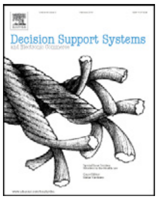
Decision Support Systems 55 (2013) 206–217

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Deriving market intelligence from microblogs

Yung-Ming Li [*], Tsung-Ying Li

Institute of Information Management, National Chiao Tung University, Hsinchu, 300, Taiwan

ARTICLE INFO

Article history:
Received 26 May 2011
Received in revised form 24 October 2012
Accepted 21 January 2013
Available online 10 February 2013

Keywords:
Social media
Microblog
Market trends
Sentiment classification
Credibility assessment
Opinion classification

ABSTRACT

Given their rapidly growing popularity, microblogs have become great sources of consumer opinions. However, in the face of unique properties and the massive volume of posts on microblogs, this paper proposes a framework that provides a compact numeric summarization of opinions on such platforms. The proposed framework is designed to cope with the following tasks: trendy topics detection, opinion classification, credibility assessment, and numeric summarization. An experiment is carried out on Twitter, the largest microblog website, to prove the effectiveness of the proposed framework. We find that the consideration of user credibility and opinion subjectivity is essential for aggregating microblog opinions. The proposed mechanism can effectively discover market intelligence (MI) for supporting decision-makers.

1. Introduction

Web 2.0 applications, such as Wikipedia, blogs, and forums, have empowered Internet users to publish their creations and opinions and spread new content via social networking. An overwhelming amount of content, which comprises life experiences, emotional expressions, criticism, and the praise of products, is all over social media platforms, while more and more people join the line of peer production. Much user-generated content is informative and valuable to business managers who are eager to learn how and in what aspects customers love or hate their products and services. Social media platforms have been argued to be important means for planning marketing strategies and managing customer relations [19,20]. As opposed to waiting for customer contact, actively collecting and analyzing customers' opinions are a suggested approach for gaining business competiveness; thus, businesses should use social media platforms as data sources for market research and align their goals with customers' tastes [26,27].

Right after the blooming of blogs, microblogs appeared and grew quickly. Microblogs descended from blogs in 2006 and have become an increasingly influential social media since. 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 term “microblog” was coined because of its 140-character limitation for each post. Microblogs have several characteristics [16,17]. First, this compactness of message length makes microblog posts easier to produce and consume. Second, 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 poorly structured. Third, the following–follower model allows one to follow and receive new posts from others without requesting permission. This subscription-like model stimulates the information spreading on microblog. Furthermore, the repost function (a.k.a. “retweet” in Twitter) makes message diffusion even faster. These characteristics make microblogs a good place to conduct e-word of mouth (eWOM) marketing. Many successful cases, such as [40,41], have shown the potential of marketing on microblogs. For example, by posting Twitter-exclusive offers to its followers, computer manufacturer Dell gained $3 million in revenue. Best Buy demonstrates another successful usage of microblogs as a real-time customer services tool, the “Twelp”, to collect customers' opinions and answer their questions. Customers could ask any questions by adding a hashtag #Twelpforce to the post. As of February 2009, @twelpforce had provided over 19,500 answers to customer inquiries.

Marketing intelligence (MI) is an important pillar of business intelligence. The MI system is designed to fulfill four needs of business managers: (1) identify opportunities and threats from the market environment; (2) help managers know more about competitors; (3) help preempt competitors' actability; and (4) aid effective marketing decision making [45]. Many MI systems are proposed to cope with traditional types of web content, such as product reviews on forums [3,12,23] or weblog usages [4]. However, few studies have effectively discovered well-rounded MI over microblogs because the microblog platform is new and has unique characteristics. Numerous posts are produced every second on microblogs, which makes them a great source of understanding customers' opinions on campaigns and the new products/services rolled out by businesses in real time.

[*] Corresponding author.
E-mail addresses: yml@mail.nctu.edu.tw (Y.-M. Li), egist1005@gmail.com (T.-Y. Li).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.01.023

To derive a MI system on microblogs, several problems are in the way. First, the volume of posts is overwhelming on microblogs. Table 1 shows our conservative estimation of the daily volume of posts that mentioned six brands and products. As shown, it is nearly impossible to read and organize every post manually because of the massive number of opinions.

Hence, an interesting problem arises: can we develop a system framework to summarize and extract valuable knowledge from opinions automatically? Several sub-problems thus emerge. First, the opinions about the topic of a user's query may focus on many different aspects. For example, when people talk about a company, they may comment on specific services, products, or even the environmental issues of the company. Therefore, it is important to know the topics concerned by the customers. Second, how should these opinions be summarized and grouped? Third, should we discriminately treat opinions that come from different expressers because of their different levels of credibility?

As in most MI systems, the data gathered from external environment are analyzed and assembled into concrete information units before providing subtle reports and helping decision making [11]. To provide advanced MI, informal text-format posts on microblogs have to be cleaned, analyzed, and polished. From microblog posts, several attributes can be extracted. We can learn what is commented and the expresser's evaluation of it. With these two basic attributes, MI is made possible. As in [4], important tasks, such as tracking customer satisfaction and competitors' performances, can be accomplished by quantifying and tracking the fluctuation of these two attributes. However, many spam accounts and extremists are reported on microblogs [22] and thereby the credibility of opinion expressers should be carefully considered.

To respond to these problems, in this research we develop a framework that achieves the following tasks. The first is topic detection, which means that the topics mentioned in the opinions associated with the queries of users should be identified and extracted. The second task is the classification of opinions. By judging the polarity of the sentiment released, the impressions held by customers can be captured. The third task, understanding the credibility of the expresser, should be assessed to provide more representative summarization. Fourth, the above three kinds of information should be aggregated adequately to reflect the true points of view of the opinions. To evaluate the effectiveness of the system framework, we conducted an experiment on Twitter.

On the path of constructing such a system, we developed adequate methodologies and reformed them to fit the communication paradigm on microblogs. Most existing information retrieval approaches deal with document-based data sources. The microblogs' characteristics mentioned above need different considerations. From a practical angle, we anticipate that the system will benefit both consumers and companies in the markets. For consumers, they could know each other's opinions on different aspects of brands, products, and services and track a product launch in real time before reading the detailed reviews. For companies, they could track users' perceptions. Therefore, labor-intensive work is minimized, while correct and deeper insights are revealed.

The rest of the paper is organized as follows. Section 2 presents related works. Section 3 demonstrates the system framework of numeric summarization of microblog opinions. Section 4 describes the experiments, along with data collection and data analysis, followed by the experimental results and discussion. Finally, Section 5 concludes and portrays future works.

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

Entity	Google	Microsoft	Sony	iPhone	iPad	Macbook
# of posts	~50,000	~10,000	~14,000	~50,000	~70,000	~7,000

* This estimation is made from the data set collected from Twitter from 2010/03/06 to 2010/03/25. We collected posts that mentioned these six entities every three minutes.

2. Related works

This section reviews related works including microblogs, feature extraction, sentiment analysis, and credibility assessment. These research fields are associated with the approaches applied in our framework.

2.1. Microblogs

As microblogs have become mainstream social platforms for information sharing, much research has revealed not only their usages and behaviors but also their hidden marketing opportunities. The uniqueness of a microblog is addressed in [16]: “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”. The length of a standard microblog message is approximately equal to the length of a typical newspaper headline and subheading [30], which makes it easy to produce and consume.

Java et al. [17] analyze the structural properties of Twitter via social network analysis to understand microblog user behavior. The authors point out that Twitter is a scale-free network. Furthermore, the authors categorize microblog users into three categories: “information seeker”, “information source”, and “friends”. As also addressed in [29], the authors find that a majority of Twitter users focus on “self”, while others focus on information sharing.

Jansen et al. [16] conduct experiments on a Twitter data set and provide two insights. First, over 20% of posts that mention a brand or product express a sentiment as well. Second, the sentiments expressed by users change over time. These observations imply the imperative need for an efficient opinion summarization framework.

2.2. Feature extraction

Feature extraction methods are used to gather product features from a set of product reviews. Several techniques have been developed to discover relevant concepts and the topics of a query [1,31]. In this section, we discuss related works and address the issue of applying them on microblogs.

