Information & Management 52 (2015) 801–812

Contents lists available at ScienceDirect

Information & Management

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

A novel social media competitive analytics framework with sentiment benchmarks

Wu He [a,*], Harris Wu [a], Gongjun Yan [b], Vasudeva Akula [c], Jiancheng Shen [d]

[a] Department of Information Technology & Decision Sciences, College of Business and Public Administration, Old Dominion University, Norfolk, VA 23529, USA
[b] Department of Management and Information Sciences, Romain College of Business, University of Southern Indiana, Evansville, IN 47712, USA
[c] Head of Business Consulting, VOZIQ, Reston, VA 20190, USA
[d] Department of Finance, College of Business and Public Administration, Old Dominion University, Norfolk, VA 23529, USA

ARTICLE INFO

Article history:
Received 29 August 2014
Received in revised form 31 January 2015
Accepted 27 April 2015
Available online 12 May 2015

Keywords:
Social media analytics
Competitive analytics
Sentiment benchmarks
Text mining
Sentiment analysis
User-generated data
Social media
Marketing intelligence
Big data
Social media monitoring

ABSTRACT

In today's competitive business environment, there is a strong need for businesses to collect, monitor, and analyze user-generated data on their own and on their competitors' social media sites, such as Facebook, Twitter, and blogs. To achieve a competitive advantage, it is often necessary to listen to and understand what customers are saying about competitors' products and services. Current social media analytics frameworks do not provide benchmarks that allow businesses to compare customer sentiment on social media to easily understand where businesses are doing well and where they need to improve. In this paper, we present a social media competitive analytics framework with sentiment benchmarks that can be used to glean industry-specific marketing intelligence. Based on the idea of the proposed framework, new social media competitive analytics with sentiment benchmarks can be developed to enhance marketing intelligence and to identify specific actionable areas in which businesses are leading and lagging to further improve their customers' experience using customer opinions gleaned from social media. Guided by the proposed framework, an innovative business-driven social media competitive analytics tool named VOZIQ is developed. We use VOZIQ to analyze tweets associated with five large retail sector companies and to generate meaningful business insight reports.

1. Introduction

In recent years, we have observed the rapid development of social media, which has drastically transformed the way in which people communicate and obtain information. Currently, social media has become ubiquitous and plays an increasingly critical role in today's business environments. A number of companies use social media tools such as Facebook and Twitter to provide a variety of services and to interact with customers. As a result, a large amount of user-generated content is available on social media sites. User-generated content offers opportunities and challenges to businesses. In the business field, consumers increasingly rely on user-generated reviews to evaluate products and services prior to making a purchase. To increase competitive advantage and effectively assess the competitive business environment, companies not only need to monitor and analyze the customer-generated opinions about their businesses but also need to track opinions about their competitors. Studies indicate marked performance growth in companies that have strong business analytics capabilities [52].

Marketing intelligence is typically performed by businesses to collect and analyze data from both internal and external information sources. The mechanism for collecting and analyzing these sources is also called business intelligence (the portion that focuses on internal business data) and competitive intelligence (the portion that focuses on external business data) [40]. Marketing intelligence is considered important to the improvement of a firm's organizational performance [45]. In recent years, due to advances in social media technology, the amount of online social media data has grown explosively. Leskovec [29] proposes that user-generated content in the form of blog posts, comments, and tweets establishes a connection between companies and consumers. Thus, companies are expected to harness this user-generated data to extract entities and themes, to understand consumer sentiment, to visualize relationships and to create their marketing intelligence

* Corresponding author. Tel.: +1 757 683 5008; fax: +1 757 683 5639.
E-mail addresses: edu">whe@odu.?edu (W. He), hwu@odu.edu (H. Wu), gyan@usi.edu (G. Yan), vakula@voziq.com (V. Akula), jshen@odu.edu (J. Shen).

http://dx.doi.org/10.1016/j.im.2015.04.006
0378-7206/

to excel in the business environment. In particular, a marketing intelligence report can include market information about the popularity of competitors' products and services, consumer sentiments on their products and services, promotional information, and/or activities offered by competitors [14].

As one consequence of social media development, social media analytics has emerged as an important area of study [16,48]. Social media analytics is concerned with using advanced informatics tools and analytics techniques to collect, monitor, and analyze social media data to extract useful patterns and intelligence [31,48]. Therefore, the development of effective and efficient analytics techniques for social media analytics becomes essential. Data mining, text analysis, and sentiment analysis techniques are frequently adopted to conduct social media analytics [3–6,19,46]. Recently, there has been strong interest in the power of social media analytics to create new value, to support decision making and to enhance competitive advantage. For example, Abrahams, Jiao, Wang, and Fan [5] use social media analytics to discover a specific vehicle defect from social media to improve their automotive quality management.

One of the challenges in uncovering actionable insights is to properly interpret the meaning of both positive and negative sentiments in unsolicited customer opinions on social media. While there are advances in improving the sentiment accuracy itself, the core challenge for businesses to identify areas of improvement based on sentiment analysis remains. Furthermore, companies often want to determine how their performance stacks up against their key competitors' performance. Effective social media benchmarking can be used as a means to compare a company's performance with that of its key competitors or with the industry as a whole. In an effort to help companies understand how to perform social media competitive analysis, transform social media data into actionable knowledge, and develop the ability to benchmark effectively, we propose a social media competitive analytics framework with sentiment benchmarks. This framework, first, calls for creating a sentiment benchmark of the industry (or of comparable businesses) and then using that industry sentiment benchmark to compare whether a business's social media sentiment is higher or lower. To further enhance the ability to drive business decisions, the framework uses industry-specific lexicons to quickly identify topics that might be of interest to businesses based on the verticals in which they operate. The entire framework leverages emerging technologies, including text mining, sentiment analysis, and social network analysis. Based on the idea of the proposed framework, new social media competitive analytics tools can be developed to identify actionable marketing intelligence.

The remainder of the paper is organized as follows. Section 2 provides a review of text mining and sentiment analysis. Section 3 proposes a social media competitive analytics framework for marketing intelligence. Section 4 describes an innovative social media competitive analytics tool named VOZIQ that we developed and offers an example of using VOZIQ for analyzing social media performance in five large retail sector companies. Conclusions and future research are given in Section 5.

2. Literature review

2.1. Text mining

Today, numerous customers and users share their experiences using various social media sites such as Twitter, Facebook and blogs. It has become a challenge for organizations to monitor and understand what people post on social media sites. Traditional content analysis methods are no longer able to meet organizations' needs to analyze the large amount of new content on a daily basis. Applying automatic methods to quickly analyze such content is increasingly needed by organizations. As users continue to post textual information on various social media sites, there is a growing interest in using text mining, sentiment analysis and social network analysis approaches to process large amounts of user-generated data and extract meaningful knowledge and insights.

