big data, on the Internet. The growing availability of high-speed Internet access has further added to the popularity of these social interaction applications.

Social media has always been associated with low quality information than the Internet. According to a research by Idris et. al. (2014), information on social media is less popular if it concerns the academic information. This is because the submitted information does not meet the requirements of academic and it is geared more towards social relationships. Many also believe that social media causes negative impact, especially to the younger generation, in relation to their education such as pollution of language. However, individuals may find these medium very useful, especially, to find out what other people's opinions on a product or service are before they decide to buy it. In addition, from the point of view of businesses and organisations, social media allows activities such as the delivery of information, promotion and the process of getting feedback from their customers to be done more effectively (Gentle & Anne , 2009; Sin, Khalil & Al - Agaga, 2012). These positive impacts motivate nearly all businesses to create facebook account, instagram account etc. to reach their customers. Analysing the customers' opinions (favorable or unfavorable opinion) about their product or services would provide powerful information for competitive analysis and marketing analysis. However, without automation, it may become a serious issue for businesses or organisations worldwide to find, make sense, monitor, filter and categorise these opinions.

2.0 Social Media

Social media as described by Wiki is "computer-mediated tools that allow people, companies and other organizations to create, share, or exchange information, career interests, ideas, and pictures/videos in virtual communities and networks". Social media is built on mobile and web-based technologies to create highly interactive platforms. It introduces substantial and pervasive changes to communication between businesses, organisations, communities, and individuals. Some of the most popular social media websites are Facebook, WhatsApp, Tumblr, Instagram, Twitter, and Snapchat. The number of accounts and posts created on social media platforms continue to grow every day.

This paper is focusing on the sentiment analysis from Twitter. Twitter is a popular microblogging social media service founded in 2006. Millions of short status messages called tweets are created and posted each day, Tweets contain up to a maximum of 140 characters. Unregistered users are able to view tweets from Twitter accounts that are set to public, while registered users are able to view tweets and post their

own. A registered user may also ‘follow’ or be ‘followed’ by other users. When following another user, the followed users’ posts will appear in the follower’s home page called Timeline. The Timeline will list down tweets according to time, with newer tweets appearing at the top, while older tweets will be pushed down the list.

Anyone can register as a Twitter user for free. As of September 2015, Twitter has 316 million active users per month and over 500 million tweets sent per day (Twitter, 2015). This shows that Twitter is still relevant and strong in the social media field, having a steady growth of users and they are increasing every day. Jaya et. al (2007) pointed out that there are four main intentions on Twitter; to talk about their daily activities, to make conversation with other twitter user using @ symbol followed by a username for replies, to sharing information/URL and to report news.

Tweets on Twitter can be accessed by an Application Programming Interface (API). A tweet is made up of several attributes: i) text - The body text of the tweet, ii) id - The unique id of the tweet, iii) author - The author of the tweet, iv) retweeted - A Boolean variable indicating if this is a Retweet, v) coordinates - Co-ordinates the tweet originated from (if available), vi) followers - The number of followers the user has, vii) following - the number of users that the user follows, viii) place - Geo information, (place name, bounding box), xi) created at - Timestamp the tweet was created at. Examples of tweets are

- *@Msdebramaye I heard about that contest! Congrats gal!!*
- *Disappointing day. My BFF’s wedding reception, my car broke down, it was raining and my phone was out :(*
- *I just finished a 3.2km run with a pace of 9/km with Nike+ GPS. #nikeplus #makeitcount*

As can be seen from the above tweets, there is only some valuable words that can help in determining the sentiment. This includes the use of emoticons and emoji. For the purpose of analysing the sentiments in this research, only text, id and author attributes are taken into account.

Tweets are man-made messages written in natural language that usually contain shortened words and slangs. Natural language and machines do not mix very well, as the computer would not be able to understand these messages without using natural language processing. One of the many challenges is the incredible breadth of topics it covers and another is the huge quantity of tweets as people tweet about anything and everything. With the rapid increase of the size of data on the internet, machines are needed to analyse the information as it can be overwhelming for individuals to do it manually. Go et al. (2009), introduced machine learning algorithms to accurately classify sentiment in tweets. This algorithm, while there are areas to be improved on, has performed reasonably well to classify sentiment in tweets.

Traditional data mining methods such as surveys, interviews and observations have been proven to be limited in terms of producing results. Such methods require more interaction with other people. While user interaction is great for research, the ratio of effort to result may not be very high. It is definitely suitable for small data sets or number of participants but may not work well on bigger datasets and number of participants without investing in more time and money. Sentiment analysis can replace and work more effectively than these traditional data collection methods.

