OMG! How do you know my feeling?
A predictive model to determine sentiments from Tweets

Yi Zhong
y6zhong@ucsd.edu

Zefeng Guo
zguo@ucsd.edu

Abstract

In this paper, we present a new method for the study of existing sentiment analysis methods in Twitter messages. In this new method, we will implement the algorithms that can determine positive and negative message. In order to achieve it, we need to analysis the existing automatic sentiment methods and the text features of social media messages, then build the new method to analysis sentiment of Twitter messages.

We will improve the existing algorithm to increase the accuracy of the prediction result. For this paper we only using hashtags and emoticons as our main feature to solve the problem. We then explain the application of the algorithm and the future research of related topic.

I. INTRODUCTION

In the past few years, there has been a huge growth in the use of microblogging platforms such as Twitter, Facebook, Tumblr. This is due to people can post message on the microblogging about their opinions on a variety of topic and express their sentiment in daily life. By the hottest topic we can clearly understand what are people talking about. Twitter is a popular social networking and microblogging website which allows users to post real-time messages called tweets. Tweets are short messages, restricted to 140 characters in length and its frequently used to express a tweeter's emotion on a particular subject. Due to the nature of this microblogging service, people will use acronyms and emoticons or other characters to indicate some special meaning. For example, tweeter's can use emotions to express their feeling. Also, user can use hash tags to mark topic. By related hashtags, you can find some "like-minded" friends. In addition to this, user of Twitter can use the "@" symbol to refer to other user on the microblogging.

There are a number of studies on how sentiments are expressed in genres such as online reviews. By determining keywords and hashtags, we can try to predict the author's sentiment. In this paper, we are focusing on build the models for classifying tweets into positive negative and neutral sentiment. Neutral sentiment refers to the text does not include emotional component.

We can see that this fits a classic predictive modeling task and we seek to exploit the synergistic relations between keywords and sentiments to implement an algorithm for automatic classification of text into positive, negative or neutral.

II. RELATED WORK

Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistic to identify and extract subjective information in source material[1]. Generally, the purpose of sentiment analysis is to determine an author's attitude for some specific topic. This attitude may express people's evaluation or judgment, emotional state when the author is writing, or the emotional effect that author wishes to have on the reader. A basic task in sentiment analysis is classifying the polarity of a given text at the documents, sentence whether the expressed opinion in a document is positive, negative or neutral. Advanced, beyond polarity sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy."

For example, someone will post on Twitter to discuss that the seafood in Seattle is good or bad. Analyzing tweets for sentiment will answer this question for you. Through catch some keywords such tasty and terrible to make the decision. From this analysis you can also learn why people think the food is good or bad, by extracting the exact word indicating why people did or didn't like the food. Because Sentiment Analysis can track a particular topic, many companies use it to track or monitor their products, services or reputation in general. For example, if someone is attacking your brand on social media, sentiment analysis will score the post as extremely negative, and you can create alerts for posts with hyper-negative sentiment scores.

III. BACKGROUND

We used two different corpora of Twitter messages in our experiments. The dataset contains over 5 million tweets which collected over a period of six months.

- **Hashtags** – Tweet containing the hashtags has a higher value than those not contain hashtags' tweet. We can said that because actually user are embedding polymerization information into the tweet through using hashtags. Except the original data set we would like to create an independent hashtags dataset to improve the algorithm. Cause we are making the predictions based on the hashtags, so first we need to removed the data that does not contained the hashtags. The existing data package contains many duplicate hashtags, we also need to remove redundant information to reduce the margin of error. In the rest of the data, we will be classified as a variety of hashtags to express different sentiments. Table 1 lists the 15 most used hashtags in our dataset.

  From table 2 we can find some obvious category hashtags such #love, #fail, #win, etc. Select the most useful hashtags that can identify different sentiments and bring them into the final package.

Hashtags	Frequency
#nyc	245,530
#nowplaying	223,482
#tfbjp	126,239
#fb	107,890
#love	103,225
#followme	98,708
#follow2befollowed	78,554
#followingfriday	63,204
#jobs	56,810
#fail	49,671
#giveaway	35,734
#news	28,334
#nyfw	20,970
#np	18,136
#win	15,209

Table 1: Most frequent hashtags in Twitter

Positive	#success #love #sleep #rest #truth #iwantfo #thanx #changinglive #health #fun #awesomeness #weightloss #romance #winning
Negative	#fail #hate #fairpay #fire #indiepub #lost #fat #cancer #tweetmyjobs #paidleave #asmsg #worse #overweight
Neutral	#deals #jobs #photography #teenchoice #media #pi #linuxkernel #polandball #mode #news #style

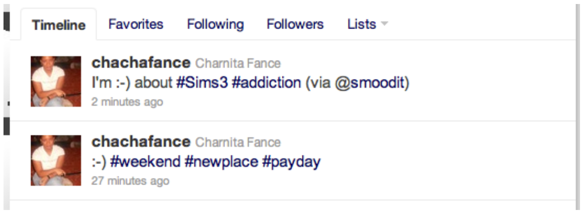
Table 2: Parts of hashtags dataset

- **Emoticon** – We prepare the emoticon dictionary by labeling over 200 emoticons listed from Wikipedia with their emotional state. For example, ":)" is determined as positive whereas ":(" is labeled as negative. We assign each emoticon a label from the following set: positive, negative, and neutral.

