TPT 1201

RESEARCH METHODOLOGY IN COMPUTER SCIENCE

<ASSIGNMENT 2>

PREPARE BY

Student ID	Student Name
1121115725	TAN CHONG RAEN

LECTURE SECTION : TC01
TUTORIAL SECTION : TT02
LECTURER: *DR. POO KUAN HOONG*

OPINION MINING ON SOCIAL MEDIA DATA

TAN CHONG RAEN 1121115725
Faculty of Computer Information,
Student of Multimedia University,
Cyberjaya, Malaysia.
Email: chongraentan@gmail.com

September 18, 2015

Contents

1 Executive Summary of Research Proposal

Social media has become a very popular micro-blogging platform which play as a communication tool among Internet users. It allows users to share or post their opinions on the platform with their mobile devices or computers. Therefore, there are a massive of textual information being generated. All of these data may help company managers or businessman to make an informed decisions. So, it is worthwhile research for analyze and found the valuable and meaningful information from these data. However, there are few problems that we challenged. As post of the micro-blog usually is short and informal. Moreover, the existing algorithms are not fit in type of text and it make the accuracy of mining become lower. So, we proposed a system architecture called *Opinion Miner* to analyze the sentiments of textual information automatically and combine the social media data with the system to do the sentiment analysis and improve the accuracy. For the algorithm of *Opinion Miner*, we distribute into four steps. There are *preprocessing, extracting, short text classifying, and training multiple classifiers*. Finally, the result of experiment as our expectation as the accuracy of the *Opinion Miner* is works better than the existing algorithm, *Unigram Model*, which accuracy is 67.58% compare to 70.39%.

2 Introduction

Nowadays, social media has become a marketing tool and being actively used by governments, people, organizations and institutions. It provide a micro-blogging platform to every internet users to share their opinion, diary of their life, or they can discuss about the main issues with each other. Not only that, some organizations or users also used the social media for e-commerce purpose. They allows their customers to do product reviews, purchase products, or even private message them to get the customer services. Therefore, we can realize that the power of the social media and its data can be help the organizations or users to keep track the customer opinions by using a sentiments analysis system to analyze the sentiment content of the reviews. Opinion is the term that can help people to make a better decision. For example, governments collects feedback that provide by its people and analyze it for getting the rating of the policies. However, the post of the microblogging user is short and informal. So it make the existing algorithm could not working well. And the accuracy of analysis system is important since it may affect the users during the decision making process. For these reasons, we purposed the new system architecture which called *Opinion Miner*. *Opinion Miner* is a system that help to improve the accuracy of the sentiment analysis. Moreover, this system able to do *Short Text Classification* and the machine would learn how to extract the social media data that contain opinion and train them in the distinct categories (i.e. positive, negative). The idea for implement this system is to apply the

domain-specific training data and develop a generic classification model with social media data.

3 Justification of Research

The purpose of this empirical research is to analyze the social media data and generate the useful information to help the users or mangers to make an informed decision. Furthermore, we proposed a system called *Opinion Miner* to improve the accuracy of the existing opinion mining algorithm. *Opinion Miner* is a system architecture that used to distribute the positive or negative opinion in the social media data. So if the experiment shows the higher accuracy result then it may help the managers or users to make better decisions.

4 Research Objectives

- To implement a system that able to do the sentiment analysis from social media data.
- To improve the accuracy of sentiment analysis on social media data.

5 Literature Review

According to the one of research paper that we reviewed(Pak & Paroubek, 2010). They stated in that paper where there are few model provided to build the sentiment classifier such as *Nave Bayes, Conditional Random Fields, and Support Vector Machine(Pang, Lee, & Vaithyanathan, 2002)*. According their research, Nave Bayes yielded the best result. In additional, Nave Bayes classifier is following the Bayes theory. Furthermore, they introduce two strategies to increasing the accuracy of sentiment analysis. First strategy is depend on formula Shannon entropy and the second strategy is apply the formula of salience. In their experiments, the result indicates that the accuracy of salience is higher than entropy. Besides that, we also review another research paper about the *Sentiment analysis of twitter data*(Agarwal, Xie, Vovsha, wen Rambow, & Passonneau, 2011). In their research, they tested for two-way classification task and three-way classification task. And they apply five different models to each task. The models that they apply are *Unigram, Senti-features, Kernel, combination of Unigram and Senti-features, and combination of Kernel and Senti-features*. During their experiment, the result shows

that the accuracy of combination of Unigram and Senti-features is the highest in two-way classification task but the result shows in the three-way classification task that the accuracy of combination of Kernel and Senti-features is the highest. According to these two research paper, we found that the accuracy of the sentiment analysis may decrease significantly when apply the Unigram model in n-way classification task (when n >=3). Furthermore, the formula of salience would help to increase accuracy of the sentiment analysis.

