Deriving Marketing Intelligence over Microblogs

Yung-Ming Li
Institute of Information Management, National Chiao Tung University, Taiwan
yml@mail.nctu.edu.tw

Tsung-Ying Li
Institute of Information Management, National Chiao Tung University, Taiwan
egist1005@gmail.com

Abstract
With rapid growing popularity, microblogs have become a great source of consumer opinions. Confronting unique properties and massive volume of posts on microblogs, this paper proposes a summarization framework that provides compact numeric summarization for microblogs opinions. The proposed framework is designed to cope with four major tasks: 1) topics detection, 2) sentiment classification, 3) credibility assessment and 4) score aggregation. The experiment is held on twitter, the largest microblog platform, for proving the efficiency and correctness of the framework. We found the consideration of user credibility and opinion quality is essential for aggregating microblog opinions.

1. Introduction
Right after the blooming of blogs, microblog appeared and grew quickly as the descendant of blog from mid 2006. Today, the largest microblog platform, twitter has over 100 million users and generates 55 billion posts per day according to its report at the end of April 2010. The name of “Microblog” is coined because of its 140-characters limitation for each post. Characteristics of microblog are widely discussed as in [6,10]. Compactness of message length makes microblog posts easier to produce and consume. Moreover, microblogs are highly accessible from many mobile devices thus users are able to share and broadcast timely information and experiences conveniently. However, the format of posts is usually informal and not well structured. The following-follower model allows one to follow and receive new posts from the others. This subscription-like model stimulates the information spreading while the repost function makes message diffused even faster.

Social websites like blogs, forums and microblogs have been argued to be important means for planning marketing strategy and CRM as stated in [9]. Opposed to waiting for customers’ reach, actively collecting and analyzing customer’s opinions are suggested for gaining business competitiveness. The authors of [14,15] states that business should take social media platforms as data sources for market research and align their goals with customers’ taste. Also we can observe that much of user-generated content is informative and valuable to business managers. In the work of [6], the authors pointed out over 20% of posts mentioned a brand also expressed a sentiment or opinion concerning the company, product or service. In addition, sentiment expressed by users is fluctuant with time. Above observations leads to the conclusion that the sentiments toward brands and products needs to be watched and carefully treated as time goes.

However, the volume of microblog posts is overwhelming. As we can see in Table 1, it’s nearly impossible to read and organize every post manually. Hence a problem interests us: How to summarize and extract valuable information automatically?

Table 1. Conservative estimation of daily posts volume on several products and brands

<table>
<thead>
<tr>
<th>Entity</th>
<th>Google</th>
<th>Microsoft</th>
<th>iPhone</th>
</tr>
</thead>
<tbody>
<tr>
<td># of posts</td>
<td>&gt;50000</td>
<td>&gt;10000</td>
<td>&gt;5000</td>
</tr>
</tbody>
</table>

For the convenience of discussion we use the term “query” to represent the name of entity that user want to know about. Several sub-problems emerge as we consider the design of the automatic summarization system. First, the opinions of a query may focus on many different aspects. It’s important to learn the concerned topics and summarize opinions respectively. Second, how to convert tons of opinions into a compact type? Third, when summarizing the opinions, should we discriminately treat opinions comes from expresser with different level of credibility? We’d like to propose a system providing well-rounded, compact and representative summarization of those opinions.

On the path of constructing such a system, we develop adequate methodologies and reform them to fit the communication paradigm on microblog. On practical angle, we anticipate that the system will benefit parties in the markets. For consumers, they could know the other’s opinions on different aspects of brands and products easily. For companies, they could track the users’ perception, Therefore, labor-intensive work is minimized while correct and deeper insights are revealed.

2. Related literature
In this section we review previous work related to this research. Researches about microblogs and twitter, topic discovering related issues, sentiment analysis and user credibility assessment are discussed.
2.1. Microblog and Twitter

Many researches have revealed not only usage and behavior on microblog but also hidden marketing opportunities. The uniqueness of microblog platform is addressed in [6]: “While the shortness of the microblog keeps people from writing long thoughts, it is precisely the micro part that makes microblogs unique from other eWOM mediums, including full blogs, WebPages, and online reviews.” Also the length of standard microblog message is approximately the length of a typical newspaper headline and subhead [16], which makes it easy to both produce and consume.