Feature extraction automatically identifies the features of products mentioned in opinions. To extract product features, the authors of [16] generate a set of frequent features by finding out frequent terms and pruning the feature set by calculating term compactness and redundancy. In [35], the Red Opal system also uses frequent nouns and noun phrases for feature extraction. Another approach applies association rule mining techniques to find out syntax rules of feature term occurrence, which could discover out how frequently a feature term occurs in some kind of syntax patterns [3,10]. Besides the extraction of explicit features, it has been shown that the detection of implicit features could improve recall and precision. Within ontology engineering communities, it has been recognized that natural language texts provide a rich source for extracting semantic relations, such as hyponyms and meronyms. The acquired meronyms and hyponyms can be applied to generate ontology. How hyponym relations can be acquired automatically using linguistic patterns has also been studied [12,36]. The topics related to a query also appear via meronym patterns since the queried entity conceptualizes the topics as its attributes.

However, these methods are not sufficiently comprehensive to discover the relevant topics of microblog posts individually. Specifically, due to the short length of a post, utilizing frequency analysis to identify features/products from microblogs may generate many noise terms (e.g. some people's name or event's title), which are not satisfied topics of a query. On the other hand, the approach of synonyms/meronym pattern recognition, which identifies topics only by sematic patterns, could not fully extract the most important and relevant features/topics. In our topic detection module, we combine and adjust the above approaches to find out the relevant topics from microblog opinions.

The features/products discovered from synonyms/meronym pattern recognition are further ranked by the TF×IDF (term frequency×inverse document frequency) scoring scheme.

2.3. Sentiment analysis

Opinion mining and sentiment analysis research aims to know the opinions of users all over the Web [31]. The major applications of opinion mining are product reviews [3,11,15,16,24,25,32,33,35], recommendation systems [37], or business and government intelligence [2,8,9]. Sentiment classification aims to identify the sentiment (or polarity) of retrieved opinions. There are two categories of approaches for this task. One approach is to develop the linguistic resources for sentiment orientation and the structures of sentiment expression, and then classify the text based on these developed resources [16]. Linguistic resource development aims to construct linguistic resources that provide subjectivity, orientation, and the strength of terms, and make it possible to perform further opinion mining tasks. WordNet expansion and statistical estimation [18], such as the point-wise mutual information method, are two major methods. The second approach for analyzing sentiment is to train and deploy a sentiment classifier, which can be built with several methodologies, such as support vector machine (SVM), maximum entropy, and naïve Bayes [46].

Recently, several works on the sentiment analysis of microblog opinions have been conducted. In [8], the authors use a predefined lexicon word set of positive and negative words to classify Twitter posts and track the sentimental fluctuation to the result of polls, such as consumer confidence survey and the job approval of President Obama in the US. The authors argue that time-intensive and expensive polls could be supplemented or supplanted by simply analyzing the text on microblogs. In [9], the authors develop an analytical methodology and visual representations that could help a journalist or public affairs manager better understand the temporal dynamics of sentiment in reaction to the debate video. The authors demonstrate visuals and metrics to detect sentiment pulse, anomalies in that pulse, and indications of controversial topics that can be used to inform the design of visual analytic systems for social media events.

To classify sentiments on microblogs, machine learning should be adequate because many new sentimental words are invented and used widely on microblogs. It is difficult to determine the sentiment polarity of many exclamations and emoticons, such as “arrrg” and “>__<” by using the common sentiment linguistic sources construction approach. With large and up-to-date training data, machine learning methods are more capable to deal with those words. In our framework, an SVM classifier was used, while we apply several heuristic preprocesses and test different features to provide a more accurate classification.

2.4. Credibility assessment

Prior to the Internet era, several important criteria, such as source, receiver, message, medium, and context, were addressed to assess the credibility of the information contained in presswork and interpersonal communication [44]. As web content exploded, the credibility of web pages was questioned and discussed. In the electronic media, the above factors are modified to fit into electronic platforms and web pages on the Internet since authority is not necessarily identified in the web environment. People use many different and new characteristics of web information objects to make their judgments of authority and credibility. Recently, the metrics for evaluating blog credibility have also been studied. In the work of [44], the authors argue that simply taking in-links into consideration is one-sided and unfairly rewards blog longevity. The authors introduce a credibility measurement for blogs that takes into consideration a blogger's basic information, message format, and reader perception.

In [42], the authors state that authority leads to credibility. A more authoritative source makes information more credible. Some research adopts link analysis on web pages and provides authority indicators, such as HITS [21] and PageRank [23]. Trust in social networking sites is another promising solution to online credibility [6]. In the context of microblogs, several indicators have been discussed to measure the influence and credibility of a user. In [7], the authors introduce three indicators: mention influence, follow influence, and retweet influence. In our framework, we make use of user credibility to enhance the quality of data and thereby to derive MI.

3. The system framework

In this section, we describe the proposed framework in detail. For convenience, we use the term “query” to represent the name of the entity that end users want to know about. A query could be a keyword, such as a brand name or a product name. The goal of this framework is to identify relevant trendy topics for a user's query and obtain a representative score of microblog customer opinions towards the targeted topics. For example, when a user queries the system with “Google”, the system should find out topic terms such as “Gmail” and “Google Maps” and provide scores on these topics. Fig. 1 displays the main modules and procedures included in our proposed system.

Before processing the analysis, we need to gather opinions on the web. A web spider is used to collect user opinions and social relations on microblogs. The preprocessed data sets are collected in content and social graph databases. A trendy topic detection module discovers the trendy topics discussed or commented on in the opinions. In the sentiment classification module, a SVM is trained and deployed as a sentiment polarity classifier. While most prior research evaluating opinion sentiment has only taken the numbers of positive and negative opinions into account, we argue that opinion subjectivity and expresser credibility should also be taken into consideration because of inconsistent user credibility and the expression of different emotions. Therefore, in the present framework we develop a subjectivity analysis module to measure opinion subjectivity and a credibility assessment module to evaluate credibility. Finally, the numeric summarization module aggregates the opinions' semantic scores (opinion subjectivity and polarity) and expresser's credibility scores, and then provides a compact and trustworthy score on each relevant topic.

3.1. Trendy topics detection module

The major task of the trendy topics detection module is to assign a tendency score of being a relevant topic to each term in the opinion set of a given query. We define Q as a set of queries and O as the set of opinions the system has collected. $q \in Q$ is a query given by end users, while $O_q \subset O$ represents a set of opinions in which a query q is mentioned. For example, a tweet is taken as an opinion if the tweet contains the name of subject we are interested in or care (such as “gmail”, “google map”). If a query is “Google”, O_q is the set of tweets containing “Google”. T is defined as the set of nouns/phrases that appear in opinion set O and $t \in T$ is a distinct term in T. The “topics” are discovered from the opinions (tweets) based on the “query” subject inputs. For example, a marketing team in Apple may be interested in knowing whether “iPhone 4” (query) has any problems on “antenna” (topic).

The Topic Tendency Score (TTS) of a term t on a query q, $TTS_{q,t}$, is calculated as:

$$TTS_{q,t} = TF_{q,t} \times IDF_{q,t} \times MPP_{q,t} \tag{1}$$

where $TF_{q,t}$ is the frequency of term t in opinion set O_q and $IDF_{q,t}$ is the inverse document frequency of term t in opinion set O. Specifically,

$$TF_{q,t} = \text{number of occurrences of term } t \text{ in opinion set } O_q \tag{2}$$