As an emerging technology, text mining aims to extract meaningful information from unstructured textual data [19,23,33]. To glean useful information from a large number of textual documents quickly, it has become imperative to use automated computer techniques [20,33]. Text mining is focused on finding useful models, trends, patterns, or rules from unstructured textual data [2,23,39]. Different from traditional content analysis, the main purpose of text mining is to automatically extract knowledge, insights, useful patterns or trends from a given set of textual documents [21,50].

Text mining techniques have been used to analyze large amounts of textual data. Morinaga et al. [36] present a framework for mining public opinions related to product reputation on the Internet. They find that text mining techniques offer both a dramatically reduced cost and increased knowledge discovery from public opinion, compared with the conventional survey approach. Kloptchenko et al. [27] use data and text mining methods to analyze the textual part of a company's financial report. They find that the mining results can be used to predict the company's future financial performance to some extent. Abdous and He [2] use text mining techniques to analyze the online questions posted by video streaming students and identify a number of learning patterns and technology-related issues. Hung [22] uses clustering analysis as an exploratory technique to examine e-learning literature and visualizes patterns by grouping sources that share similar words, attribute values and coding rules.

Some major applications of text mining include: cluster analysis, categorization, information extraction (text summarization), and link analysis [22,50]. In particular, cluster analysis is a key application of text mining and includes four main building blocks: feature selection, the clustering algorithm, the validation of the results, and the interpretation of the results [17]. By dividing a population into clusters that are different from one another (maximal distance between clusters) but whose members are similar (minimal distance within each cluster), cluster analysis can enhance the understandability of datasets and support effective decision making [25]. Currently, there are a wide range of tools that can be used for text mining and analytics, such as IBM SPSS Modeler (formerly Clementine), Semantria, Lexalytics, Leximancer, Clarabridge and SAS Enterprise Miner. Due to the powerful capabilities of text mining, it is believed that applying text mining to textual data, including messages posted on social media such as blogs, can yield interesting findings [6,11,21].

2.2. Sentiment analysis

Sentiment analysis is the computational detection and study of opinions, sentiments, emotions, and subjectivities in texts [30,32,37]. As a special application of text mining, sentiment analysis is concerned with the automatic extraction of positive or negative opinions from texts [37]. Given that texts often contain a mix of positive and negative sentiments, it is often useful to identify the polarity of sentiment in texts (positive, negative, or neutral) and even the strength of the sentiments expressed [37,44]. Sentiment analysis mainly relies on machine learning techniques, such as Support Vector Machine (SVM), Naive Bayes, Maximum Entropy and Matrix Factorization, to classify texts into positive or negative categories [30,38].

There is a growing interest in using sentiment analysis methods to mine user-generated data. Sentiment analysis has been used to determine the attitude of customers and online users on some specific topics, such as consumer product (e.g., books, movies, electronics) reviews, hotel service reviews, public relations statements, and financial blogs. Bollen et al. [8] use sentiment analysis to mine a large corpus of Twitter messages to determine the mood of the Twitter population on a given day. They find that the mood of the Twitter population is able to predict the movement of the Dow Jones Industrial Average (DJIA) on the following day with a claimed 87.6% accuracy. Sprenger and Welpe [41] use the sentiment analysis of tweets gathered for the top 100 stocks in the Standard & Poor's index (S&P 100) and are able to show a consistent correlation between Twitter sentiment and stock market returns. Ludwig et al. [34] use text mining methods to "extract changes in affective content and linguistic style properties of customer book reviews on Amazon.com." They find a positive asymmetrical relationship between positive affective content and the conversion rates on the website. Duan, Cao, Yu, and Levy [15] use the sentiment analysis technique to mine 70,103 online user reviews posted in various online venues from 1999 to 2011 for 86 hotels in Washington, DC. Sentiment analysis helps them decompose user reviews into five dimensions to measure hotel service quality, and the sentiment analysis results show a high level of accuracy in capturing and measuring service quality dimensions, compared with existing text mining studies. Klein, Altuntas, Riekert, and Dinev [26] propose a novel approach to extracting financial instrument-specific investor sentiment from a set of blog articles. Their results suggest that extracting investor sentiment about future returns of financial instruments is a useful approach for investment managers and other stakeholders in the financial industry. Lee et al. [28] use both data mining and sentiment analysis techniques to analyze the dataset collected from MyStarbucksIdea, one of the most popular online open innovation communities. Specifically, they use sentiment analysis to extract the sentiment contained in each idea and comment collected from the MyStarbucksIdea website. The experimental results have been used to help them build a recommendation system that can help firms identify prospective ideas for innovation from among a large amount of ideas. Stieglitz and Dang-Xuan [42] suggest that companies should pay more attention to the analysis of sentiment related to their brands and products in social media communication.

2.3. Competitive intelligence and analytics

Kahaner [24] defines competitive intelligence as "a process of monitoring the competitive environment, with a goal to provide actionable intelligence that will provide a competitive edge to the organization." The main goal of competitive intelligence (CI) is to monitor a firm's external environment for information that is relevant to its decision-making process [12]. CI allows a company to identify its competitors' strengths, weaknesses, strategies and other areas and in turn help the company improve its strategic decisions against its competitors [9,43].

CI tools and techniques can be classified into two categories: data collection tools and data analysis tools. Companies often use certain tools such as web search engines and web crawlers to collect data and then use other tools to analyze the data. CI analysis tools and techniques are mainly based on text mining, web mining and visualization technologies [9]. Traditional CI tools are mainly developed to collect data from webpages, blogs, online reports, emails and online text reviews. For example, Chen, Chau and Zeng [12] develop a tool called CI Spider that can collect webpages from sites specified by the user, perform data analysis and provide the user with a comprehensive view of the websites based on user interest. Xu, Liao, Li and Song [47] develop a graphical model to extract and visualize comparative relations between products from customer reviews on the Amazon website for competitive intelligence.

With the rapid growth of social media in business, we have observed large numbers of customer-generated social media data that contain information about competitors. Unfortunately, the aforementioned tools or methods have not been updated in a timely fashion to conduct social media competitive analytics with sentiment benchmarks for industry-specific marketing intelligence.

A recent literature review reveals that there are only a few studies that focus on social media competitive intelligence and analytics in business, despite the fact that a considerable amount of research is devoted to analyzing the data presented in social media [6]. He, Zha and Li [19] analyze the unstructured textual content on the Facebook and Twitter sites of the three largest pizza chains: Pizza Hut, Domino's Pizza and Papa John's Pizza. Their results reveal the business value of comparing social media content. Dey, Haque, Khurdiya and Shroff [14] use text mining to gather competitive intelligence about competing products and companies. They find that social media data can be used to derive competitive intelligence, study the correlation between rival brand promotion events and sales data, and consumer sentiments. To date, there is a lack of published articles in the literature that discuss social media-based competitive intelligence and analytics frameworks or tools specifically designed for collecting, mining, analyzing and visualizing social media data from different companies or brands. In this regard, the VOZIQ tool that we present in this paper is a pioneering effort.