Towards user behavior of microblog, the authors of [7] analyze structural properties of twitter via HITS algorithm and social network analysis measurements. The authors point out that twitter’s social network is a scale-free network in which user’s in-degree distribution follows power law distribution. Furthermore, the authors categorize user on microblog into three categories: “information seeker”, “information source” and “friends”. In [6], the authors conducted experiment on a twitter dataset and provide two insights. First, over 20% of posts mentioned on a brand or product expresses sentiment as well. Second, the sentiment expressed by users change over time. These observations implicitly claim the imperative need of an efficient opinion summarization system framework.

2.2. Feature extraction and Meronyms acquisition

Since we’d like to know the subject that customers express opinion on, several research fields related to finding out relevant concepts of a given query are reviewed. Feature extraction methods are used to gather product features out of a set of product reviews. Meronyms and Hyponyms acquisition techniques are used for generating ontology with free text.

Feature extraction is to automatically identify features of products mentioned in opinions. Previous works provide strong motivations for developing automatic feature extraction while some may argue that the product features can be obtained from manufactures. Two main reasons are that the terminology used by manufactures may be different to the terms used by customers and customers may comment on unexpected features that manufactures have never thought about [5].

To deal the task of production feature extraction, the authors of [5] generate a set of frequent features by finding out frequent terms as preliminary feature set then prune the feature set by calculating compactness and redundancy of term phrases. In [20], Red Opal system also uses frequent noun and noun phrases feature extraction. Another approach applies association rule mining to find out rules, which indicate possible occurrence and position of feature terms in sentences as in [3,12].

Within the lexical and ontology engineering communities, it has been recognized that natural language text is a rich source for extracting semantic relations, such as hyponyms and meronyms. For instance, [4,22] have studied how hyponym relations could be extracted.

We believe combining above approaches could find out relevant topics from microblog opinions efficiently.

2.3. Opinion mining and sentiment analysis

Complex marketing intelligence application is made possible only when the system obtains the ability to learn the feeling expressed from customers. Opinion mining and sentiment analysis researches aim to know “what do people think” in text format opinions all over the Web [17]. Major applications of opinion mining related techniques are product review mining [5,9,13,22], recommendation systems [23] and business intelligence [2].

Sentiment analysis is to identify the sentiment of retrieved opinions. One approach is to develop linguistic resources of sentiment orientation and structures of sentiment expression as in [5]. WordNet expansion and statistical estimation [8] such as point-wise mutual information (PMI) method are two major methods. The second approach to analyze sentiment is to use machine learning classifier, which can be built with SVM, Maximum Entropy or Naïve Bayes.

Recently, there are works on sentiment analysis of microblog opinions. In [2], the authors use a predefined lexicon word set of positive and negative words to classify twitter posts and track the fluctuation of sentiment to the result of polls. The authors argue that time-intensive and expensive polls could be supplemented or supplanted by simple analyzing text on the microblog. In [4], the authors develop an analytical methodology and visual representations that could help a journalist or public affairs person better understand the temporal dynamics of sentiment in reaction to the debate video.

In the paper, the machine learning methods are examined with automatic train set retrieval method. Promising result is shown on sentiment analysis for microblog with machine learning methodology.

2.4. Credibility assessment

In the electronic media, people used many different, and new characteristics of Web information objects, like user profile, friend number and web page layout etc., in order to make their judgments of
authority and credibility. In the work of [25], the authors argue that simply taking in-links into consideration of credibility is one-sided and unfairly reward blog longevity. The authors introduce a credibility measurement of blog take blogger’s basic information, message format and reader perception into consideration. In [26], the authors state that a more authority source makes the information more credible. The HITS algorithm calculates hubness and authority from the link structural [10]. Pagerank proposed in [11], calculate a single score of each web page.

In researches of microblogs, indicators have been discussed for measuring user influence. In [1], the authors mentioned three indicators: mention influence, follow influence and retweet influence. In [27], a modified pagerank algorithm, twitterrank, is proposed. The twitterrank can be used to measure topic-sensitive influence of twitter users.

In our work, we consider the characteristics of spam user and repost behavior to evaluate user’s credibility. In the experimental result, the credibility assessment is shown of being an important factor on summarizing opinions on microblog.

3. System framework

In this section, we look at the flow of the system framework first then explain modules involved independently. On the system architecture, we follow the framework defined in [21] and fill the system with adequate tuned methodologies for microblog context. The goal of the framework is to obtain a representative score of customer opinions on microblog towards relevant topics around a user query. To achieve this goal, following modules are presented as shown in figure 3.1. Prior to any processing, we need to gather opinions on the web. A web spider is used to collect users’ opinions and relation on microblogs. Once the data is ready, we are able to do further analysis.