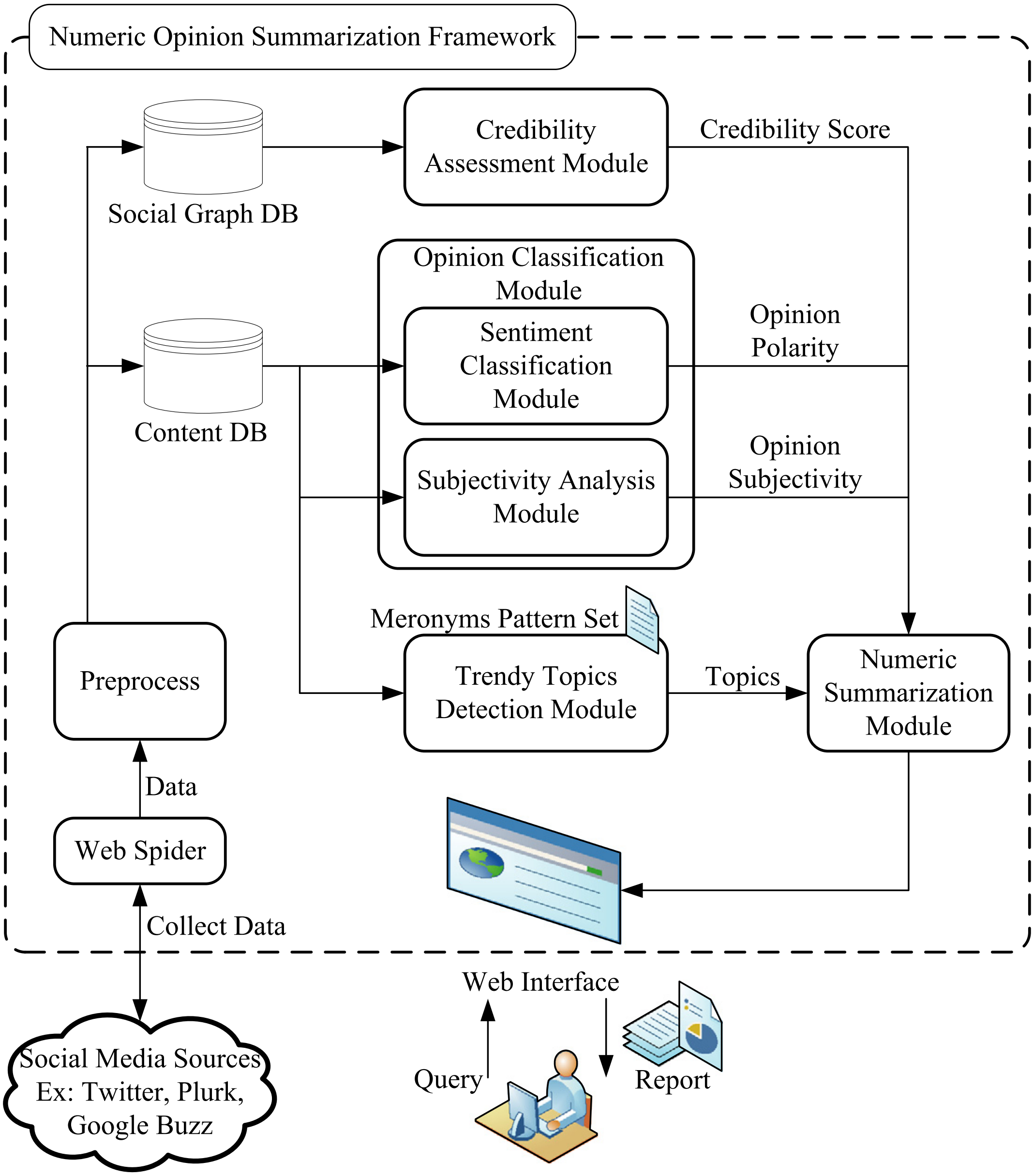

Numeric Opinion Summarization Framework
Social Graph DB
Credibility Assessment Module
Credibility Score
Opinion Classification Module
Sentiment Classification Module
Opinion Polarity
Content DB
Subjectivity Analysis Module
Opinion Subjectivity
Meronyms Pattern Set
Preprocess
Trendy Topics Detection Module
Topics
Numeric Summarization Module
Data
Web Spider
Collect Data
Web Interface
Social Media Sources Ex: Twitter, Plurk, Google Buzz
Query
Report

Fig. 1. Architecture of the numeric opinion summarization framework.

$$IDF_{q,t} = \log\left(\frac{|O|}{\left|\left\{O_q : t \in O_q\right\}\right|}\right), O_q \subset O \tag{3}$$

The consideration of TF and IDF is based on the assumption that the relevant topic terms of a specific query q (e.g. Google) should appear often in O_q and should be less frequent across O since O represents all opinions cover different queries (includes "google", "microsoft", "sony", etc.). The last factor, $MPP_{q,t}$, stands for the portion that a term appears with a pattern, which is in the predefined set of meronym patterns, P, with which people express meronym and hyponym relations. A qualified topic term should appear frequently in some meronym pattern. To improve the precision of topic detection, we utilize the meronym pattern matching method [35,36] in the module. Besides, we added some more patterns based on our observations on a bunch of tweets outside of the dataset used. For example, a post "Battery of iPhone is not good." matches meronym pattern "PART of ENTITY". "Battery" matches token PART while "iPhone" matches token ENTITY in the meronym pattern. With this evidence, we could gain confidence that "battery" is a part of "iPhone" and also a discussed topic of "iPhone". $MPP_{q,t}$ is calculated as Eq. (4).

$$MPP_{q,t} = \frac{\text{number of occurrences of } t \text{ in } O_q \text{ with patten in } P}{TF_{q,t}} \tag{4}$$

For each query q, we calculate the TTS for each term t and rank the terms by their TTSs. With the TTS-ranked terms, we select the top k terms as the relevant topics TP_q for further summarization processes.

3.2. Opinion classification module

Since the ultimate goal of our system is to provide numeric scores for opinions, we propose an approach that converts the format of an opinion from text into a numeric value. In the framework, the opinion classification module is used to identify the polarity and subjectivity of opinions and combines them as a Semantic Score (SS) for the aggregation of the final opinions.

3.2.1. Subjectivity analysis module

Although microblog posts are short, it is still likely that a post contains more than one sentence and that multiple subjects are mentioned in a sentence. To evaluate the subjectivity of an opinion, we need to know how strong an opinion is on a relevant topic. From previous literature [3,11], opinions could be classified into two categories: objective and subjective. Objective opinions are usually descriptions of the basic information about an entity and lack emotional or subjective viewpoints. Subjective opinions, by contrast, express more personal perspectives. Since our purpose is to integrate users' viewpoints on certain topics, subjective opinions are more important. Generally, a larger portion of emotional words will be used in sentences when people are expressing their own feelings relative to the description of objective information. Hence, we define the Opinion Subjectivity (*OS*) of a post *o* as the average emotional and sentimental word density in all sentences in post *o* that mentions topic *t*.

To evaluate the subjectivity level of opinions, we prepare a subjective word set, which includes emotional and sentimental words via a 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 advance [39] and then query WordNet for synonyms and antonyms recursively for word set expanding to the depth of six degrees. Once we have the subjective word set, Φ, the opinion subjectivity for a post *o* related to a topic *t*, $OS_{o,t}$, is formulated as:

$$OS_{o,t} = \left(\sum\nolimits_{s\in S_t^o} \frac{|U_s \cap \Phi|}{U_s}\right) / |S_t^o| \qquad (5)$$

where

U_s the set of unigrams pertained in sentence
S_t^o the set of sentences in opinion *o* which is mentions topic *t*.

3.2.2. Sentiment classification module

To convert a text opinion into a numeric value, the identification of polarity expressed in an opinion is an important step. Machine-learning methods, such as Naïve Bayes and SVM, perform well in sentiment classification [33]. In this module, an SVM model is trained and used for the classification of opinion polarity. There are three tasks when using an SVM for classification. First, the features of the data have to be selected. Second, a data set used for training has to be labeled with its true classes. Third, the best combination of parameters and model setting has to be found. Upon SVM feature selection, we tested various features shown in Table 2. Unigrams and bigrams are distinct one-word and two-word tokens sliced from the opinion text. As a micro-blogging message is restricted to be short, the chance of repetitive occurrence of a feature in the same message is small. Therefore, in this research, all of these features are counted in a presence-based binary value {0,1}. "1" stands for the appearance of the feature in a post, while "0" stands for its absence.

SVM is a supervised machine learning method; thus, a set of training data is required for finding good MMH (Maximum Margin Hyperplane). In previous research, a collection of documents (e.g. review articles) has been reviewed and labeled by human experts and then used as the training data. However, a microblog post is much shorter than is an article and the number of features provided is also smaller. Here, we first investigate whether we could use emoticons as indicators of the sentiment expressed in opinions. We collected data from Twitter that was queried with two kinds of emoticons: returned posts with ":)" were labeled with "+1", which stands for positive polarity, and posts with ":(" were labeled with "–1", which means negative polarity. We found that 87% posts were labeled correctly. Hence, our training data were collected in this automatic manner in order to include more features. Finally, we adopted a grid search [13] to find out the best combination of parameters c and γ for the SVM with a Radial Basis Function kernel. With the trained SVM, the polarity of opinion *o*, $polarity_o \in \{+1,-1\}$, which stands for positive and negative sentiment respectively, can be predicted.