6 Research Methodology

In this research, we need to combine data of Twitter, one of the most popular social media, with our system for do the sentiment analysis and build a manually labeled data as the training data (Ding, Liu, & Yu, 2008) in our model. After all prepared statement is done, we fetch the tweets from Twitter, and perform the first step in our system architecture, *Preprocessing*, to pre-process all the tweets. In this process, our system transformed all the words of tweet into lower case and eliminating whether the tweets are English, less than five words after greeting words, or just a URL. Moreover, this system detect all the words that start with and replaced them with USER since is represent a user in Twitter. In additional, this system convert repeated characters to a character such as sooooooooo covert to so. After that, we apply a tool *Tree Tagger* in our system to provide a *Part-Of-Speech* tag to every word in text. In the following step, we apply *Nave Bayes (NB) classifier* on the training data to classify whether the tweets is the group of opinion or non-opinion. And its formula is shows in the equation(1).

$$P(tk|c) = \frac{T_{ct} + 1}{\sum_t (T_{ct} + 1)} \tag{1}$$

Next is the Short Text Classification, we apply two different algorithm during this step. There are Mutual Information (MI)(Dumais, Platt, Heckerman, & Sahami, 1998) and X^2 Feature Selection (X^2). For the formula of MI is state in equation(2) where for each class C and each feature F, there is a score to measure how much F can contribute to making a correct decision on class C.

$$MI(C;F) = \sum\sum P(C,F) log \frac{P(C,F)}{P(C)P(F)} \tag{2}$$

In the other hand, the formula of X^2 is declare in the equation(3) where N is the total number of training sentences. N11 is the number of co-occurrences of the feature F and the class C. N10 is the number of sentences containing the feature F but that are not in class C. N01 is the number of sentences in class C that do not contain feature F. N00 is the number of sentences not in C and that do not contain feature F.

$$X^2(F,C) = \frac{N(N11 - N00 - N10N01)^2}{(N11+N01)(N11+N10)(N10+N00)(N10+N00)} \quad (3)$$

After complete the classification of short text, this system train the classifiers in the distinct categories with different training data that we labeled manually before.Furthermore, we develop many binary classifiers with all of the training data and Nave Bayes algorithm for this system.

Finally, we can start to evaluate the system and shows the result to predict the semantic orientations on Twitter. Now, we distribute the experiments into three parts. First is prepare the data sets. Next is experimental result of each step and the last is the comparison between existing model and Opinion Miner. For the first part, we have prepared two data sets that are Training Data and Testing Data. In additional, our data is collected from Twitter with Twitter API and we set the language of the data is English. Moreover, we only crawl tweet of three different categories from Twitter that are Movie, Camera, and Mobile Phone.

Next, this system extract the tweets that are contain opinion and filter out all the non-opinion tweets. After extracted the tweets, we will process to *Short Text Classification*. During this process, we apply two different feature selection to classify the text, which are Mutual Information and X^2 Feature Selection, and the top N features with highest accuracy will be apply for the features set to do the test. After that, we determine the semantic analysis of tweet by using a model of Naive Bayes.

Last but not least, we will show the result of Opinion Miner and existing model and compare them in a table.

TABLE I : RESULT OF OPINION MINER

	Positive (318)	Negative (66)	Non-opinion (126)
Positive (318)	282	28	64
Negative (66)	15	21	6
Non-opinion (126)	21	17	56

TABLE II : RESULT OF UNIGRAM MODEL

	Positive (318)	Negative (66)	Non-opinion (126)
Positive (318)	250	19	59
Negative (66)	20	40	13
Non-opinion (126)	48	7	54

TABLE III : ACCURACY OF RESULT

	Accuracy
Opinion Miner	70.39%
Unigram	67.58%

References

Agarwal, A., Xie, B., Vovsha, I., wen Rambow, & Passonneau, R. (2011). Sentiment analysis of twitter data. , 3038.

Ding, X., Liu, B., & Yu, P. (2008). A holistic lexicon-based approach to opinion mining. , 231-240. doi: 10.1145/1341531.1341561

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. , 148-155. doi: 10.1145/288627.288651

Pak, A., & Paroubek, P. (2010, April). Twitter as a corpus for sentiment analysis and opinion mining. , 1320-1326.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? : Sentiment classification using machine learning techniques. , 7986.

Project Assessment:

Title of your research projec	OPINION MINING ON SOCIAL MEDIA DATA
Member of your project	TAN CHONG RAEN
Executive Summary (5 marks)	
Introduction (3 marks)	
Justification of Research (3 marks)	
Research Objectives (3 marks)	
Literature Review (6 marks)	
Research Methodology (8 marks)	
References (2 marks)	