Topic detection module discovers the topics discussed or commented in the opinions. In Sentiment Classification module, a Support Vector Machine is trained and deployed as a sentiment polarity classifier. When it comes to the aggregation of online opinions, most prior researches only take positivity and negativity of opinions into account, we argue that opinion quality and expresser credibility should also be taken into consideration due to inconsistent user credibility and different emotion density expressed. For this concern, opinion quality is measured in opinion quality module while credibility assessment module evaluates credibility of opinion expresser. The numeric summarization module aggregates above information and provides scores on topics related to the query.

3.1. Data collection and preprocessing

For analyzing purpose, the opinions and related information need to be collected. The opinions could be retrieve via API calls available on most microblog platforms. In the framework, opinions content are collected into content database while basic information of expressers and relations among them are stored into social graph database. After opinions are collected, a copy of opinion text is POS-tagged by Stanford Part-of-Speech tagger trained with Wall Street Journal corpus.

3.2. Relevant topics discovering module

Relevant topics discovering module is designed to assigns a tendency score of being a relevant topic to each term appeared in the opinion set of a given query. From now, we defined $Q$ as a set of queries of which the system has collected related opinions $O$ and $q \in Q$ which represent a subset of $O$ which is comprised by the opinions mentioned $q$. $T$ is defined as the nouns set appears in $O$ and $t_i \in T$ is distinct term in $T$. Topic Tendency Score (TTS) of $t_i$ on $q$, $TTS_{t_iq}$, is calculated as following.

$$TTS_{t_iq} = TF_{t_iq} \times IDF_{t_iq} \times MPP_{t_iq}$$

In above equation, $TF_{t_iq}$ is the term frequency of $t_i$ in $O_q$ and $IDF_{t_iq}$ is inverse document frequency of $t_i$.

$$TF_{t_iq} = \text{number of occurrence of } t_i \text{ in } O_q$$
The consideration of TF and IDF is based on the assumption that relevant topic terms of a specific query \( q \) should appear often in \( O_q \) and should less frequently across \( O \). The last factor, \( MPP_{t, q} \), stands for the portion that a term appears with a pattern in the predefined set of meronym patterns, \( P \). We believe the adoption of meronym pattern matching will improve the precision of topic discovering. \( MPP_{t, q} \) is calculated as following equation.

\[
MPP_{t, q} = \frac{\text{number of } t \text{ occurrence in } O_q \text{ with pattern } P}{TFF_{t, q}}
\]

For each query \( q \), we calculate TTS for each term \( t_q \) and rank terms by their TTS. With the TTS-ranked terms, we take top \( k \) as relevant topics \( TP_q \) to make summarization.

### 3.3. Semantic Score (SS) evaluation module

Semantic Score Evaluation Module identifies polarity and quality of opinion and combines them as a Semantic Score (SS) for final opinion aggregation.

### 3.3.1 Opinion quality

Though the posts of microblog are short, it is still common that a post contains more than one sentence while multiple subjects mentioned in a sentence. Therefore, we’d like to know how strong an opinion is regards to a relevant topic. Since our purpose is to integrate user’s viewpoints on certain topics, obviously subjective opinions are more important.

We hypothesize that large portion of emotional words will be used in the sentences by users when they’re expressing their own feelings relative to only describing objective information. Hence, we define Opinion Quality (\( OQ \)) of a post \( o \) as average emotional and sentimental words density in every sentence in \( o \).

To evaluate quality based on our definition, we prepare a subjective word set which includes emotional and sentimental words via word set expansion with WordNet. WordNet is an online semantic lexicon in which synonyms and antonyms of words are defined. We define a seed set of subjective words suggested in [24] in advance. Then we query WordNet for synonyms and antonyms recursively and expand the final subjective word set.

Once we have the subjective word set, for post \( o \), opinion quality \( OQ_o \) is calculated as following equation. In the aggregation module, the \( OQ \) could be used to alleviate inability of SVM classifier to filtering out neutral opinions.

\[
OQ = \frac{\text{number of unigrams found in subjective word set}}{\text{number of unigrams pertained in } S}
\]

Here, \( S \) stands for sentences pertained in post \( o \).

### 3.3.2. SVM sentiment classification module

In order to convert text opinion to numeric value, identification of polarity expressed is an important step. In this module, a SVM is trained and used for opinion polarity classification.

There are three tasks when applying SVM for classification task. First, features of data have to be defined. Second, a data set used for training has to be labeled with its true classes. Third, best combination of parameters and model setting has to be found.