3. Our proposed method

3.1. The proposed framework

Fig. 1 lists the proposed social media competitive analytics framework with sentiment benchmarks for industry-specific marketing intelligence. We propose to identify the leading companies in particular industries, such as technology, banking, retail and real estate, compare their social media mentions for competitive analysis, and create industry-specific sentiment benchmarks for marketing intelligence and decision making. The data collection technique simply involves using publicly available APIs from social media sites and, if APIs are not available, crawling specific websites and parsing HTML as needed to collect review comments. The result of the sentiment benchmark analysis can be used to compile reports that show the variances between a company's key performance metrics and its peer group benchmarks. Each variance can either show in which areas a company is genuinely excelling or show a potential problem area to be fixed and to highlight the opportunity to improve the company's overall performance. Below is a detailed description of the procedures.

First, organizations need to select a few leading companies in their target industry and identify their social media sites for competitive analysis. The sentiments of these companies will become the basis for social media benchmark to compare against. Businesses can choose comparable businesses based on geographical proximity, revenue range, customer base, or products and services offered. Social media such as Facebook, Twitter, and blogs offer a huge amount of user-generated content on the Internet. Because there are numerous social media sites, organizations need to decide which social media site(s) they are going to use in terms of conducting competitive analytics. In addition, a business should establish effective and realistic benchmarks to measure and monitor their social media efforts against competitors. Some examples of social media measurements and metrics include:

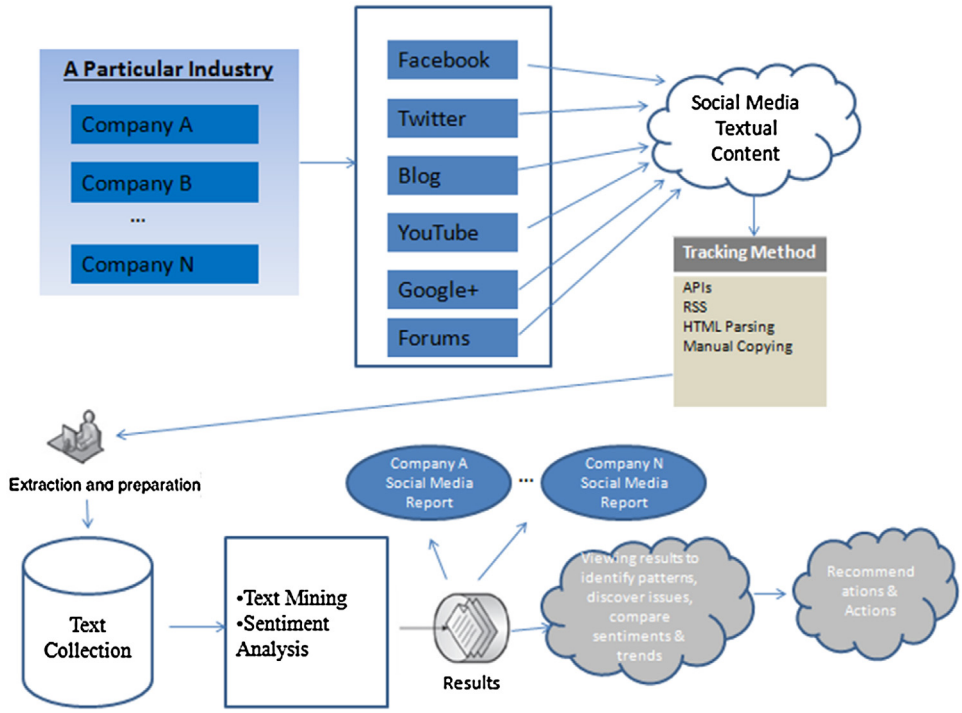


Fig. 1. A social media competitive analytics framework with sentiment benchmarks for industry-specific marketing intelligence.

number of fans/followers; number of postings, comments, likes, tweets and retweets; frequency of posting; posting and response time, etc. In addition to quantitative measurements, it is also necessary to establish qualitative metrics to assess what is being said in the texts, such as sentiments or emotions, and why those items were shared. A business can use these measurements and metrics to compare their social media efforts against competitors' social media efforts and to determine what they can do, either to get ahead or to make improvements.

Second, organizations need to monitor the selected social media sites and collect user-generated data posted on those sites constantly. He, Zha and Li [19] recommend that companies use social media monitoring tools such as Google Alerts, Social Mention, Quora, HootSuite, and Advanced Twitter Search to constantly monitor their own social media presence and their competitors' social media presence. Some social media sites, such as Twitter, Facebook, and YouTube, offer application programming interfaces (APIs) for data tracking. Organizations can have their programmers create custom applications and track the data using these APIs. By contrast, blogs and online forums typically do not provide APIs for data tracking. However, most blogs and online forums offer RSS feeds that can be easily tracked. For those without RSS functions, manual copying or web crawling techniques such as HTML parsing can also be used to collect data, although they may be more time-consuming [42].

Third, a data pre-processing step is needed to ensure that raw data are transformed into a usable format, mainly by data cleaning. Subsequently, a combination of various text mining, sentiment analysis, and traditional social network analysis techniques can be used to examine the data sets to gain insights into users' social media activities and sentiments. For example, organizations can use these methods to identify influential online users, to understand consumer sentiments about brands or about specific promotions or campaigns, to discover potential issues or problems posted by consumers, to extract emotionally charged topics, and to recognize business risks [42]. Organizations can use existing software tools such as SAS Miner, IBM SPSS Modeler, Leximancer, NVivo, LingPipe, Semantria, Lexalytics, SentiStrength, Clarabridge, Pajem, and UCnet to help them perform text mining, sentiment analysis, and social network analysis. Many of these tools can effectively extract key concepts, perform query searches, generate categories, and help to quickly gain insights from the textual data. Query searches are mainly used to test ideas and to find interesting patterns, connections, and unusual information based on the research questions.

Finally, the results of the social media analytics should be carefully reviewed and then used to derive insights, create intelligence reports, support decision making or make recommendations. The results of the analysis should be presented in a way that the user can understand [35]. In particular, marketers can review the results to discover new knowledge (e.g., brand popularity) and interesting patterns, to benchmark industry sentiments and categories, to understand what their competitors are doing and how the industry is changing in various categories, and to use the findings and improved understanding to achieve competitive advantage against their competitors [14,18]. Decision makers can also use the findings to develop new products or services and to make informed strategic and operational decisions.

3.2. Key issues and solutions

As customers share an increasing amount of opinions about their experiences with the products and services they use, both positive and negative, there are several important and challenging issues for businesses to which computer automation analysis technologies can be applied. For example, one of the key challenges for companies while using social media sentiment benchmarks is to identify actionable areas from uncontrolled opinions expressed in an extremely large, uncontrolled open forum such as social media. Customers typically talk freely about inherently positive topics, such as promotions and new enhancements, and about inherently negative topics, such as returns and damaged products. When analyzing social media sentiment at the company and brand level without industry benchmarks, companies tend to focus on

areas where there is the highest negative sentiment. This may not be the right approach because the fact that businesses cannot continuously devise promotions or the best return policy is not necessarily a bad thing, from a business perspective. Each of these inherently positive or negative areas, first, needs to be compared across industry peers to accurately gauge whether a business should consider it as an area that should be addressed. Our proposed framework addresses this key industry challenge with the help of industry social media sentiment benchmarks by topic. We propose to address this scenario and to help businesses to focus on key opportunities by immediately comparing each topic against the same topic across the industry for other companies or brands.