Table 2
Feature set used for testing the performance of the SVM classification.

Feature	Unigram	Bigram	Unigram + bigram	Subjective word set
Frequency or presence?	Presence	Presence	Presence	Presence

Finally, with derived subjectivity and the polarity of opinions on a topic *t*, we can calculate the semantic score SS as:

$$SS_{o,t} = Polarity_o \times OS_{o,t}, \text{ where } SS_{o,t} \in [-1,1] \qquad (6)$$

Notice that opinion subjectivity *OS* could be used to alleviate the inability of the SVM classifier to filtering out neutral opinions.

3.3. Credibility assessment module

In addition to the consideration of a post's semantic information, information on the opinion expresser is also crucial for finding out the liberal aggregated score on relevant topics. An opinion provided by a more credible source should be taken more 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. Therefore, the credibility assessment module was designed to measure Credibility Score (*CS*), which reflects the credibility of an opinion expresser.

Previous research on the credibility of information on the web has provided a set of factors that should be taken into consideration [28,34]. For example, source, content, format, presentation, currency, accuracy, speed of page loading, and even URL are crucial indicators. However, many of these factors are not applicable on microblogs. As a result, we consider two important factors, source and content, which act as reasonable proxies in the context of microblogs. Source credibility means the information comes from a credible source. Content credibility suggests that the information content is rational, reasonable, and believable.

To measure the credibility of a user, we calculate the user's follower–followee ratio (the number of the user's followers divided by the number of users followed by the user). A user with relatively more followers will obtain a higher source credibility score since most users tend to follow other users who provide fair and informative content. The adoption of the follower–followee ratio was also based on another observation: spammers usually follow a lot of users, while few followed users follow spammers back. The use of the ratio could make spam accounts less important in the score aggregation stage. Assume there are *N* users in the social network *SN*. *SN* can be represented as an $N \times N$ adjacent square matrix. If user *i* follows user *j* then $SN_{i,j} = 1$, otherwise $SN_{i,j} = 0$. Note that *SN* is asymmetric. The network adjacency matrix is formed by the network constructed from the opinion expressers in the opinion set *O*. The source credibility score of user *i*, f_i^{SN}, is defined as:

$$f_i^{SN} = \min\left(\frac{\sum_{j\neq i}^{N} SN_{j,i}}{\sum_{j\neq i}^{N} SN_{i,j}}, 1\right) \qquad (7)$$

Notice that f_i^{SN} is used to evaluate the credibility of an expresser, rather than the popularity of an expresser. While some users likely maintain a relatively small network (tens to a hundred followers/ followees), these users' opinion should also be taken into scoring if some another people have accepted them as friends and are willing to listen opinions from them. However, an expresser with credibility should have at least some level of popularity, therefore, we can set a threshold of friend links and remove those users with very little followers and followees. In this research, the threshold number of the follwees and the threshold ratio of followers/followees are set to be 10. Notice that the threshold can be humanly adjusted according to the structure of constructed social network and available amount of reviews. A higher threshold will reduce the number of nodes and corresponding amount of reviews while the credibility level of the qualified users increases.

In addition, repost frequency should be an adequate proxy for measuring the content credibility of posts from users. On most microblog platforms, users can repost posts from others with no modifications or comments added. Since users cannot add personal opinions to the reposted posts, it is believed that there is a high agreement shown between the posts and the users that repost them. Therefore, the repost rate of a user's posts can be used as a measure of content credibility. We define the content credibility score of user *i* in a time period *TP* as:

$$r_i^{TP} = \frac{\text{number of posts reposted of user } i \text{ in time period } TP}{\text{number of posts of user } i \text{ in time period } TP} \quad (8)$$

Finally, the credibility score of user *i* is the geometric mean of source credibility score f_i^{SN} and content credibility r_i^{TP} as shown in Eq. (9).

$$CS_i = \sqrt{f_i^{SN} \times r_i^{TP}} \quad (9)$$

Notice that some interested parties (e.g. a company promoting its products or attacking its opponent's products) in the microblogging sphere may attempt to affect the analysis. To prevent the improper abuse of credibility, users with exceptional high credibility could be identified and removed. We can determine the threshold ratio of follower/followee based on the observations on spam or official user accounts. In this research, we set the ratio to be 100. Notice that when an expresser is removed, his/her opinions will also be removed and will not be considered in all relevant opinion analysis modules.

3.4. Numeric summarization module

These examinations of online text sources aim to obtain subtle numeric information for representing market trend intelligence. With the numeric evaluation of the semantic score of opinion and the credibility score of an opinion expresser analyzed by the modules described above, we can then aggregate them against relevant topics discovered on users' queries such that the texts on microblogs can be quantified into a traceable unit. The final score for a topic *t* with respect to a query *q* is formulated as:

$$Score_{q,t} = \frac{\sum_{o \in O_{q,t}} \left(SS_{o,t} \times CS_i \right)}{\sum_{o \in O_{q,t}} \left(\left| SS_{o,t} \right| \times CS_i \right)} \quad (10)$$

where $O_{q,t}$ is the set of opinions mentioning topic *t* for a given query *q* and user *i* is the expresser of an opinion *o*.

As we can see in the formulation, an opinion gains a higher score when its expresser is more credible and this opinion contains more subjective terms. The consideration of credibility could weaken the interference of spam accounts and untrustworthy opinion sources. Opinions with few subjective viewpoints would not cause large changes in aggregated scores. This would filter out objective information and focus the aggregation process on the opinions of customers expressing their tastes and perceptions.

4. Experiments

In this section, we detail the experiments conducted to verify the effectiveness of the proposed framework. We chose Twitter as the platform on which the experiments were performed because it is the largest microblog (more than 105 million registered users and 180 million unique visitors). In addition, because of the support of Twitter and popularity of smartphones, 37% of users exploit the Twitter service via their mobile devices. Besides the basic function of posting messages, Twitter provides two valuable functions: repost (a.k.a. “retweet” on Twitter) and search. The repost function makes it easy to repost other users' messages. This feature makes information spread quicker on Twitter [5]. The search function works as a filter for marketers and consumers. Twitter's search engine receives around 600 million search queries per day. The appropriate use of the search function could greatly reduce the time searching users' opinions on Twitter.

4.1. Data collection, preprocessing, and data description

The data sets for the experiments were collected from Twitter by querying its search API for English posts. We have two periods of data collection: period #1 was from 2010/03/05 to 2010/03/25 (20 days) and period #2 was from 2010/05/13 to 2010/05/23 (10 days). A set of queries, which contains three brands and three products, was defined beforehand. With these target queries, we searched on Twitter and gathered posts that mentioned them every 3 min. The data collected in period #1 were used for preliminary studies and treated as the training data and the test data for the preparation of the SVM sentiment classifier. Data collected in period #2 were used for the verification of topic detection and the aggregation of topic scores. The target queries and number of opinions collected are shown in Table 3. After the related opinions had been collected, two main preprocessing procedures were carried out: (1) a copy of the opinion text was POS-tagged using a Stanford Part-of-Speech (POS) tagger, which was trained with Wall Street Journal corpus [38] for further semantic analysis; and (2) the social networks of opinion expressers according to their follower and followee relationships were constructed for further credibility analysis.

4.2. Topic detection effectiveness evaluation

To evaluate the performance of the proposed topic detection module, we adopted precision and recall rates defined as follows.

$$\text{Precision rate} = \frac{\text{Number of relevant topic terms the system retreives}}{\text{Number of topic terms the system retreives}} \quad (11)$$

$$\text{Recall rate} = \frac{\text{Number of relevant topic terms the system retreived}}{\text{Number of all relevant topic terms in the dataset}} \quad (12)$$

Table 3
Target queries and number of posts collected for the experiments.

	Brand			Product			
Target query	Google	Microsoft	Sony	iPhone	iPad	Macbook	
# in period 1	121,834	31,025	119,848	128,574	84,293	73,337	558,911
# in period 2	519,083	130,267	21,681	560,514	517,258	50,763	1,799,566
Sum	640,917	161,292	141,529	689,088	601,551	124,100	2,358,477

Table 4
Meronym pattern for topic detection.