Upon SVM feature selection, we test various features shown in table 2. Unigrams and bigrams are distinct one-word and two-word tokens sliced from opinion text. Subjective word set is the word set expanded with algorithm shown in 3.3.1. All of these features are counted in presence-based binary value.

<table>
<thead>
<tr>
<th>Feature</th>
<th>unigram</th>
<th>bigram</th>
<th>+bigram</th>
<th>subjective word set</th>
</tr>
</thead>
</table>

SVM is a supervised machine learning method. A set of train data is required for finding MMH. In previous researches, a collection of documents (e.g. review articles) are reviewed and labeled by human experts then used as train data. However, a micro blog post is much shorter relative to an article and number of features provided is also smaller. We conduct an elementary investigation to know if we could make the usage of emoticons as indicators of sentiment expressed in opinions. We collect data from twitter with query it with two kinds of emoticons. Returned posts with “:)” are labeled with “+1” which stands for positive polarity and posts with “:(” are labeled with “-1” which means negative. Then reviewer team formed by 2 graduate students and 1 PhD student checks the correctness of labels assigned. We found that 87% posts are labeled correctly. Hence, our train data is collected in this automatic manner with which lots of work are saved and more features we could have.

With the trained SVM, polarity of opinion \( o \), \( polarity_o \in \{+1,-1\} \) is predicted. With quality and polarity of the opinions, we calculate SS as following.

\[
SS_o = polarity_o \times OQ_o
\]

### 3.4 Credibility Score (CS) evaluation module

An opinion provided by a more credible source should be taken seriously as opposed to one expressed
by a less credible source, such as a “troller” or a “spammer” for the reason that we want to obtain a fair score. Credibility Score Evaluation module is designed to measure Credibility Score (CS), which reflects credibility of opinion expresser.

While many of factors provided in prior works are not applicable on microblogs, we only take two factors, source and content, that reasonable proxy can be found in the context of microblog. Source credibility means the information comes from a credible source (user). Content credibility means the information content is rational, reasonable and believable.

To measure the credibility of a user, we calculate the user’s follower-followee ratio. (Number of the user’s followers over number of users followed by the user) A user with relatively more followers will obtain higher source credibility since it is imaginable that most users tend to follow users providing fair and informative content. The adoption of the follower-following ratio is based on the other observation: spammers often follow many users while little of people follow them back. Use of the ratio could prevent the spam opinions from affecting aggregate scores. Nevertheless, many news sources and celebrities have very large number of followers. This leads to a large follower-followhee ratio. For instances, Obama and CNN’s accounts have more then 3.5 million followers on twitter. If the ratio is used in the scoring function directly, opinions from these sources and celebrities have very large number of followers over number of users followed by the user) A rational, reasonable and believable.

We define source credibility score of user $i$, $f_i$, in social network SN, which is in $N\times N$ adjacent square matrix. If user $i$ follows user $j$ then $N_{ij}=1$ otherwise $N_{ij}=0$. Note that SN is asymmetric.

$$f_i^{SN} = \min_{j=0,j\neq i}^{N} SN_{ij}, 1$$

Second, reposts frequency should be an adequate proxy for quality of posts. In most microblog platforms, users repost posts from the others with no modification. Since users could not add any personal opinions to the reposted posts, it is believed that highly agreement is shown between the posts and the users repost them. We define $r_i$, score as the portion of reposted posts published by user $i$ in a time period $t$.

$$r_i = \frac{\text{number of posts get reposted of }i \text{ in } t}{\text{number of posts of }i \text{ in } t}$$

Finally, in time $t$, CS for user $i$ in social network SN is the geometric mean of $f$ and $r$ as following.

$$CS_{SN} = \sqrt{f_i^{SN} \times r_i}$$

### 3.5 Numeric summarization module

Finally, a weighted additive aggregation formula is used for aggregating score for topic $t$ of query $q$ as shown below:

$$\text{Score}_{i,q}^{SN,T} = \frac{\sum_{o \in O_{q,t}} (CS_i^{SN,T} \times SS_o)}{\sum_{o \in O_{q,t}} (CS_i^{SN,T} \times |S_o|)}$$

SN is the social network of opinion expresser, $T$ is a sampled time period for calculating Credibility Score and $O_{q,t}$ is the opinions set mentioned $q$ and topic $t$.

### 4. Experiment study

In this section, we provide information about the experiments conducted to verify the efficiency and correctness of the framework. We choose twitter as the platform that experiment held. The most decisive reason is that twitter is the largest microblog and having most features of microblogs.