Another challenge for companies is to fully understand what customers think about their products and services. When there are any problems in their operations, companies want to determine the reasons for those problems. Using sentiment analysis, companies are able to discover the root causes and drivers of positive and negative opinions.

To mine the reasons, we propose a solution based on data mining. There are two types of mining algorithms. The first is called supervised text mining. Using text mining, given a set of documents, we assume that there are N categories. We can manually determine N categories and put N_i text records for the ith category. These N categories with the classified text are used as training data. New text messages are sorted into categories. The second type is called unsupervised text mining. There is no need for human processing of text messages. All of the messages are clustered by data mining algorithms. For each cluster, a tag known as a topic is computed. The cluster can then be treated as a category. Once the categories are determined, the sorting of new text messages can be performed using text mining algorithms. Both the unsupervised and the supervised classification methods have advantages and disadvantages. The unsupervised method can be executed independently, but it sometimes generates less meaningful results. The supervised method requires human effort and intelligence, which may be labour intensive. Thus, we suggest combining both the unsupervised method and the supervised method as a hybrid method. First, we can use the supervised method to classify text comments into several main sentiment categories. In each category, we can then adopt the unsupervised method to conduct more in-depth analysis to discover new subcategories, themes, topics, or patterns. In our proposal, the main categories can often be defined as sentiment types: strong positive, positive, neutral, negative, strong negative. This proposed method quickly helps in discovering topics that are applicable across the industry, not only those specific that are to one brand. Over a period of time, effective lexicons can be developed to discover emerging topics (bottom up with unsupervised learning) and eventually convert them into structured categories (top down, with Boolean keyword-based rules). The feedback from the businesses that we interviewed during the methodology development has been positive, given that businesses generally want to track relevant topics and be made aware of any unforeseen emerging topics. Our proposed hybrid approach covers both needs highly effectively by combining the benefits of both unsupervised and supervised topic discovery methods.

In our proposed method, text messages are first classified into five sentiment categories. Next, we apply a topic analysis algorithm to mine the messages within each category to identify the causes that may explain its sentiment. For example, if a bank found that 40% of the comments it received were negative, its need for analysis after this discovery would be to arrive at the root cause and determine why there were so many negative comments. To this end, we can extract all the negative comments and then use unsupervised text clustering algorithms to perform some text mining, generating several clusters that will be tagged with keywords. The keywords associated with each cluster can then be treated as possible reasons for the negative comments.

Fig. 2 illustrates a hybrid approach that combines both the top-down (the supervised method) and the bottom-up method (the unsupervised method) for classifying texts into industry-specific categories of interest. The top-down method filters comment texts by logic filter. The logic filter mainly includes three types of rules: logic conjunction (i.e., logic AND), logic disjunction (i.e., logic OR) and logic exclusive (i.e., logic NOT). Users can determine which rules they want to apply to their text collection. For each category, the comment text that matches with the logic rules is placed in the appropriate category. The top-down method can specifically filter out the comment texts that do not match a user's interest. However, this method may not cover new topics and emerging categories that are related to the problem to be solved. The bottom-up method analyzes comment texts using text analysis methods such as natural language processing techniques (e.g., Latent Dirichlet Allocation, N-Gram) and clustering. The bottom-up method can be used to automatically process, mine, and cluster texts into the specified categories. Because some categories may not be of interest to the user, combining the two methods provides a benefit not only to obtain the best classification but also to help the user focus on the most relevant categories that meet his/her interests.

3.3. Key word optimization and the modified N-Gram

3.3.1. Modified Chi-square feature selection

A large number of messages can generate thousands, millions, even billions of words, which can result in cascading computational difficulty (e.g., the analyzing and training model) and the "curse of dimensionality" [7] if all of the words are counted in computing. An intuitive idea is to extract the features of each message and then to process similarity based on the extracted features instead of the entire messages. There are two steps used to extract representative features: (1) electing feature sets and (2) extracting feature values.

One of the difficulties in electing feature sets is to determine representative words without sacrificing the performance of the algorithm. To improve the accuracy and to reduce the computing time, we need to delete irrelevant or redundant features (words) in the messages to reduce the number of features. In the literature, there are several methods to use when electing representative word sets, for example, chi-square, local/global document frequency, information gain, etc.

We define the following tuples as possible permutations of feature selection [49]:

- $D_A = t, c_i : t \in c_i$.
- $D_B = t, \square\, c_i : t \in \square\, c_i$.

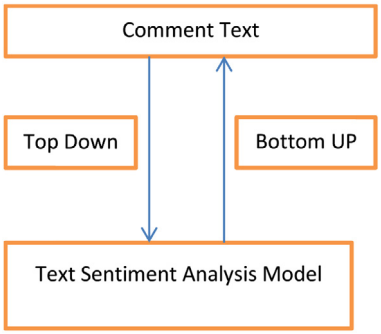


Fig. 2. Our proposed hybrid method.

- $D_C = \square\, t, c_i : t / \in c_i$.
- $D_D = \square\, t, \square\, c_i : t / \in \square\, c_i$.

where t is the term/word in a message and c_i represents a category. The above D_A, D_B, D_C, D_D are the four dependency tuples. The D_A and D_D are positive dependency between t and c_i. The D_B and D_C are negative dependency between t and c_i.

The chi-square method of feature selection can be defined as follows:

$$\chi^2(t, c_i) = \frac{N[P(t, c_i)P(\bar{t}, \bar{c}_i) - P(\bar{t}, c_i)P(t, \bar{c}_i)]^2}{P(t)P(\bar{t})P(c_i)P(\bar{c}_i)}$$

where N is the total number of documents, and each message is saved in one document. Chi-square assumes that t and c_i are independent of each other. According to the value of $\chi^2(t, c_i)$, t is independent of c_i if the value is very small, i.e., $\chi^2(t, c_i)$ is treated as a measurement error. Otherwise, t is not independent of c_i, i.e., t is the member of c_i.

Ideally, we need to select all of the terms that are the most indicative members of c_i. For each category c_i, the size of positive feature set F_i^+, noted as L_{F_i+}, present $0 < L_{F_i+} = L_{F_i}$ where L_{F_i} is the size of feature set F_i for c_i. Similarly, we can define the size of the negative-feature set F_i^- as L_{F_i+}, which satisfies $L_{F_i-} = L_{F_i} - L_{F_i+}$. Note that $F_i = F_i^+ \cup F_\iota^-$.