Emoticon	Polarity
:) :-) :] :3	positive
:(:-(	negative
:|	neutral

Table 3: Part of the dictionary of emoticons

IV. DATA



chachafance Charnita Fance
I'm :-) about #Sims3 #addiction (via @smoodit)
2 minutes ago

chachafance Charnita Fance
:-) #weekend #newplace #payday
27 minutes ago

We are using the 2009 twitter dataset which a subset of Standford dataset. The data set contains 477 million tweets which period of June to December 2009. Twitter contains a lots of short message information which created by users of this kind microblogging platform. The contents of the messages contains a strong personal thoughts and characteristics.

For the data collection, we are using the existing twitter data set and recent tweets via Twitter API. The file format includes six fields: (1) URL information; (2) The id of the tweet; (3) The date of the tweet; (4) The query, the value of this field will be NO_QUERY if there is no query existing in the dataset; (5) The user that tweeted; (6) the text of the tweet which we choose the range of Twitter message is less than 140 characters, average length of each tweet is around 80 characters or 14 words. After we get the data we do a data preprocessing in order to get the clean data and transform it to the format we need. Since the dataset contains a lots of information, we are removing some parts such URLs and user handles that we don't need it. Also we have language detection to discard tweets not in English.

Due to our goal is trying to analysis the sentiments; we will split between three sets of texts:

1. texts that containing positive emotions, such as happiness, amusement or joy
2. texts that containing negative emotions, such as sadness, anger or disappointment
3. objective texts that only state a fact or do not express any emotions

V. IMPLEMENT METHOD

A. Selection of Feature

In order to realized in this work we have to determining the polarity. We can make an assumption that what kinds of words or symbols can be determine to different sets. N-grams are among frequently used features employed when the task determining polarity is solved. In this work, any sequence of letters is a word, and any sequence of n words in n-gram of n order. To identify a set of useful n-grams, first we need to remove the stopwords. We remove articles ("a", "an", "the", "to", "of") from the dataset. Then we perform rudimentary negation detection by attaching the word not to a word that precedes or follows a negation term (such as not and no). This process will help to improve the accuracy of the classification, since the negation has a special to express the emotion and judgment. For example, the sentence "It cold today :(" includes only two kinds of n-grams which n-1 order and n-2 order: It, cold, today, It cold, cold today.

B. Selection of method

We have two methods can used to solved this problem. One is Support Vector Machines (SVMs) which purpose is to search for a linear hyper plane in a feature space dividing all entities into two classes [2]. The other method is Naïve Bayes. Since the Naïve Bayes classifier yielded the better result we decide to take it as our selected method. The main advantage of this classifier is its low computational complexity and optimality, provided there is real independence of feature. Naïve Bayes classifier is based on Bayes' theory [3].

$$P(s|M) = \frac{P(s) \cdot P(M|s)}{P(M)}$$

where s is determine as a sentiment, M is represent a tweet that contains the legally message. Since we assume that positive, negative, and neutral sentiment all are the equal sets, we can simplify the equation:

$$P(s|M) = \frac{P(M|s)}{P(M)}$$

The classifier based on POS distribution estimates probability of POS-tags presence within different set of texts and uses it to calculate posterior probability [4]. Even though POS is dependent on the n-grams, we still can make an assumption of conditional independence of n-gram features and positive information for the calculation simplicity:

$$P(s|M) \sim P(G|s) \cdot P(T|s)$$

where G is a set of n-grams representing the text information, T is a set of POS-tags of the message. We assume that n-grams are conditionally independent:

$$P(G|s) = \prod_{g \in G} P(g|s)$$

In probability theory, two events are conditionally independent given a third event precisely if the occurrence or non-occurrence of both two events is independent in their conditional probability distribution. Analogically, based on the definition we can assume that POS-tags are conditionally independent:

$$P(T|s) = \prod_{t \in G} P(t|s)$$

$$P(s|M) \sim \prod_{g \in G} P(g|s) \cdot \prod_{t \in G} P(t|s)$$

Finally, we calculate log-likelihood of each sentiment:

$$L(s|M) = \sum_{g \in G} \log\big(P(g|s)\big) + \prod_{t \in G} \log\,(P(t|s))$$

C. Selection of training set

Sentiment	Number of tweets
Positive	129
Negative	68
Neutral	54

Table 4: Calculated result of part of training set

Since the dataset contains large information, we are using 1/10 of the corpus as our training set for testing our algorithm. Then the rest can be dedicated towards training whatever algorithm we are using to classify sentiment. We compute accuracy of the classifier on the whole evaluation dataset [5], i.e.:

$$accuracy = \frac{N(correct\ classifications)}{N(all\ documents)}$$

We can measure the accuracy across the classifier's decision:

$$decision = \frac{N(retrueved\ documents)}{N(all\ doctments)}$$

The value of the decision shows what part of data was classified by the system. We are trying to using the training dataset with a very simple Naïve Bayes classification algorithm and the result were 75% accuracy, given that a guess work approach over time will achieve an accuracy of 50%. Even its not perfect that a simple approach could give out 50% better performance than guess work essentially, but given that generally 10% of sentiment classification by human can be debated, the maximum relative accuracy any algorithm analyzing over all sentiment of a text can hope to achieve is 90%.