Experiment design and processes are organized as following. In 4.1, method of data collection and characteristics of data are described. In 4.2, performance of topic discovering module is provided. Next, our test result of different feature set of SVM for sentiment classification is presented in 4.3. After that, statistical results are shown in 4.4 for proving our correctness of aggregated score.

### 4.1 Data collection method and description

Our data is collected by querying twitter for English posts via Search API. We have 2 periods of data collection, period 1 is from 2010/03/06~2010/03/25 and period 2 is from 2010/05/13~2010/05/23. A set of target queries is defined beforehand, which contains two brands and two products. With these target queries, we search on twitter and gather posts mentioned them every 3 minutes. The target queries and number opinions collected are shown in table 3.

Here we hypothesis that the user expression will have no significant change between period 1 and period 2. The data collected in period 1 is used for preliminary studies and treated as train data and test data for preparation of SVM sentiment classifier. Data collected in period 2 is used for verification of topic discovering and topic scores aggregation. The reason that period 1 is shorter than period 2 is trying to reflect the condition that we could use a smaller amount of training data to cope with future larger opinions.
Table 3. Target queries and posts number (k) collected for experiment

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td># in period 1</td>
<td>121</td>
<td>31</td>
<td>128</td>
<td>73</td>
</tr>
<tr>
<td># in period 2</td>
<td>519</td>
<td>130</td>
<td>560</td>
<td>50</td>
</tr>
<tr>
<td>Sum</td>
<td>640</td>
<td>161</td>
<td>689</td>
<td>124</td>
</tr>
</tbody>
</table>

*A: Google, B: Microsoft, C: iPhone, D: Macbook

4.2 Evaluation of topic discovering

We use precision, recall as performance metrics of topic discovering module. Precision and recall is defined as following.

\[
\text{Precision} = \frac{\text{Number of correct topic terms that system retrieved}}{\text{Number of topic terms that system retrieved}}
\]

\[
\text{Recall} = \frac{\text{Number of relevant topics retrieved by the system}}{\text{Number of all relevant topics in dataset}}
\]

Opinions collected for each query are combined into a document. With 4 queries we prepared 4 documents, so that the system can calculate term frequency (TF) and inverse document frequency (IDF).

However, the calculation of MPP requires a list of meronym patterns. We collect patterns from [19,22] and add several patterns, which might be useful in microblog posts. These patterns are shown in table 4.

In table 4, “Y” token matches the target query and “X” token matches possible topics relevant. The “(*)” token stands for wildcards, which could be any 1 or 2 terms. Each post is matched with these meronym pattern, if any of these patterns are matched then the meronym pattern frequency of topic “X” to query “Y” is increased as well as its MPP value.

Since we want to calculate precision and recall, we need to know the relevant topics in the data that should be found. Nevertheless, there is no efficiency and simple way to find out all true relevant topics hidden except checking the opinions by human. So, we draw out smaller subsets comprise by 2000 posts for each query by random sampling from all data collected in period 2 then mark out the topics that should be found manually. With the true relevant topics for each query, we evaluate precision and recall, which are shown in figure 2. In figure 2, we compare our TSS ranking with other two baseline topic detection methods. One is frequent noun/phrases proposed in [5]. The other baseline extracts hashtags (e.g. “wave” in #googlewave, “docs” in #docs) as topics terms and ranked them by occurrence times. From the recall-precision plot, we could see that generally TTS-ranked yields better result over other two baselines. The
precision of TTS-ranked is higher at the same recall level. Beyond our expectation, extracting hashtags for topic terms works poorly. This may be due the mix usage of hashtag. Except for using hashtag to mark topics, people often use hashtag with names of celebrity and places that makes the precision of hashtag lower. Another reason is that multiple words are conjunct for the use of hashtag. For instances, “iphone snow leopard” is abbreviated to “#iphonesnowleopard”. It’s not easy to separate words from the abbreviation and this largely weaken the usefulness to find topics terms with hashtags.

### Table 4. Meronym patterns used

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y X</td>
<td>Y has (*) X</td>
</tr>
<tr>
<td>X on Y</td>
<td>Y with (*) X</td>
</tr>
<tr>
<td>X for (*) Y</td>
<td>Y come with X</td>
</tr>
<tr>
<td>Y’s(*) X</td>
<td>Y equipped with X</td>
</tr>
<tr>
<td>X of (*) Y</td>
<td>Y contain(s)(ing) (*) X</td>
</tr>
</tbody>
</table>

Next, we look at point precision when top k terms are picked as topic terms which is shown in table 5.