If the distribution of terms in each category is even, we simply choose L_{F_i+} term t-s with the highest indicator, say, $\chi^2(t, c_i)$. In reality, the number of categories and the distribution of terms in each category can significantly affect the effectiveness of the methods. Defining a function $\Im(, c_i)$ as the term t with larger value of $\Im(t, c_i)$ is the term that more likely belongs to the category c_i. One of most effective feature selection methods is [49]:

1. select L_{F+} greatest term t–s where $\Im(t, c_i)$ is sorted in decreasing order;
2. select L_{F-} smallest term t–s where $\Im(t, \square\, c_i)$ is sorted in increasing order;
3. optimize L_{F+} and L_{F-} for better performance.

The chi-square method may create a significant problem. Only the presence of terms is considered, while the frequency of the terms in a document is not considered. The frequency of terms, however, often conveys the greater importance of the term. Therefore, we have modified the chi-square method to consider the frequency of terms. We present

$$\Im(t, c_i) = a\frac{\chi^2(t, c_i)}{\sum_{j=1}^{N_c} \chi^2(t, c_j)} + b\frac{f(t, d_i)}{\sum_{j=1}^{N_d} f(t, d_j)}$$

where N_c is the total number of categories, $f(t, d_i)$ is a function that shows the frequency of term t in document d_i, N_d is the total number of document, and $a \in R_{\geq 0}$, $b \in R_{\geq 0}$ are coefficients on the condition $a + b = 1$ ($R_{\geq 0} = \{x \in R : x \geq 0\}$).

3.3.2. Modified N-Gram model

One difficult challenge is to determine how many categories should be focused on for each individual business. Our method is the following: we adopt a unigram, bigram, and trigram (i.e., $n = 1$, 2, 3 in our modified N = Gram, respectively) to find the keywords for categories. In this manner, we have a basic idea of how the text comments are clustered.

The basic idea of the N-Gram [10] method is to compute the probability that the term t will appear after specific sequence of n terms $t_1, t_2, \ldots, t_n$, i.e., $P(t_i|t_1, \ldots, t_{i-1})$.

The unigram model used in the natural language processing model can be represented as the probabilities of m terms in a context, i.e.,

$$P(t_1, t_2, \ldots, t_m) = P(t_1)P(t_2|t_1)P(t_3|t_1t_2)\cdots P(t_m|t_1t_2, \ldots, t_{m-1})$$

$$P_{\text{unigram}}(t_1, t_2, \ldots, t_m) \approx P(t_1)P(t_2)P(t_3)\cdots P(t_m)$$

In the unigram model, the probability of hitting every term only depends on itself.

Because there are limited terms in documents and computing is extremely expensive, the N-Gram is often assumed to have the Markov property, i.e., $P(t_i|t_1, \ldots, t_{i-1}) \approx P(t_i|t_{i-(n-1)}, \ldots, t_{i-1})$. That means that the probability of observing the ith term t_i in the context history of the continuous preceding $i-1$ terms can be roughly approximated by the probability of observing t_i in the shortened context history of the continuous preceding $n-1$ words (nth order Markov property). For an expression of terms that is composed by m terms, we write

$$P(t_1, t_2, \ldots, t_m) = \int_{i=1}^{m} P(t_i|t_1, \ldots, t_{i-1}) \approx \int_{i=1}^{m} P(t_i|t_{i-(n-1)}, \ldots, t_{i-1})$$

The words bigram and trigram language model denote n-gram language models with $n = 2$ and $n = 3$, respectively. For a term in a bigram, we can write

$$P_{\text{bigram}}(t_i|t_{i-1}) = \frac{P(t_{i-1}, t_i)}{P(t_{i-1})}$$

where $P(t_i|t_{i-1})$ is the probability of the fact that t_i will appear given that the condition t_{i-1} has appeared.

Because the bigram method generates a considerable amount of 0 values (a considerable amount of term combinations that are not reasonable and logical), it is hard to compute a new term combination that is reasonable and logical. Therefore, we define a new probability to represent $h_\Theta(t_i) = \alpha P_{\text{bigram}}(t_i|t_{i-1}) + bP_{\text{unigram}}(t_1)$, where $a \in R_{\geq 0}$, $b \in R_{\geq 0}$ are coefficients with condition $a + b = 1$ ($R_{\geq 0} = \{x \in R : x \geq 0\}$. It is easy to prove that $0 = h_\Theta(x) \leq 1$.

Therefore, we can define the target function of the document as

$$\text{cost}_y(x) = -(y\log h_\Theta(x) + (1-y)\log(1 - h_\Theta(x))), \quad (1)$$

where y only have two possible values 0 or 1. $y = 1$ means that the term is not in the category; $y = 0$ means otherwise.

The cost function is below:

$$\min_\Theta \frac{1}{N_t}\sum_{i=1}^{N_t} y^{(i)}\text{cost}_1\left(\Theta^T x^{(i)}\right) + \left(1 - y^{(i)}\right)\text{cost}_0\left(\Theta^T x^{(i)}\right) + \frac{1}{2N_t}\sum_{j=1}^{N_c}\Theta^2, \quad (2)$$

where N_t is the total number of terms.

In text mining, we normally want to validate our confidence in the indicators. Cross-validation is often used as a method to show confidence. Cross-validation avoids the use of the entire sample data as a training data set, instead only using part of the entire data set for training purposes. When training is finished, a data mining/machine learning algorithm predicts the base on any unseen data that are not part of the training data set. More specifically, the basic steps of k-fold cross-validation are as follows: (1) partitioning all of the samples into k groups in which $k-1$ is used as the training sets and 1 is used as the test set; (2) repeatedly swapping the testing set with a training set that has not been used. The goal is to examine how well the training algorithm learns the sample set.

We first validated our implementation of the N-Gram model using 20 news groups [1] that had well-accepted sample data.

There were 4747 examples in total. We divided the sample corpus into a number of evenly sized groups called folds. For example, with 10 folds, there were 4273 training cases and 474 test cases each. We also assumed that the results would be calculated with 95% normal approximation to the binomial confidence interval per run and with no smoothing on the binomial estimate. The result is shown in Fig. 3. The accuracy of our implementation of the N-Gram is approximately or above 94%. To avoid the fact that most or all of the training data in a category had been seen by the training procedure, we reshuffled the corpus using a random number along with a fixed random seed. Therefore, the experiment is repeatable.

We showed the classic metrics of text mining. We performed classification of the sentiment samples that were obtained from the Internet. In our sample set, we had 500 positive samples and 1000 negative samples. Fig. 4 describes the results of the positive comments versus all comments. We noticed that the *accuracy* of the N-Gram is approximately 0.8242. That means that most of comments were correctly classified by the N-Gram. The *precision* value showed that the N-Gram correctly recognized 79% of the comments out of all of the comments marked by the models. The *recall* values of the two models represented the percentage of positive comments that had been marked positive in all the processed comments. The *recall* value of the N-Gram showed that only approximately 22% of the positive messages were marked as negative messages. The values of *accuracy* showed that the N-Gram correctly marked positive message and negative messages at a success rate of 82.42%.