Y X	Y has (*) X
X on Y	Y with (*) X
X for (*) Y	Y come with X
Y's (*) X	Y equipped with X
X of (*) Y	Y contain(s)(ing) (*) X

All opinions collected for each query were combined into one document to calculate the term frequency (TF) and inverse document frequency (IDF) measures. However, the calculation of MPP requires a list of meronym patterns. Here, we collected the patterns from [35,36] and added several patterns that might be useful in microblog posts. These patterns are shown in Table 4, in which the "Y" token matches the target query and the "X" token matches the possibly relevant topics. The "(*)" token stands for wildcards, which could be any two terms at most. If any of these patterns was matched, then the meronym pattern frequency of topic "X" to query "Y" was increased as well as its MPP value.

To calculate the precision and recall rates, we need to know the relevant topics in the data that should be found. Nevertheless, there is no efficient and simple way to find out all the relevant topics hidden except by checking the opinions one by one. Thus, we drew out smaller subsets comprising 2000 posts for each query by random sampling from all the data collected in period 2 and then marked out the topics that should be judged manually. There are three coders responding for the coding work. Relevance of a discovered topic for tweets is judged by the three coders and a qualified topic must be with the unanimous consensus of them. In our experiment, the average consensus rate is 0.978 and the percentage of qualified topics to the total number of topics annotated by each coder is (0.987, 0.995, 0.982).

In Fig. 2, we compare our TSS ranking with two other baseline topic detection methods. One is the frequent noun/phrases approach proposed in [14] without the feature pruning. The other method extracts hashtags (e.g. "wave" in #Googlewave, "docs" in #docs) as topic terms and ranks them by occurrence frequency.

From the precision-recall plots, we can see that generally the TTS-ranked approach outperforms the two other baseline approaches. The precision rate of the TTS-ranked approaches is higher at the same recall level. TTS provides a good ranking for topic terms since we can see that most curves of the precision-recall rate plot go downward when the recall rate increases. Beyond our expectations, extracting hashtags for topic terms works poorly, while marking the topics of posts is recommended. This phenomenon may be owing to the mixed usage of hashtags because users often also add hashtags to address the names of celebrities and places. This lowers the precision rate of the hashtag extracting method. Another reason is that many hashtags are a combination of many terms. For instance, "iphone snow leopard" is abbreviated to "#iphonesnowleopard". "google city tour" is abbreviated

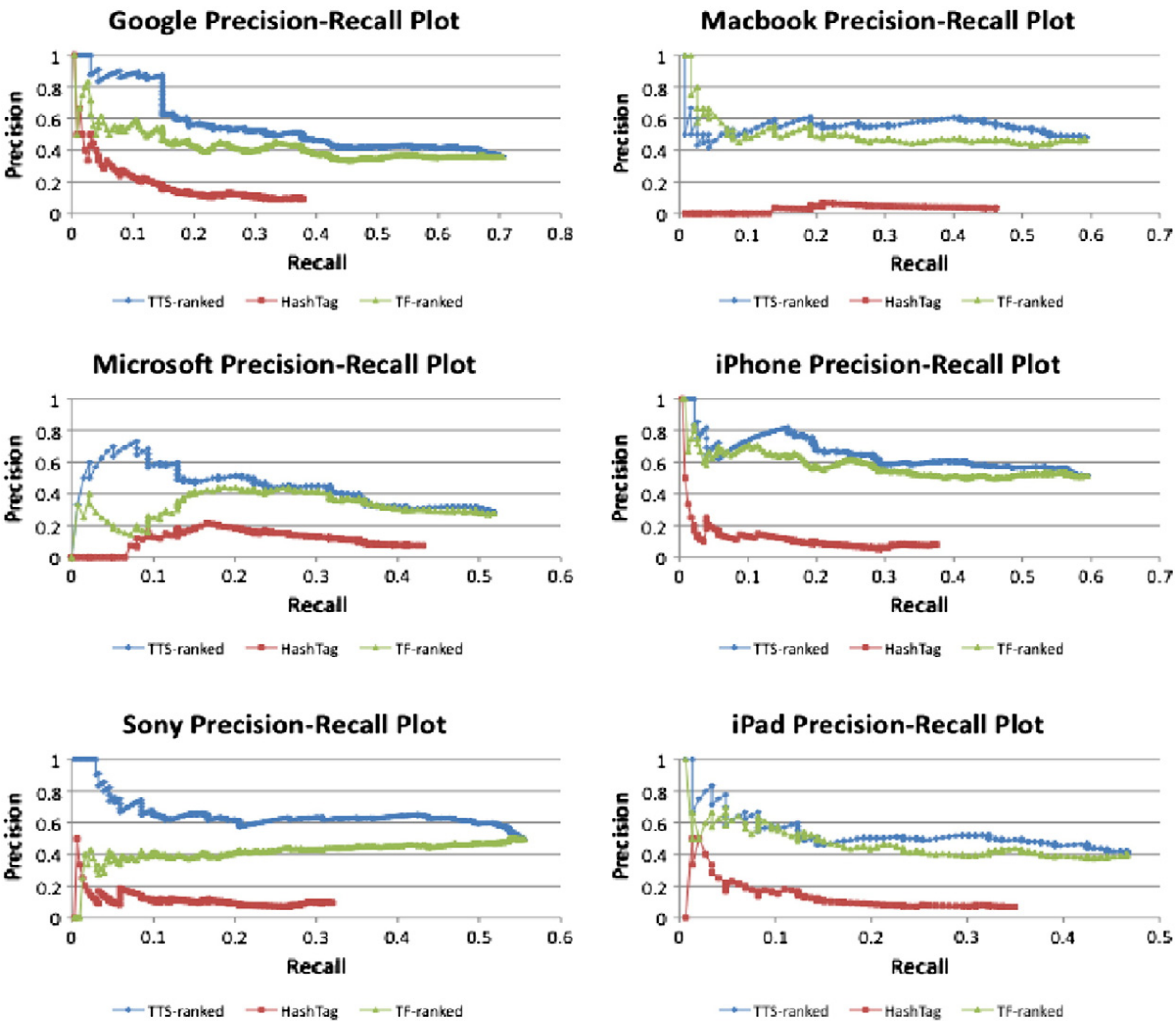


Fig. 2. Precision-recall plots for target queries.

Table 5
Precision rate (%) with the top *k* terms ranked with TSS picked as a relevant topic.

K Query	10	20	30	40	50	Average
Google	100	100	97.4	94.3	94.5	97.3
Microsoft	66.4	69.5	73.7	67.4	65	68.7
Sony	67.3	69.9	59	61.6	60.6	63.7
iPhone	100	100	94.7	83.1	80.7	91.7
iPad	39.5	62.2	55.2	53	52.1	52.4
Macbook	63.7	60	56	57.6	58.7	58.7
Average	72.8	76.9	72.7	69.5	68.6	72.1

to "#googlecitytour". Thus, it is not easy to separate words from the abbreviation and this significantly weakens the accuracy of identifying topic terms using the hashtag method.

The precision rate with respect to the top *k* terms picked is shown in Table 5. The TSS scoring function works very well on certain target queries (Google, iPhone) while yielding a normal performance on the others. A possible reason is that our scoring function weights terms on their frequency of appearing in meronym patterns. An example is the "Google" query, as we observed that many services and products mentioned are in a "Y X" pattern, such as "Google Maps" or"Google Docs". Another observation is that a POS tagger trained with document-based corpus performed poorly on tagging microblog posts because of highly different text formats. Many phrases were incorrectly identified as a relevant topic because of wrong POS tagging. For example, the incorrect topic "cool stickers" was identified by the query "Macbook" since the POS tagger tagged "cool" as a noun. Although the posts in microblogs are less structured and formal, the average precision rate (72%) of microblogs is still comparable to that of the review articles published in product review websites or blogs.

4.3. Sentiment classification with SVM classifier

In this section, we describe the steps of SVM training and evaluate the accuracy of the trained SVM provided. As mentioned in subsection 3.2.2, the training data set was gathered automatically based on emoticons. For preparing the training data set, we drew out 11,929 posts from the data collected in period 1 with two emoticons ":)" and ":(". Posts that contained ":)" were labeled with a positive class "+1", while posts that contained ":(" were labeled with a negative class "–1". If the posts contained both ":)" and ":(", then we discarded them. After the automatic labeling process, there were 7510 positive posts, 3,947 negative posts, and 236 discarded posts. Then, positive and negative posts were randomly split into five groups. The first four groups with 9165 posts altogether were used as the training data set and the remaining 2292 posts were used as the test data set.