### Table 5. Precision (%) with top k terms picked as relevant topics

<table>
<thead>
<tr>
<th>Query</th>
<th>K</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>100</td>
<td>100</td>
<td>97.4</td>
<td>94.3</td>
<td>94.5</td>
<td>97.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td>66.4</td>
<td>69.5</td>
<td>73.7</td>
<td>67.4</td>
<td>65.0</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td></td>
<td>100</td>
<td>100</td>
<td>94.7</td>
<td>83.1</td>
<td>80.7</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td></td>
<td>63.7</td>
<td>60</td>
<td>56</td>
<td>57.6</td>
<td>58.7</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td></td>
<td>82.5</td>
<td>82.4</td>
<td>80.5</td>
<td>75.6</td>
<td>74.7</td>
</tr>
</tbody>
</table>

In table 5, we notice an interesting result. The TSS scoring function works well on certain target queries while yielding usual results on the others. A possible reason is that our scoring function weights terms on its frequency of appearing in meronym patterns. If the meronym pattern that the term usually pertained in are not included in our meronym pattern lists, then the term will get very low MPP score as well as its TSS even though its TF and IDF are high. On the contrary, if the topic term appears with meronym patterns defined in our pattern list often then we got very good result. An example is the “Google” query, we observed that many services and products mentioned are in “Y X” pattern, such as “Google Maps”, “Google Docs”.

### 4.3 Sentiment classification with SVM

In this section, we describe steps of SVM training and accuracy of trained SVM provided. As mentioned in 3.3, the train data is gathered automatically by querying data with target query and emoticons. For the purpose of train and test SVM, we draw out 11,929 posts from the data collected in period #1 with two emoticons “:)” and “:(. Posts contains”:” are labeled with positive class “+1” while posts contains “:(“ are labeled with negative class “-1”. If the posts contains “:)” and “:( at the same time, then we discarded it directly. After automatic labeling, there are 7,510 positive posts, 3,947 negative posts and 236 discarded posts. Then positive and negative posts are randomly split into 5 folds, 4 folds containing 9,165 posts are used as train data and the rest 2,292 posts are test data.

Before the use of these dataset, several preprocesses are made for features reduction. First, we remove target query and topic terms in order to avoid classifier classifies sentiment by particular query or topic. Second, numbers in posts are replaced with a special token “NUMBER_OR_VERSION”. Third, we add a prefix “NOT_” to any words after a negation word in every sentence. The negation words we defined are “not”, “never” and every words end with “n’t”. Last, besides token “NUMBER_OR_VERSION”, all other words are stemmed with Porter Stemming algorithm.

Next, we extract different feature sets and evaluate the accuracy provided by them. Unigrams and bigrams are one-word and two-word tokens extracted from preprocessed posts. Subjectivity word set is the word set expanded with the algorithm mentioned in section 3.3.1. Two types of accuracy are reported. First accuracy is 5-fold cross validation accuracy and the second is accuracy yield when we use trained SVM to predict the sentiment of test data. In addition to SVM classifier, we also provide accuracy of a Naïve Bayes classifier with identical feature sets.

### Table 6. Classifier accuracy from various features sets

<table>
<thead>
<tr>
<th>Features</th>
<th># of Features</th>
<th>Naïve Bayes Test Data</th>
<th>SVM 5-fold Cross Validation</th>
<th>SVM Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>unigram</td>
<td>11,802</td>
<td>71.7%</td>
<td>90.4%</td>
<td>88.1%</td>
</tr>
<tr>
<td>bigram</td>
<td>40,830</td>
<td>63.7%</td>
<td>77.5%</td>
<td>72.0%</td>
</tr>
<tr>
<td>unigram bigram</td>
<td>52,632</td>
<td>60.74%</td>
<td>87.4%</td>
<td>81.3%</td>
</tr>
<tr>
<td>subjective word set</td>
<td>4,206</td>
<td>34.2%</td>
<td>67.7%</td>
<td>63.6%</td>
</tr>
</tbody>
</table>

As we could see in table 6, the result is similar to the works of [18]. The simple unigram feature set provides best accuracy both in Naïve bayes and SVM classifier. And the SVM provides better accuracy over Naïve bayes in the task of sentiment classification. The reason that unigram out-perform the other feature sets in the environment of microblog, a possible
explanation is that there are more informal and newly invented terms are used to express sentiment and this fact reflects on the accuracy difference between unigram and subjective word set feature sets since the subjective word set is derived from a dictionary. Hence, in the topic score aggregation, we adopt SVM model with unigram features to classify sentiments.