4. Developing an innovative social media analytics tool

Based on the aforementioned framework and methods, new social media competitive analytics tools can be developed to enhance marketing intelligence. One of the authors has been working on social media competitive analytics and over the past few years has developed an innovative social media competitive analytics tool named VOZIQ. Several programmers have been involved in the development of the tool. In particular, we use Apache Solr for text searches, Hadoop for big data analysis, MySQL for storing processed data and JavaServer Pages (JSP) for web-based visualization. This tool is currently available for online access at http://www.voziq.com. The development team has recently moved the tool to the Amazon AWS cloud platform to offer cloud-based and subscription-based business services and for on-demand scaling to process large volumes of social media data. VOZIQ is currently being used by multiple businesses as a Software as a Service (SaaS) solution. VOZIQ focuses on competitive analysis, using data from social media websites such as Twitter, Facebook, blogs, forums, news, etc., by first creating a social media sentiment benchmark, as outlined in Fig. 4.

Fold	Accuracy	Variance
0	0.97	+/-0.02
1	0.97	+/-0.01
2	0.97	+/-0.02
3	0.98	+/-0.01
4	0.97	+/-0.01
5	0.96	+/-0.02
6	0.96	+/-0.02
7	0.97	+/-0.02
8	0.97	+/-0.02
9	0.98	+/-0.01

Fig. 3. 10-fold cross-validation for news group.

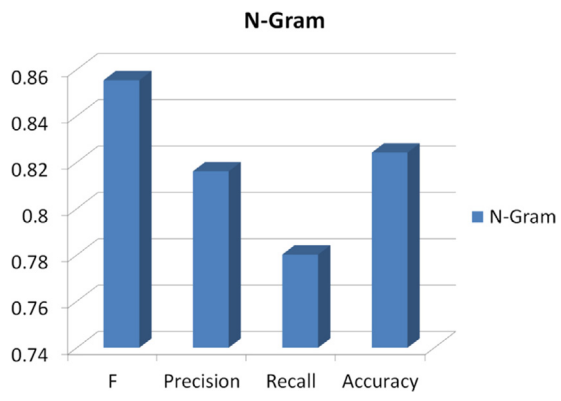


Fig. 4. Classic metrics of text mining.

To illustrate the value of our tool, as an example, we present the use of VOZIQ for analyzing the Twitter messages associated with five large companies in the retail industry (Costco, Walmart, Kmart, Kohl's, and The Home Depot) and generating business insight reports described below.

As a popular social media platform, Twitter enables its users to share and discover topics of interest with a network of "followers" in real time. Twitter allows users to send and read 140-character short messages known as "tweets." There are currently over 500 million active registered Twitter and 1.2 billion Facebook users. Blogs, forums, and news combined generate hundreds of millions of messages that can be mined in a systematic way to extract business intelligence. Many people use these platforms to talk about their daily activities and interests (e.g., favourite brands, customer service experiences, product issues) and to seek or share information. Using the search API provided by these platforms, we can easily gather mentions for selected companies in the same industry.

VOZIQ provides a number of distinctive features:

- *The VOC3 sentiment benchmark report*: To develop this report for any business, VOZIQ gathers social media mentions from that business's voice of customers, competitors, and competitors' customers (VOC3 Listening). It further classifies these social media mentions into various categories using both the top-down and bottom-up categorization approaches discussed earlier to identify relevant categories for each industry. Finally, by discovering sentiment within these categories, it helps businesses discover actionable competitive intelligence. For example, Fig. 5 shows how Costco compares in sentiment across various product, service, and marketing categories. Peer ranks and ranges are displayed to show how one company stacks up against other companies in the industry. Drilldown information at the category level makes this information highly actionable because each company can clearly understand where they are leading and where they need to further focus to obtain a leadership position in its particular industry.
- *Share of voice*: VOZIQ compares social media comments from different companies within the same industry. For example, we compared social media comments that mentioned one or more of the following retail brands: Costco, Kmart, Kohl's, Home Depot, and Walmart, as shown in Fig. 6. By comparing the number of name mentions in Fig. 6, we can see that Walmart has been mentioned the most during the entire period of study. A comparison of social media mentions provides us with a clear and visual understanding of the market attention for different companies within the same industry.

We also collected the revenue data for the five retail brands in 2013 and 2014 from Bloomberg and CSIMARKET. Fig. 6 shows the share of revenue for 2014 for the five retail brands. By comparing

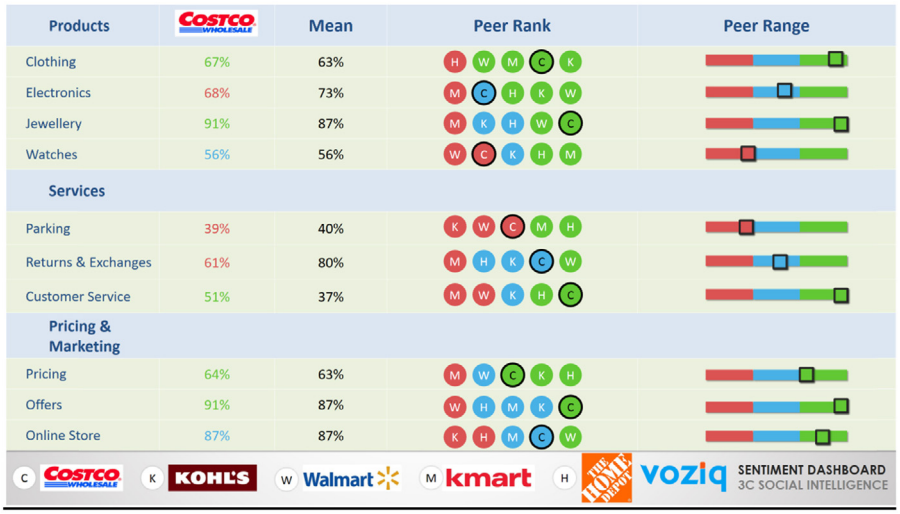


Fig. 5. VOC3 social media sentiment benchmark report for Costco.

the volume of mentions and the share of revenue data over time (such as a week, a month, or a year), it is possible to detect hidden patterns and/or the impact of social media mentions on a business' performance. For example, Figs. 6 and 7 show that Walmart had not only the largest revenue but also the most mentions on social media, compared with other brands, during our survey. By comparing their share of voice on social media with other business metrics across industry peers, interested businesses can identify both potential opportunities and problem areas.