3.0 Sentiment Analysis

Sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes. It involves identification, extraction and classification of opinion and emotions expressed in natural language. The task is technically challenging and practically offers a powerful technology to a number of problem domains such as business intelligence, marketing, biomedical and crime prevention. For example, businesses always want to find public or consumer opinions about their products and services. Potential customers also want to know the opinions of existing users before they use a service or purchase a product. Literature shows that there are numerous innovations for sentiment analysis have been proposed and developed such as Semantic analysis (Tetsuya & Jeonghee(2003)), meaning of adjectives (Alexandra & Patrick, 2010) etc.

The goal of sentiment analysis is to extract subjective information. Laurent (2015) presents an overview of sentiment analysis process, which is based on Twitter in Figure 1. Twitter sentiment analysis only works for tweets from Twitter, therefore, in order to get the relevant tweets out of the many, data mining from Twitter is required. Twitter sentiment analysis is to be able to automatically classify a tweet as a positive or negative tweet sentiment.

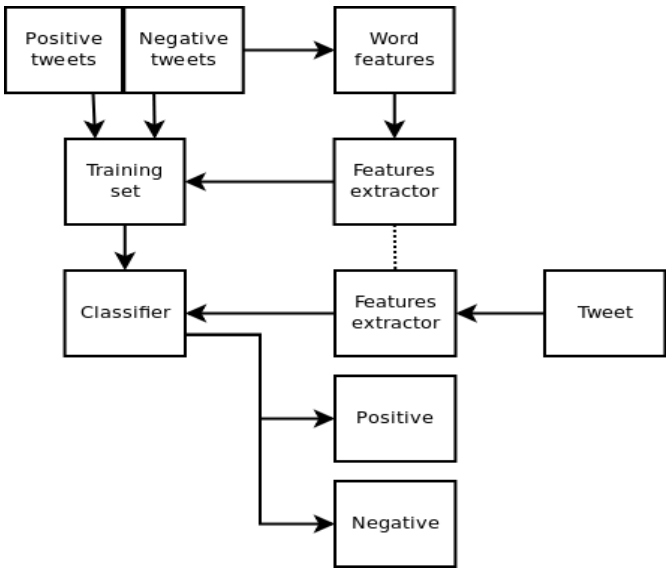

Figure 1 – Overview of Twitter sentiment analysis (Laurent, 2015)

The purpose of sentiment analysis can be described from two different perspectives; an organisation and the user.

3.1. Organisation's Perspective

An organisation could use sentiment analysis to research public opinion on their product or services for marketing purposes. The organisation can also analyse customer satisfaction of the newly introduced product through customers' feedback. These feedbacks can be used for product and service improvement. It is common to use an ad campaign to promote new products. Ideally, the success of this promotion can be determined by the number of peoples' tweet about these products. However, when a lot of people tweet about a product, it may not necessarily be positive statement. People may be talking about it due to the negative side of the product. Reading through all these tweets to get a conclusion is time consuming. Sentiment analysis will quickly reveal the actual sentiment behind those tweets without having a person to look over all related tweets.

There are several other ways that an organisation can use to get cutomers' feedback about their product. Some of the common tools are survey, interview, email, usability test and social listening. The data collected would need to be compiled and analysed to see the pattern of the customers. All these, will take time and resources and may not fit financially.

3.2 User's Perspective

Users in the context of this project are those individuals who can benefit from utilising the sentiment analysis based on their chosen tweet subjects. Sentiment Analysis can be used by the user to research products or services before making a purchase by finding out what people think of a certain product or service. For example, you want to buy a newly released mobile phone in the market such as iPhone6 Plus. You could look up at Tech websites like Phonearena, which usually provide tech expert reviews on mobile phones. But then, you may also want to know what the general public feels about it and what their experiences are when using this phone. During the release of the iPhone6 Plus, tech website gave it good reviews but a huge faulty in the device was found by users of the phone which is the sturdiness of the phone (Osborne, C., 2014).

4.0 System Design and Implementation

4.1 System Architecture

The sentiment analysis system is implemented in 3 main components; the classifier, Twitter-Python interaction and website interface. Figure 2 below shows the architecture of the system.