Before using these data sets, several preprocesses were employed to reduce the number of features. First, we removed the target query and topic terms to avoid the classifier classifying sentiment by particular queries or topics. Second, numbers in posts were replaced with a special token "NUMBER_OR_VERSION". Third, we added a prefix "NOT_" to any words after a negative word in every sentence. The negative words we defined were "not", "never" and every words end with "n't". Lastly, all other words were stemmed with a Porter Stemming algorithm [43].

Next, we extracted different feature sets and evaluated their accuracy levels. Unigrams and bigrams are one-word and two-word tokens extracted from the preprocessed posts. Two types of accuracy were reported. The first accuracy was a five-fold cross-validation accuracy and the second was an accuracy yield when we use trained SVM to predict the sentiment of the test data. In addition to the SVM classifier, we also provided the accuracy of a Naïve Bayes classifier based on identical feature sets.

As shown in Table 6, our result is similar to the work of [33]. The simple unigram feature set provides the best accuracy both in Naïve Bayes and in the SVM classifier. The SVM approach provides a better accuracy rate over Naïve Bayes in the task of sentiment classification. The reason that the unigram feature set outperforms the other feature sets in microblogs is because more informal and newly invented terms are used to express sentiment and this fact negatively affects the accuracy of the subjective word set feature set because the subjective word set is derived from a dictionary. Besides, we can also observe that the results generated by the binary representation scheme of features are better than those generated by TF and TF × IDF representation schemes. Because the micro-blogging messages are generally very short, TF representation scheme is very similar to the binary representation scheme and their accuracy rates are quite close. However, as the IDF of a feature is becoming high, the TF/IDF weight of popular feature becomes low and the prediction accuracy deteriorates. Hence, in the topic score aggregation process, we adopted the SVM model with unigram features represented in binary scheme to classify sentiments.

4.4. Score aggregation correctness evaluation

In this section, we examine whether the summarized topic score based on our model is aligned with the real score assigned by users directly or not. However, there is no way to obtain a real numeric evaluation of the mentioned topics in the opinion expressers' minds, so we

Table 6
Accuracy from various features sets of SVM classifier.

	# of features	Naïve Bayes	SVM	
		Test data (%)	5-fold cross validation (%)	Test data (%)
Representation scheme — binary				
Unigram	11,802	71.7	90.4	88.1
Bigram	40,830	63.7	77.5	72
Unigram + bigram	52,632	60.74	87.4	81.3
Subjective word set	4206	34.2	67.7	63.6
Representation scheme — TF				
Unigram	11,802	70.3	89.9	87.7
Bigram	40,830	64.3	75.8	71.9
unigram + bigram	52,632	60.6	86.3	82.2
subjective word set	4206	33.1	68.3	63.5
*Representation scheme — TF*IDF*				
Unigram	11,802	67.4	82	79.4
Bigram	40,830	62.2	73.2	67.4
Unigram + bigram	52,632	54.6	72.7	72.5
Subjective word set	4206	30.3	61.8	62.2

Table 7
Questionnaires and responses.

	Target query questionnaire					
	Brand			Product		
	Google	Microsoft	Sony	iPhone	iPad	Macbook
# of topics	20	20	20	20	20	20
# of responses	88	70	59	40	33	63
# of valid topics	15	8	13	8	3	2

Table 8
MAEs of different aggregation methods on target queries.

Aggregation Method	NoWeight	WeightQuality	WeightCredibility	Proposed
Average MAE	0.5711	0.5515	0.5056	0.3384

utilize statistical approaches to verify the consistency between the scores summarized by the proposed system and general numeric evaluations from the public. Six questionnaires, each of which corresponds to a target query, were issued. In a questionnaire, the top 20 topics of the target query found by the topic detection module in the data collected in period #2 were listed. The respondents were requested to rate the topics on a five-point Likert scale (1 = very bad impression and 5 = very good impression). Moreover, to ensure that only the experienced users evaluated the topics of the target query, respondents were clearly informed to skip the topics that they had no ideas about and/or experiences.

All questionnaires were issued to the fans' pages related to the target queries on Facebook. There were three main reasons for issuing the questionnaire on Facebook. First, spreading information and obtaining responses to questionnaires is not easy on Twitter. Although Twitter is a big platform, there are no explicit user groups or communities formed by a specific brand or product. Hence, it is not straightforward to find out relevant Twitter users and ask them to fill in a questionnaire. Second, Facebook is now the largest social networking site, and the discussion and reviews on the fan pages are active. From the fans' pages, we could gather users' numeric evaluations on topics. Third, from a practical point of view, we could reach both Facebook and Twitter on the Internet. Given these concerns, we decided to select Facebook as an adequate proxy for Twitter and public opinions. The questionnaires were issued and collected in two weeks from 2010/05/24 to 2010/06/07. Table 7 outlines the information on the six questionnaires and responses.

Few responses were obtained on the “iPad” and “iPhone” questionnaires because these two products are relatively new to the market. A data set with too small a sample size may not effectively reflect the true numeric evaluation on topics by the public; therefore, only 49 topics with more than 30 responses and opinions in our data set were taken as valid and used. In this experiment, two metrics were considered. The first one was the mean absolute error (MAE). In statistics, the MAE error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The MAE is defined as:

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |f_i - y_i| \tag{13}$$

where f_i is the prediction and y_i the true value.

Since the scores provided by the aggregation module were continuous values ranging from −1 to +1, we conducted a linear transformation on the aggregated scores using the same five-point Likert scale as used in the questionnaires. The scaled score S' was transformed by following formulation:

$$S' = 3 + 2 \times S \tag{14}$$

where S is the score provided by the numeric aggregation module.

We compared the MAEs of all the valid topics calculated by our aggregation method and three other benchmark score aggregation methods. The first aggregation method weights nothing and it is formulated as:

$$NoWeight_{q,t} = \frac{\sum_{o \in O_{q,t}} Polarity_o}{|O_{q,t}|} \tag{15}$$

The second aggregation weights opinion subjectivity only and it is formulated as:

$$WeightSubjectivity_{q,t} = \frac{\sum_{o \in O_{q,t}} Polarity_o \times OS_{o,t}}{\sum_{o \in O_{q,t}} |OS_{o,t}|} \tag{16}$$

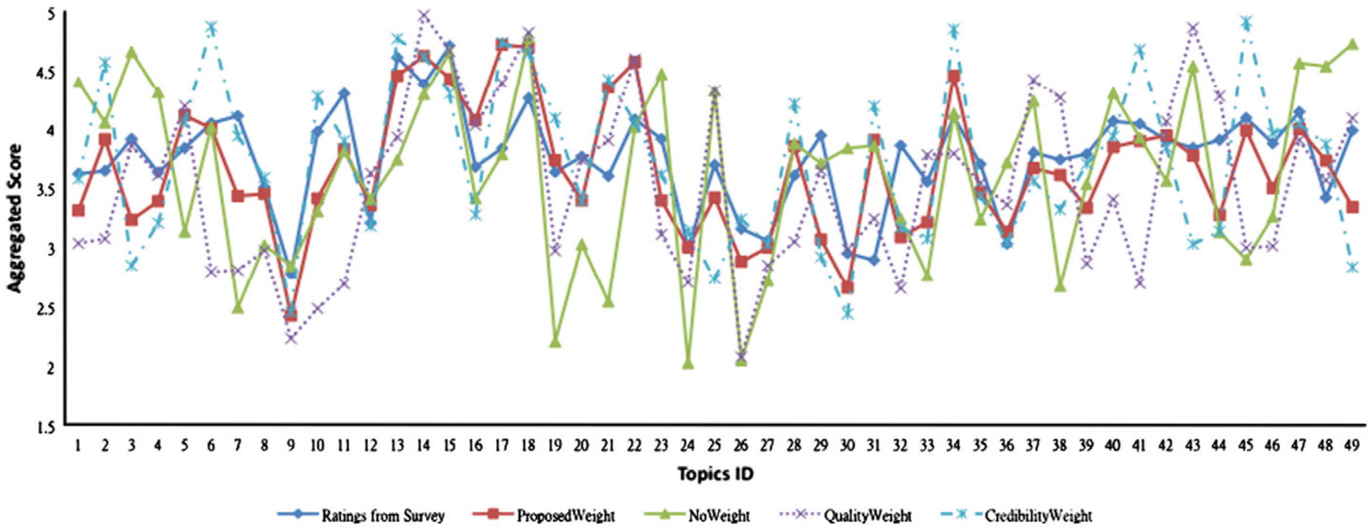


Fig. 3. Comparisons of aggregated topics scores.