- *Media distribution*: VOZIQ compares the share of different social media platforms in mentions of companies in the same industry. Using this report, one can compare the contribution of various social media platforms to the total number of name mentions for a particular brand in an industry. This can be effectively used to compare the performance of the brand across various social media sites. For example, Fig. 8 shows the contribution of social media sites to the total name mentions of the brand Costco. It can be noted that Costco receives more mentions on forums than on other social media sites.
- *Themes and theme sentiment*: VOZIQ compares the key themes within each brand and offers a breakdown by sentiment. Themes are noun phrases extracted from text and are the primary means of identifying the main ideas within the content. They allow users to discover emerging topics because they primarily use a bottom-up approach to discover interesting content within the mentions. For example, Fig. 9 lists some positive themes and some negative themes that were identified from the social media data set related to Costco during the measurement window. For example, hot dogs and cash cards generate a considerable amount of positive sentiment for Costco, while free samples generate both positive and negative mentions, perhaps based on their availability, non-availability, or taste. Such themes provide companies good insights that allow them to strengthen their services and to address issues.
- *Category clusters*: Category clusters are customized and optimized categories based on a hybrid approach of using top-down and bottom-up methods. The challenging part of the customized and optimized categories is to find the best number of categories and to find the representative topics of the categories. As one example, we present the topic maps of comments related to Costco. The cluster analysis forms a list of clusters. For each cluster, we can find a topic that can be treated as a category. The topic is determined by the highest frequency of keywords. We use larger circles to represent the higher frequency of certain keywords, as shown in Fig. 10. Thus, it can be concluded from Fig. 10 that *Food Items* was the most discussed topic for Costco. Similarly, Figs. 11 and 12 depict the most mentioned topics in

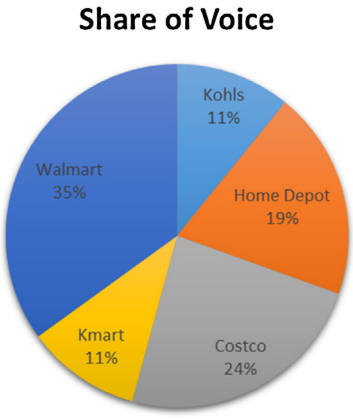


Fig. 6. Comparison of the volume of mentions for different companies.

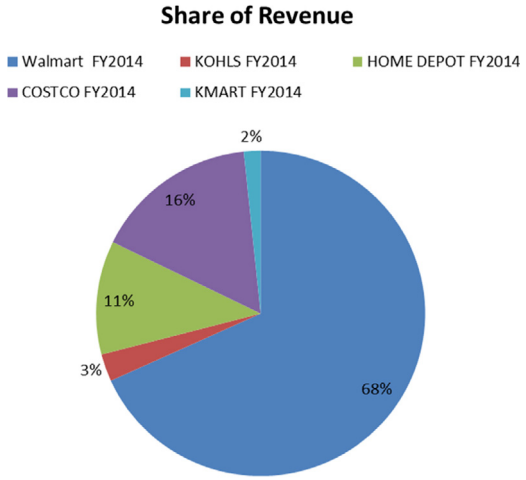


Fig. 7. Share of revenue of five brands in 2014.

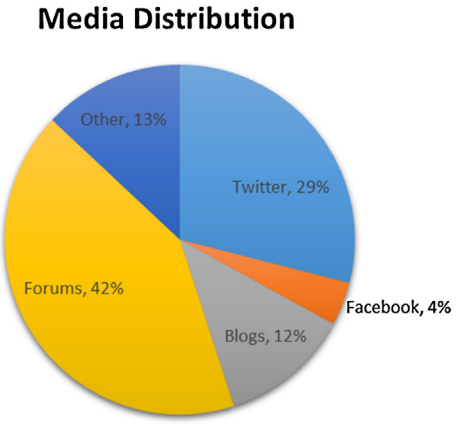


Fig. 8. Media distribution for Costco.

positive and negative discussions, respectively. Such visual representations present powerful tools to pinpoint the strengths and weaknesses of each brand. Another useful feature of this map is the use of different colors for various categories. Categories of a similar type are assigned the same color. This feature can be used to clearly visualize the frequency of mentions in categories of a similar type. For example, in Fig. 11, green circles depict categories related to *Pricing and Promotion*. When these are compared, it can be observed that *Offers* obtain the highest number of mentions, compared with other price and promotion-related categories.

- *Category correlations*: These are the correlations of different categories can help us identify pairs of brand/product-related topics that are associated with each other by customers on social media. We compute the correlation values for each pair of categories. We place thicker arc lines between categories if their correlation values were larger, as shown in Fig. 9. In Figs. 11 and 12, we present the correlations between categories for positive and negative comments, respectively. These topic maps help companies to find related categories and potential drivers for issues.

5. Implications

The paper proposes an innovative approach for efficiently collecting, monitoring, and comparing social media data on various

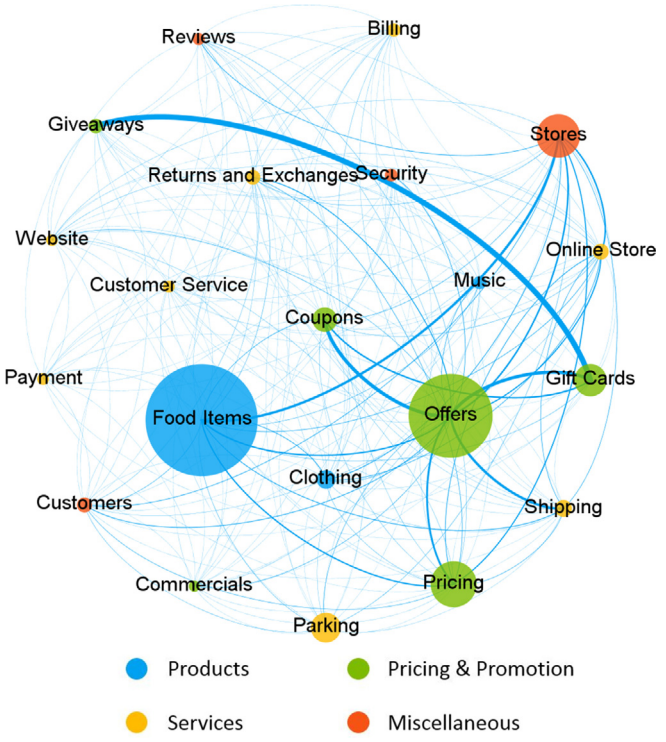


Fig. 10. A generated topic map of Costco social media mentions.

brands or companies in any industry. It uses social media sentiment benchmarks by analyzing user-generated feedback from social media for comparable businesses. The generated results can be organized into business insight reports to help companies to improve their social media marketing effectiveness, to generate sales leads, to gain competitive advantage, and to maximize their return on investment (ROI). A tool is developed based on the proposed approach, and it is in the stage of commercialization. To date, this tool has evolved into a web-based social media competitive intelligence analytics tool that offers a cloud-based and subscription-based business service. Both large businesses and small businesses can subscribe to this social media competitive intelligence analytics service online and can receive customized business competitive insight reports automatically, according to their specified schedule and service levels. The

Fig. 9. An example of positive and negative themes found for Costco.