4.4 Score aggregation correctness evaluation

In this section, we aim to examine whether the summarized topic score is consistent to the real score assign by users directly or not. However, there is no way to obtain real numeric evaluation of mentioned topics in the opinion expressers’ minds so we introduce statistical approaches to verify consistency between general numeric evaluations from the public and the scores summarized by the system.

Four questionnaires are issued. Each of them corresponds to a target query. In the questionnaire, top 20 topics of the target query found by topic discovering module in data collected in period #2 are listed. The respondents are requested to rate the topics in a 5 point Likert scale. 1 point stands for the user having very bad impression and 5 points stands for having very good impression on that topic. Moreover, the respondents were clearly informed to skip the topics that they have no ideas and experiences instead of answering them. This is to ensure that only experienced users and customers evaluated the topics of the target query. The questionnaires are provided in the appendix.

All questionnaires were issued to English fans pages related to the target queries on facebook. There are two reasons of issuing questionnaire on facebook. First, to spread information and obtain responses of customers evaluated the topics of the target query. This is to ensure that only experienced users and customers evaluated the topics of the target query. The questionnaires are provided in the appendix.

In this experiment, two metrics is considered. First one is mean absolute error (MAE). In statistics, the mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The MAE is defined as following. As the name suggests, the mean absolute error is an average of the absolute errors $e_i = f_i - y_i$, where $f_i$ is the prediction and $y_i$ the true value.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - f_i|$$

When calculation of MAE, the average score of a topic is taken as the True value $y_i$ and score provided by the system is the prediction $f_i$. Then we compute average MAE of all valid topics for our aggregation method and three baseline score aggregation methods. The three baseline aggregation methods weights score differently. First aggregation method weights nothing, second aggregation weights opinion quality only and third method weights credibility of opinion expresser only. They are calculated as following equation.

$\text{NoWeight}_{i}^{T} = \frac{\text{polarity}_i}{|O|}$

$\text{WeightQuality}_{i}^{T} = \frac{\text{polarity}_i \times SS_i}{|O|}$, $\text{polarity}_i \in [+1,-1]$.

$\text{WeightCredibility}_{i}^{T} = \frac{\text{polarity}_i \times CS_i^{T}}{|O|}$, $\text{polarity}_i \in [+1,-1]$.

Table 8 shows MAE of our aggregation methods and three baselines on 49 valid topics. As we can see that weighting on opinion quality and credibility makes the aggregated scores closer to the user rated scores since the average MAE over 49 topics is lower. Furthermore, the aggregation method proposed in section 3.5 results in a minimum MAE while opinion quality and credibility are considered at the same time. The MAE shows that the topic scores aggregated by our method are closest to the public viewpoint.

<table>
<thead>
<tr>
<th>Target Query Questionnaire</th>
<th>Brand</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>no. of topics</td>
<td>20 A</td>
<td>20 B</td>
</tr>
<tr>
<td>no. of responses</td>
<td>88</td>
<td>70 C</td>
</tr>
<tr>
<td>no. of valid topics</td>
<td>15</td>
<td>8 D</td>
</tr>
</tbody>
</table>

* A: Google, B: Microsoft, C: iPhone, D: Macbook

However, with too small sample size may not reflect true numeric evaluation on topics by the public, therefore in the following statistical verification, only 49 topics with more than 30 responses and opinions in our dataset are taken as valid and used.

In this experiment, two metrics is considered. First one is mean absolute error (MAE). In statistics, the mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The MAE is defined as following. As the name suggests, the mean absolute error is an average of the absolute errors $e_i = f_i - y_i$, where $f_i$ is the prediction and $y_i$ the true value.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - f_i|$$

When calculation of MAE, the average score of a topic is taken as the True value $y_i$ and score provided by the system is the prediction $f_i$. Then we compute average MAE of all valid topics for our aggregation method and three baseline score aggregation methods. The three baseline aggregation methods weights score differently. First aggregation method weights nothing, second aggregation weights opinion quality only and third method weights credibility of opinion expresser only. They are calculated as following equation.

$\text{NoWeight}_{i}^{T} = \frac{\text{polarity}_i}{|O|}$

$\text{WeightQuality}_{i}^{T} = \frac{\text{polarity}_i \times SS_i}{|O|}$, $\text{polarity}_i \in [+1,-1]$.

$\text{WeightCredibility}_{i}^{T} = \frac{\text{polarity}_i \times CS_i^{T}}{|O|}$, $\text{polarity}_i \in [+1,-1]$.