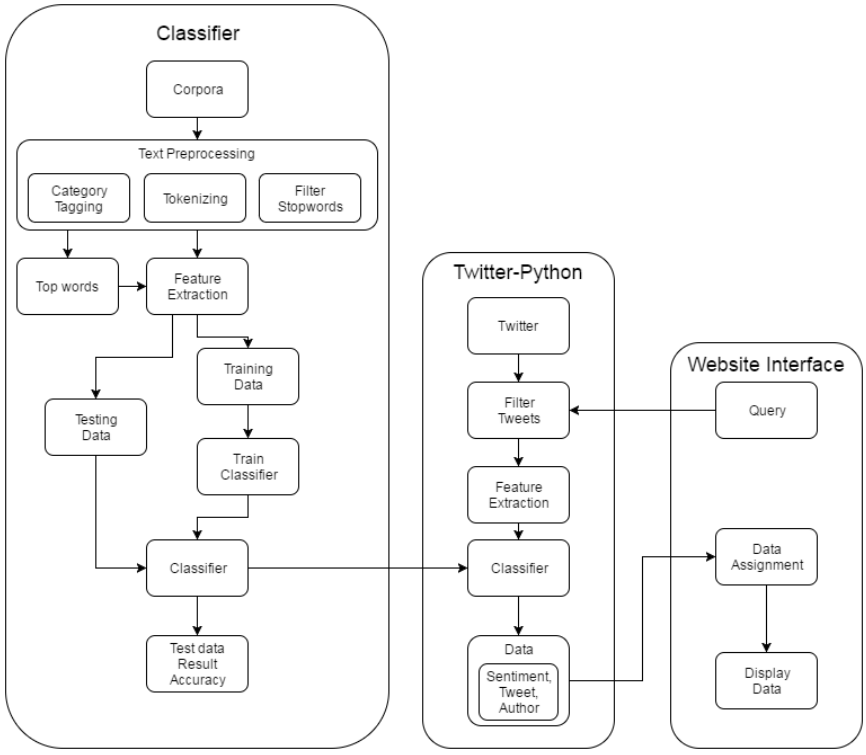

Figure 2 – Sentiment analysis system architecture

A supervised classifier was built from a trained Linear SVC algorithm. The classifier uses the Twitter corpora from Natural Language Toolkit (NLTK) containing about 20,000 tweets and the data is split into two sets; the training set and testing set. The larger the size of corpora used, the higher the accuracy of the result. This is due to having more data to sample, but process may be slower. The corpora will be filtered using text processing. A feature extractor is needed to identify what features are relevant in the tweet. A dictionary is created to store these relevant features. A good feature extractor directly determines how successful the classifier will be. A list of word features need to be extracted from these tweets by sorting every distinct word, ordered by frequency of appearance. This makes up the training set for the classifier. The classifier calculates the probability of the frequency of each feature (either positive or negative tweet) to measure the likelihood of the tweets to be a positive or a negative tweet when this word is seen as part of the input. The classifier is then trained based on the training set and tested with the testing set to determine it accuracy. This means that as a result, the training set is separated into positive and negative tweets and are labeled 'pos' for positive and 'neg' for negative tweet.

The second component of the system is Twitter-Python. This component is used for classification purposes. A search keyword will be sent from the website to Twitter-Python component. It is used to search Twitter for relevant tweets, which are the feed to the classifier so that it can sort out the sentiments of those tweets. These tweets are previously unseen data, which will go through a process called tokenisation. Tokenisation is a process of splitting strings to substrings/words. The words are then changed to lowercase to reduce the amount of data for processing. The words will be cleaned so that unnecessary words such as stop words e.g. I, we, are, the, etc., repeating more than 2 letters (e.g.happppppyyyy), and punctuation are removed. The stop words do not indicate any sentiment and for this research, stop words list are taken from NLTK, with some additional stop words added to improve the result, such as ?, http, https, 'll etc. The same classifier will be used to finds out whether the input tweet has a positive or negative sentiment.

These two components are linked together using a webpage. The webpage is a user interface, which accepts the search keyword from the user. Upon receiving the keyword, the webpage accesses the Twitter API to connect to Twitter and collect tweets for analysis. Once the type of the sentiment has been identified and all tweets are labeled as 'pos' or 'neg', the tweets along with other data are then sent back to the webpage. The webpage then renders the result received from the Twitter-Python component for the user to view.

4.2. Development Tools

The sentiment analysis system was developed and tested using a computer with 4 core Intel processor i5-4690K @3.50GHz, 16GB DDR3 RAM and 240GB flash storage. A higher specification could be beneficial as some scripts used by the system require a larger memory to achieve significantly faster execution performance.

The software used was Anaconda, Tweepy, Django and Knockoutjs. Anaconda is a Python distribution software that will install Python 2.7.11, and Phyton libraries and packages needed for the machine learning and natural language processing, and Spyder, a Python development environment program which include a text editor and built in Python console for testing and running scripts. Tweepy is a Python library for accessing the Twitter API to extract tweets from Twitter. Django is a Python Web framework that follows the MVC (model view controller) architectural pattern, used to easily integrate components of the system to the web interface. KnockoutJS is a Javascript library that follows the MVVM (Model view viewmodel) architectural pattern. It allows you to build dynamic and interactive single page web applications.