VI. RESULT

First we test the impact of different n-gram orders on the classifier performance. We can see the comparison result from Figure 1. For this comparison we only used three different n-gram orders, when using the bigram can achieve the better accuracy. As the result we can assume that bigrams can have more accurately to determine the expression of sentiments.

Figure 2 is the result of comparison of using negation words and without using negation words. From the figure 2 we can clearly see using the negation words will have more impact to determine the accuracy of sentiments. This also verified that negation words should not be included in the analyzing range of the sentiments.

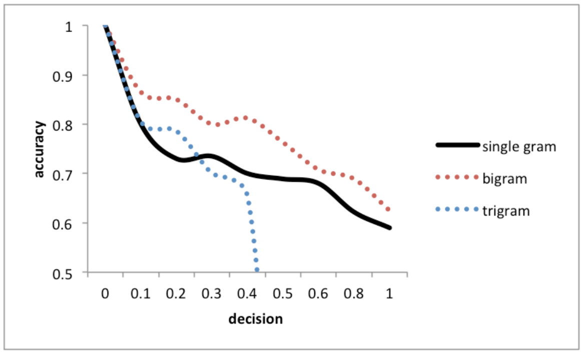


Figure 1: The comparison of using different n-gram order

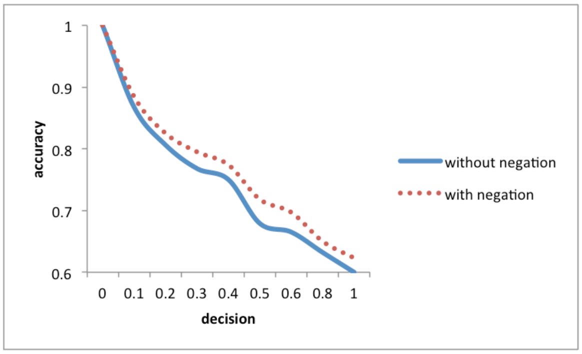


Figure 2: The comparison of using negation words and without negation words.

VII. CONCLUSION

The most attractive aspect of data mining is it allows you to discover some new knowledge from the existing information. Microblogging has become one of the most popular and indispensable tools for the social communication. It contains a wealth of information and database that can provide a good resource of data mining and analysis. In our research, we propose an idea that can automatically determine the sentiment through the text. We determine that there are three sentiments of the documents which positive, negative and neutral. The classifier is based on the multinomial Naïve Bayes classifier that uses N-gram and POS tags as feature. Since the better result on the evaluation data is calculated by using the n-gram, we still cannot assume that POS tags are less useful on microblogging data. Some POS-tags might be strong indicators of emotional text. By classifier the different tags, we calculate and analyze the distribution and the relation between the tags and sentiments. Meanwhile, we also found that there is some tags contain a strong emotional point.

We proved that it is useful for using the hashtags to collect the training data. However, there might be better way to collect the training data using different features. This can be verify in the future research.

VIII. FUTURE WORK

Analyzing the linguistic used in Twitter is a very interesting subject. We are only using the hashtags and emoticon to do the sentiment analysis in this paper. Without the hashtag, single word that contains inside the tweet can also determine the class. It might can use other algorithm to judged the sentiments changes by a sentence of grammar.

REFERENCE

[1] http://en.wikipedia.org/wiki/Sentiment_analysis
[2] Alpaydin, Ethem. *Introduction to Machine Learning*. Cambridge, MA: MIT, 2010.
[3] Anthony, Hayter. *Probability and Statistics for Engineers and Scientists.* Belmont, CA, 2007.
[4] Meral, Meric, and Banu Diri. "Sentiment Analysis on Twitter." *2014 22nd Signal Processing and Communications Applications Conference (SIU)* (2014)
[5] Psomakelis, Evangelos, Konstantinos Tserpes, Dimosthenis Anagnostopoulos, and Theodora Varvarigou. "Comparing Methods for Twitter Sentiment Analysis." *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (2014)
[6] Bahrainian, Seyed-Ali, and Andreas Dengel. "Sentiment Analysis and Summarization of Twitter Data." *2013 IEEE 16th International Conference on Computational Science and Engineering* (2013)
[7] Shannon, Claude Elwood, and Warren Weaver. *The Mathematical Theory of Communication*. Urbana: U of Illinois, 1949.
[8] Liu, Ting, Zhimao Lu, and Sheng Li. "Word Sense Disambiguation Based on Improved Bayesian Classifiers." *Journal of Electronics (China) J. of Electron.(China)* 23.3 (2006): 394-98.
[9] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2) (2008):1–135.