One step further, we’d like to know whether it is statistically significant that the scores are consistent with public opinions. We introduce second statistical method, pair-wise t-test. Paired t-test is a statistical technique that is used to compare two population
means in the case of two samples that are correlated. Here we hypothesize that the numeric summarization of microblog opinions should be consistent with the ratings given directly by questionnaire respondents since they should both represent the viewpoint of the public. We ran paired t-test for each aggregation method with the scores obtained from the questionnaire and the results are shown below in Table 9.

The paired sample t-test is to test the hypothesis that two samples are indifferent in mean. The null hypothesis is the mean difference between paired samples is zero. Nevertheless, in our case, we are trying to prove that the scores aggregated by the system is the same to the scores provided by users directly so the result of the t-test shouldn’t be significant and not rejecting null hypothesis if scores aggregated reflects true public opinion. As shown in Table 9, the scores aggregated by NoWeight and WeightQuality aggregation methods are significant, which means the results are different from the public opinions. WeightCredibility and the proposed weighting method are both insignificant, which means these two aggregation methods provide close aggregation of topic scores to the public opinions.

Table 9. Paired sample t-test result of user rating to different aggregation methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error Mean</th>
<th>t</th>
<th>sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Weight</td>
<td>.210</td>
<td>.636</td>
<td>.091</td>
<td>2.308</td>
<td>.025</td>
</tr>
<tr>
<td>Weight Quality</td>
<td>.202</td>
<td>.658</td>
<td>.094</td>
<td>2.151</td>
<td>.037</td>
</tr>
<tr>
<td>Weight Credibility</td>
<td>.136</td>
<td>.559</td>
<td>.080</td>
<td>1.702</td>
<td>.095</td>
</tr>
<tr>
<td>Proposed</td>
<td>.099</td>
<td>.416</td>
<td>.059</td>
<td>1.659</td>
<td>.104</td>
</tr>
</tbody>
</table>

From different performance of score aggregation, we could infer three points. First, it is crucial to take opinion expresser’s credibility into account while there are numerous noise and extreme opinions on microblog. Second, weighting on opinion quality could ease the inability of SVM to filter out neutral opinions while opinion expresser’s credibility is still needed to take into account. Third, weighting opinion quality and expresser credibility could alleviate the interference of less-relevant or neutral opinions and opinions expressed by less credible sources.

5. Conclusions and future work

While number of posts produced is overwhelming and the text format is not structural enough, tedious and labor-intensive works make the process of tracking customers’ sentiment inefficient. In this paper, we proposed a system designed to summarize text opinions into numeric scores upon relevant topics of the entity that the users is interesting in. With the system, users can learn public opinions toward many aspects of entities that users are interesting in easily.

The contribution and findings are summarized as follows. First, we find that combination of meronym patterns and term frequency information could help provide good precision on topic discovering. Though microblog posts are less structural, the precision from the proposed topic discovering module is comparable to previous literatures whose data sources are usually in more formal structures. Second, we test performance of SVM as a sentiment classifier. The result is similar to previous works that unigram features provided best accuracy regards to other feature sets. Third, via statistical procedures, we find that weighting opinion on opinion quality and expresser credibility is significant for obtaining realistic aggregated score.

In additional to the academic aspect, our work provides marketing practitioners ability to monitoring market reaction against their products. With the numeric summarization available, one could leverage the numeric opinion data for trend monitoring, campaign performance evaluation and competitor comparison over time. Also, the high-level tracking systems can be visualized and integrated into intelligence dashboard for managers.

Several extensive works can be studied. First, meronym patterns play a significant role in the topic discovering module. Nevertheless, the list of meronym patterns is collected from previous researches with heuristic patterns added. A better approach is to apply data mining methods to find frequently used meronym pattern in microblog. Second, sentiment classification of microblog posts is post-wide. However, there are many posts contain only one sentence expressed sentiment while other sentences are talking non-relevant subjects. Our system is unable to distinguish the sentences by which sentiment is expressed. Extending sentiment classification to sentence-wide may provide more delicate result. Third, our credibility measurement does not take user’s profile or basic information into account. However, detailed profile may be a factor of credibility. Besides, evaluation of authority may be a good proxy for credibility. Performance of adopting credibility like HITS or pagerank should be checked.

6. Acknowledgement

This research was supported by the National Science Council of Taiwan (Republic of China) under the grant NSC 99-2410-H-009-035-MY2.
7. References