5.0 Experiment and Result

The experiment aims to evaluate whether the training data is useful for training sentiment classifier for Twitter and to evaluate the effectiveness of the features selection approach for the tweets.

Unigram approach is applied in the feature selection of the classifier to validate tweets. This approach adds a single word feature to the feature list. For future implementation, other approach may also be used to compliment unigram, such as bi-gram, which uses 2 or more words as a feature. For example, ‘don’t like’ (one feature in bi-gram) compared to 3 features (‘do’, ‘not’ and ‘like’) in unigram.

The input of the system is collected from Twitter. Twitter API only provides the past 2 weeks tweets prior the searching date and the available tweets are only from those users who set their profiles to public. For training the classifier, the first 11000 tweets (training data) are used and the remainder will be used as the testing set. With 10000 words in the feature list, the tests show that the accuracy improves (refer to table 1). This indicates that the higher amount of training data and higher feature words used will provide a higher accuracy.

Table 1 – Accuracy achieved using various combinations of training set and feature words

Training data	Feature words	Accuracy
10000	1000	56.06%
10000	5000	62.34%
11000	5000	64.11%
11000	10000	70.33%

More often than not, tweets are very informally written and often contain sarcasm, for example

These shoes are good that they make blisters on my feet.

The above tweet contains the word 'good' which indicates that it is a positive tweet. However, it is actually a negative tweet. This can skew the data and train the algorithm in the wrong way. False positive (negative tweet wrongly indicated as positive) and false negative (positive tweet wrongly indicated as negative) sentiment classification will occur and this can be seen from the accuracy of the classifier. 70% accuracy means that the remaining 30% makes up the total of both false positive and negatives classification.

6.0 Conclusion

Social media may only be seen by many as a platform to communicate feelings and opinion about a particular subject. As can be seen from the above experiment, tweets can be classified as positive or negative. This separation can help organisations to find out what is good about their products and what aspects of the product or service that needs improvement. A number of organisations, particularly those in the retail business have started to realise that posts made by their customers on their social mediums may be beneficial for them to make some improvement to their product, services and businesses.

Similarly, tweets can also be utilised by non-profit making organisations such as government ministries, to find out what the public or customers feel about their products and services. In order to get customers' feedback, using traditional ways would be filling up the feedback form with proper structure and may include the person's identification. With Twitter, people can speak freely about their opinions and thoughts and it fulfills the need of a faster mode of communication.

The experiment on Twitter sentiment analysis conducted in this research shows that the supervised classifier (trained Linear SVC) is insufficient to analyse the sentiment in the microblogging domain. The classifier is able to classify tweets with just above 70% accuracy. This may be caused by the amount of

data in the corpus provided by NLTK which has about 20,000 tweets. Some of the words encountered in the tweet may not be covered in the feature list and it gets assigned to the default classification label, which happens to be positive, in this case. Hence, the training set and the feature list are crucial for the success of the classifier. This accuracy rate could be improved by having more training data and more feature list. However, one should note that it will take longer time to train the data especially if a large number of tweets are used.

The system has a potential to be up scaled and commercialised. The main areas that could be improved are the classifier algorithm, the social platform and the corpus.

Reference

Gentle, A. (2009). Conversation and Community: The Social Web for Documentation. Fort Collins, CO: XML Press.

Idris, A.R.B., Bee, O.G., Zakaria, M.Z.M. & Pontian, J. (2014), Saluran Maklumat Akademij Pelajar Tahun Akhir di Sebuah IPT., Sains Humanika 2 (4)

Java, A., Song, X., Finin, T. and Tseng, B., (2007). Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pp. 56-65.

Sin, S. S., Khalil, M. D., & Al-Agaga, A. (2012). Factors affecting Malaysian young consumers' online purchase intention in social media websites. Procedia-Social and Behavioral Sciences, 40, pp. 326-333.

Laurent, L. (2015). Twitter sentiment analysis using Python and NLTK | Laurent Luce's Blog. Retrieved 26 September 2015, from http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/

Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis, pp.415-463, Springer Science.

Tetsuya, N. & Jeonghee, Y. (2003). Sentiment Analysis: Capturing Favorability Using Natural Language Processing In Proceedings of the 2Nd International Conference on Knowledge Capture, pp. 70-77.

Alexander, P. & Patrick, P. (2010). Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives. In Proceedings of the 5th International Workshop on Semantic Evaluation. Stroudsburg, pp. 436-439.