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

	Paired difference					*t*	df	sig. (2-tailed)
	Mean	Std. Dev.	Std. error mean	95% confidence interval of the difference				
				Lower	Upper			
NoWeight	0.20973	0.63620	0.09089	0.02699	0.39247	2.308	48	0.025
WeightQuality	0.20212	0.65785	0.09398	0.01317	0.39108	2.151	48	0.037
WeightCredibility	0.13587	0.55882	0.07983	0.02465	0.29638	1.702	48	0.095
Proposed	0.09863	0.41604	0.05943	0.02087	0.21813	1.659	48	0.104

The third method weights credibility of opinion expresser only and it is formulated as:

$$WeightCredibility_{q,t} = \frac{\sum_{o \in O_{q,t}} Polarity_o \times CS_i}{\sum_{o \in O_{q,t}} |CS_i|} \qquad (17)$$

Fig. 3 shows the topic scores aggregated by different weighting methods. The result generated from our proposed weight method was the closest to the user ratings from the survey. The scores calculated based on the NoWeight method fluctuated more than those calculated according to other approaches did.

Table 8 displays the average MAEs of our aggregation methods and the three benchmark methods. The MAE statistics also indicate that the topic scores aggregated by our method are closest to the public viewpoint.

We further conducted a pair-wise *t*-test to verify whether it is statistically significant that the scores are consistent with public opinions. Here, we hypothesized that the numeric summarization of microblog opinions should be consistent with the ratings given directly by questionnaire respondents since they should both represent the viewpoints of the public. As shown in Table 9, the scores aggregated by the NoWeight and WeightSubjectivity aggregation methods are significantly different from public opinions. WeightCredibility and our proposed weighting method are both insignificantly different. This implies these two aggregation methods provide an aggregation of topic scores close the public opinions. The results reveal several implications. However, the proposed method is still better than is the WeightCredibility method because of having a smaller mean difference. First, it is crucial to take the opinion expresser's credibility into account as there are numerous extreme opinions on microblogs. Second, weighting on opinion subjectivity could alleviate the inability of SVM to filter out neutral opinions while the opinion expresser's credibility still needs to be taken into account. Third, weighting opinion subjectivity and expresser credibility could alleviate the interference of less relevant or neutral opinions and opinions expressed by less credible sources. Fourth, although weighting on credibility could yield a result consistent with the public viewpoint, combining credibility and opinion subjectivity is helpful to further minimize the difference between public scores and aggregated scores.

4.5. Topic-specific sentimental polarity

The current analysis of opinion polarity is opinion-specific. To extend the sentiment classification to topic-specific level analysis may generate more subtle results. One of the possible approaches is to break each post into sentences by the subject and calculate/accumulate scores regard to the subject. A post with multiple sentences could be separated into different posts, each of which contains one of the divided sentences and only the post containing the topic is analyzed. Sentences are separated by the symbols, such as ",".''!''?". Table 10 outlines the results generated by the revised sentimental classification approach. Comparing the results shown in Table 8 and Table 10, we can find that the performance (MAE) is only slightly improved (from 0.33384 to 0.33381) as that most of the opinions expressed in tweets include only one sentence. According to our collected dataset, only 1.27% of the tweet opinions contain more than one sentence. That is, the experiment shows that in analyzing the microblogs, which length is limited to be short, the accuracy of opinion-specific polarity is quite close to that of topic (sentence)-specific polarity.

5. Discussion and conclusion

Research on market intelligence (MI) systems strives to gather, analyze, and provide insightful knowledge for business decision-makers. In addition to the analysis of internal information such as customer purchasing history and financial numbers, businesses should scan external factors such as customers' tastes and competitors' activities. With more and more customers expressing their opinions on brands and products via social media platforms such as microblogs, it is increasingly important to discover and trace useful insights from microblog platforms to provide MI for business managers and consumers. However, the number of posts produced is overwhelming and the text format is not structured enough to make the process of knowledge extraction efficient. In this paper, we proposed a system designed to summarize text opinions into traceable numeric scores in which users are interested. Within the proposed system framework, it is easy to establish a monitoring system to track external opinions on different aspects of a business in real time. The proposed system allows decision makers to understand market trends by tracking the fluctuation of sentiments on particular topics presented in the proposed numeric summarization format. Most importantly, the aggregated scores are representative of public opinions.

5.1. Research contributions

The contributions of this research are summarized as follows. On theoretical aspects, first, because microblog posts are less structured than are traditional blog articles or documents, we improve the precision of topic detection in microblogs by combining the refined meronym patterns and term frequency. The precision of the proposed topic detection module is comparable with the current literature in which experimental data sources are usually in more formal structures. Second, the performance of an SVM as a sentiment classifier for microblogs is justified, although the text of microblogs is short. The result is similar to previous works that unigram features provide the best accuracy of other feature sets. Another noteworthy part is our survey reveals that it is applicable to use emoticons as a proxy for expressed sentiments, which allows us to feasibly quantify a huge number of opinions expressed in microblogs. Third, as the microblog message can be disseminated quickly over the social networks of users,

Table 10
Enhanced (Topic-specific) MAEs of different aggregation methods.

Aggregation method	NoWeight	WeightSubjectivity	WeightCredibility	Proposed
Enhanced average MAE	0.518	0.5522	0.5042	0.3381

in order to detect and avoid the spamming problem, we develop a model to quantify the credibility of an expresser. The scores aggregated from microblogs are mostly close to public viewpoints when author credibility and opinion subjectivity are taken into consideration. This result makes it possible to automatically gather public opinion from microblogs rather than performing market research.

On managerial aspects, using the proposed system, marketers could learn what topics are interesting to customers in real time as well as cost efficiently. Furthermore, the sentiments on these topics can be easily traced over time. Brand managers and marketers have to read many posts to know how users feel about their services and products and have to repeat "search-and-read" loops to know how customers evaluate their products. Marketers could effectively comprehend changes in customers' attitudes by time period and specific campaigns or events. Numeric aggregation also makes it effective and clear to compare the business to its competitors. Managers could develop a competitive advantage this way. Furthermore, the proposed system prevents the information used to make marketing decisions being interfered by irrelevant opinions. In a more practical context, a BAM system based on the proposed system could provide dashboards for external perceptions of the business; hence, managers could make subtle and informed decisions with a better understanding of outside information. For example, when a marketing campaign is conducted, marketing managers could understand the popularity, reaction, and engagement of customers to the campaign immediately.

5.2. Research limitation and future works

There are some limitations in our research. First, owing to the API call limitation of our experimental platform, the number of opinions used in system evaluation is limited. Second, because of the constraints of time and human resources, the valid topics and target queries used for this evaluation are limited. However, we believe that several extensive works can be studied. First, meronym patterns play a significant role in the topic detection module. Nevertheless, we use a list of predefined meronym patterns. A better approach is to apply data mining techniques to find out the most frequently used meronym patterns in microblog platforms. Besides, different microbloggers may use different terms for the same topic. The analysis of term synonyms could be an interesting future study topic. Second, the accurate analysis of opinion expressers can effectively prevent malicious spamming behaviors. Our current credibility measurement does not take a user's profile or basic information into account. Detailed profile and basic information may be a factor in credibility. Third, the evaluation of authority may be a good proxy for credibility. The performances of the adopted algorithms such as HITS and PageRank for credibility examination should be checked. Fourth, the system could be extended to include competitive intelligence discovering. An extended work could be performed to examine the predicted rankings from customers on the homogeneous products or services provided by different companies. With this customer ranking, enterprises can know their relative strengths and weaknesses and plan business strategies accordingly.

Acknowledgement

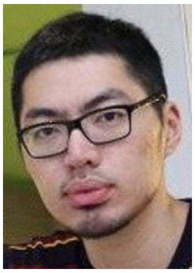
This research was supported by the National Science Council of Taiwan (Republic of China) under the grant NSC 99-2410-H-009-035-MY2.

References

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, Topic detection and tracking pilot study: final report, Proceedings of the DARPA Broad- cast News Transcription and Understanding Work- shop, 1998.
[2] N. Archak, A. Ghose, P. Ipeirotis, Show Me the Money!: deriving the pricing power of product features by mining consumer reviews, ACM SIGKDD Conference on Knowledge Discovery and Data Mining ACM, 65, 2007.
[3] Xue Bai, Predicting consumer sentiments from online text, Decision Support Systems 50 (4) (2005) 732–742.
[4] A.G. Biichner, M.D. Mulvenna, Discovering internet marketing intelligence through online analytical web usage mining, SIGMOD Record 27 (1998) 54–61.
[5] D. Boyd, S. Golder, G. Lotan, Tweet, tweet, retweet: conversational aspects of retweeting on twitter, Proceedings of the HICSS-43, Kauai, HI2010, 2010.
[6] J. Caverlee, L. Liu, S. Webb, The SocialTrust framework for trusted social information management: architecture and algorithms, Information Sciences 180 (1) (2010) 95–112.
[7] M. Cha, H. Haddadi, F. Benevenuto, K. Gummadi, Measuring user influence in twitter: the million follower fallacy, ICWSM'10: Proceedings of international AAAI Conference on Weblogs and Social Media, 2010.
[8] B. Connor, R. Balasubramanyan, B. Routledge, N. Smith, From tweets to polls: linking text sentiment to public opinion time series, Proceedings of the International AAAI Conference on Weblogs and Social Media 2010, 2010.
[9] N. Diakopoulos, D. Shamma, Characterizing debate performance via aggregated twitter sentiment, Proceedings of the 28th international conference on Human factors in computing systems Atlanta, Georgia 2010, 2010.
[10] X. Ding, B. Liu, L. Zhang, Entity discovery and assignment for opinion mining applications, 15th ACM SIGKDD international conference on Knowledge discovery and data mining ACM, Paris, France, 2009, pp. 1125–1134.
[11] W. Duan, B. Gu, A.B. Whinston, Do online reviews matter? An empirical investigation of panel data, Social Science Research Network (SSRN) Working Paper Series, 2005, http://ssrn.com/paper=616262, version as of January, 2005.
[12] M. Hearst, Direction-based text interpretation as an information access refinement, Text-Based Intelligent Systems, Lawrence Erlbaum Associates, 1992, pp. 257–274.
[13] C. Hsu, C. Chang, C. Lin, A practical guide to support vector classification citeseer, Taipei, 2003.
[14] M. Hu, B. Liu, Mining opinion features in customer reviews, 19th National Conference on Artificial Intelligence (AAAI-2004), 2004, pp. 755–760.
[15] N. Hu, I. Bose, Y. Gao, L. Liu, Manipulation in digital word-of-mouth: A reality check for book reviews, Decision Support Systems 50 (3) (2011) 22–30.
[16] B.J. Jansen, M. Zhang, K. Sobel, A. Chowdury, Twitter power: tweets as electronic word of mouth, Journal of the American Society for Information Science and Technology 60 (2009) 2169–2188.
[17] A. Java, X. Song, T. Finin, B. Tseng, Why we twitter: understanding microblogging usage and communities, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, San Jose, California 2007 ACM, 2007.
[18] H. Kanayama, T. Nasukawa, Fully automatic lexicon expansion for domain-oriented sentiment analysis, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia 2006, Association for Computational Linguistics, 2006.
[19] A.M. Kaplan, M. Haenlein, Users of the world, unite! The challenges and opportunities of social media, Business Horizons 53 (2010) 59–68.
[20] W. Kim, O.-R. Jeong, S.-W. Lee, On social web sites, Information Systems 35 (2010) 215–236.
[21] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (1999) 604–632.
[22] B. Krishnamurthy, P. Gill, M. Arlitt, A few chirps about twitter, Proceedings of the Proceedings of the first workshop on Online social networks, Seattle, WA, USA 2008 ACM, 2008.
[23] P. Larry, B. Sergey, R. Motwani, T. Winograd, The pagerank citation ranking: bringing order to the web, Technical Report Stanford University, 1998.
[24] D. Lee, O. Jeong, S. Lee, Opinion mining of customer feedback data on the web, 2nd international conference on Ubiquitous information management and communication ACM, 2008, pp. 230–235.
[25] Q. Li, J. Wang, Y.P. Chen, Z. Lin, User comments for news recommendation in forum-based social media, Information Sciences 180 (24) (2010) 4929–4939.
[26] W.G. Mangold, D.J. Faulds, Social media: the new hybrid element of the promotion mix, Business Horizons 52 (2009) 357–365.
[27] P. Manoj, B.W. Andrew, Research issues in social computing, Journal of AIS 8 (2007) 336–350.
[28] M.J. Metzger, Understanding how internet users make sense of credibility: a review of the state of our knowledge and recommendations for theory, policy, and practice, Internet Credibility and User Symposium, Seattle, Washington 4020 (805) (2005) 1–32.
[29] S. Milstein, A. Chowdhury, G. Hochmuth, B. Lorica, R. Magoulas, Twitter and the micro-messaging revolution: communication, connections, and immediacy—140 characters at a time O'Relly, 2008.
[30] M. Naaman, J. Boase, C.-H. Lai, Is it really about me? Message content in social awareness streams, Proceedings of the 2010 ACM conference on Computer supported cooperative work, Savannah, Georgia, USA 2010 ACM, 2010.
[31] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (2008) 1–135.
[32] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing 2002 Association for Computational Linguistics, 2002.
[33] A. Popescu, O. Etzioni, Extracting product features and opinions from reviews, Natural Language Processing and Text Mining (2005) 9–28.
[34] S.Y. Rieh, D.R. Danielson, Credibility: a multidisciplinary framework, in: B. Cronin (Ed.), Annual review of information science and technology, Medford, New Jersey: Information Today, 41, 2007, pp. 307–364.

[35] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, C. Jin, Red opal: product-feature scoring from reviews, 8th ACM Conference on Electronic Commerce ACM New York, NY, USA, 2007, pp. 182–191.
[36] H. Sundblad, Automatic acquisition of hyponyms and meronyms from question corpora, ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering (OLT 2002) Citeseer, Lyon, France, 2002.
[37] J. Tatemura, Virtual reviewers for collaborative exploration of movie reviews, Proceedings of the 5th International Conference on Intelligent User Interfaces, New Orleans, Louisiana, United States 2000 ACM, 2000.
[38] K. Toutanova, D. Klein, C. Manning, Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network, HLT-NAACL 2003, 2003, pp. 252–259.
[39] P. Turney, M. Littman, Unsupervised learning of semantic orientation from a hundred-billion-word corpus, Technical Report Erb-1094 National Research Council, 2002.
[40] Twitter.Com Twitter 101 - Case Study: Best Buy's Twlpforce, 2010.
[41] Twitter.Com Twitter 101 - Case Study: Dell, 2010.
[42] B. Ulicnya, K. Baclawskia, A. Magnusb, New metrics for blog mining, SPIE Defense & Security Symposium, Orlando, FL, 2007.
[43] C. Van Rijsbergen, S. Robertson, M. Porter, New models in probabilistic information retrieval, British Library, London1980.
[44] C. Wathen, J. Burkell, Believe it or not: factors influencing credibility on the web, Journal of the American Society for Information Science and Technology 53 (2002) 134–144.
[45] S. Wright, J.L. Calof, The quest for competitive, business and marketing intelligence, European Journal of Marketing 40 (2006) 453–465.
[46] R. Xia, C.Q. Zong, S. Li, Ensemble of feature sets and classification algorithms for sentiment classification, Information Sciences 191 (6) (2010) 1138–1152.

Yung-Ming Li is a Professor at the Institute of Information Management, National Chiao Tung University in Taiwan. He received his Ph.D. in Information Systems from the University of Washington. His research interests include network science, Internet economics, and business intelligence. His research has appeared in IEEE/ACM Transactions on Networking, INFORMS Journal on Computing, European Journal of Operational Research, Decision Support Systems, International Journal of Electronic Commerce, Electronic Commerce Research and Applications, ICIS, WTIS, among others.



Tsung-Ying Li is a software engineer at Chunghwa Telecom in Taiwan. He received his M.S. degree from the Institute of Information Management, National Chiao Tung University in Taiwan and his B.S. degree in Information Management and Finance from the National Chiao Tung University, Taiwan. His research interests focus on social media, electronic commerce, and business intelligence.