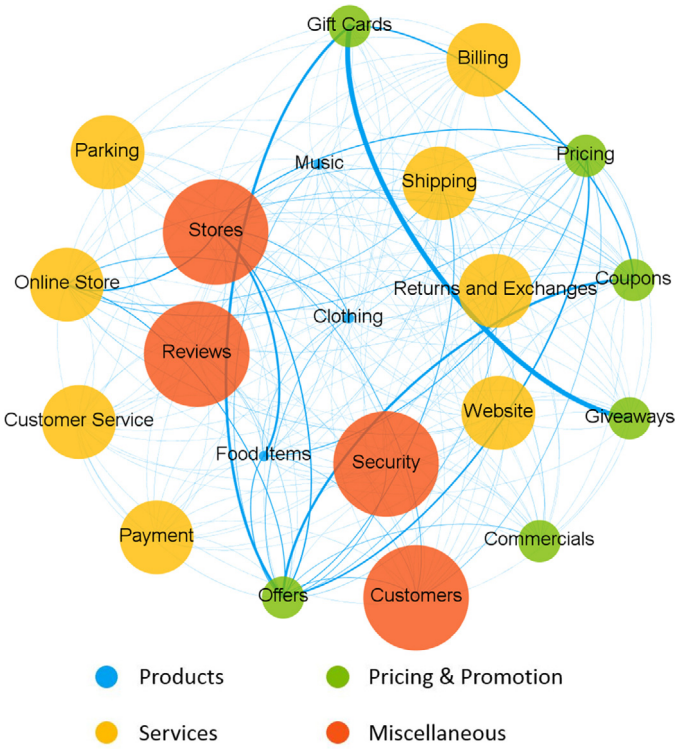


Fig. 11. Positive topic map of Costco social media mentions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

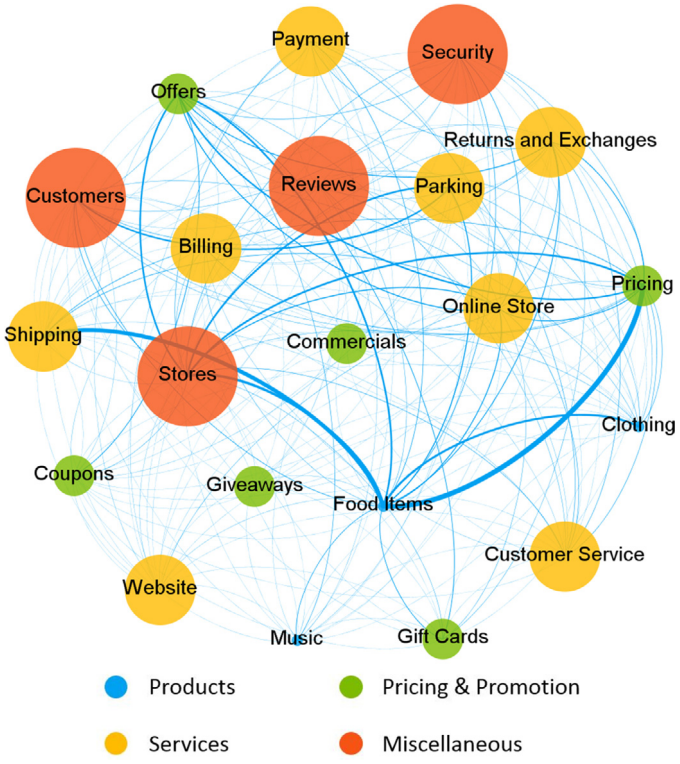


Fig. 12. Negative topic map of Costco social media mentions.

insight reports can be customized by customer preference and can offer different levels of detail, depending on customers' business needs and budget. As an increasing number of companies adopt social media platforms, there is a strong need for them to monitor and understand what their competitors are doing on social media on a daily, weekly, or monthly basis. Existing commercial social media analytics tools are not designed to process or to compare massive social data and visualize results in industry-specific contexts, allowing them to develop industry-specific benchmarks to make social intelligence actionable. Conducting comparative intelligence analysis of large amounts of social data is a very difficult task for companies. In particular, marketing professionals typically do not have the skills to perform this type of analytics. Because today's business environment is so fierce, we believe that developing business-oriented social media competitive intelligence analytics tools using our hybrid categorization models will have broad economic prospects for both large and small businesses.

Furthermore, industry-specific competitive intelligence analytics using big data processing is an innovative technology and has the potential to generate huge business revenues for economic development. For example, large companies typically have millions of social media mentions per month, requiring industry-specific insights and benchmarks to compare with peer groups and to derive actionable insights. On the other hand, many small businesses, such as pizzerias or restaurants, use Facebook and Twitter now. However, small business owners do not have the time or skills to monitor what their competitors are doing, whether they are doing new promotions or posting recipes on social media. Our approach can help these businesses monitor their competitors' social media sites and gain timely business insights and perspective to achieve a competitive advantage. As an increasing number of businesses are adopting social media, monitoring, analyzing, and comparing social data on corporate social media platforms, it is becoming a crucial advantage for businesses to understand the effectiveness of their marketing efforts, support their decision making, improve the efficiency of their product lines, and enhance their customer service, which will lead to the creation of new economic growth opportunities.

6. Conclusion

Increasingly, companies are using social media to promote products and services and to communicate with customers [51]. More and more customers are also using social media to interact with businesses and to share their experiences and opinions, all at a speed previously unheard of. As a result, a considerable amount of user-generated data are created every day. To obtain valuable insights from these data and to acquire a competitive advantage in the market, companies need to develop strong social media competitive analytics skills to differentiate themselves from their competitors. Studies indicate that organizations that focus on analytics significantly outperform their peers on the key business metrics of growth, earnings, and performance [52]. It is believed that social media competitive analytics can help organizations to realize the strengths and weaknesses of their products and services, to enhance business effectiveness, and to improve customer satisfaction [13,19].

As a main contribution, this study presents a novel social media competitive analytics framework that uses sentiment benchmarks and relevant methods for industry-specific marketing intelligence. In particular, the framework proposes to identify top companies in a particular industry and then create sentiment benchmarks that use text mining, sentiment analysis, and traditional social network analysis approaches to process large amounts of user-generated data on their social media sites for competitive analysis and comparison. This framework, with its focus on sentiment benchmarks, is novel and practical, and it contributes to the social media literature, given that we have not found a published academic article with a similar framework despite our extensive literature review. Furthermore, it is noted that, although many existing text

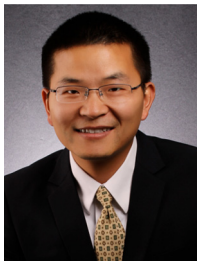
mining, sentiment analysis, and social network analysis tools are available, many end users lack the technical knowledge to use these tools. Thus, a user-friendly social media competitive analytics tool is urgently needed. To fill this need, one of the authors and his staff have developed an innovative social media competitive analytics tool named VOZIQ. An example using VOZIQ for the analysis of the social media sites of five large retail companies and the generation of business insight reports is presented. The results indicate that our proposed framework is feasible and may have wide applicability in business. Firms can use our proposed framework and tools such as VOZIQ to guide their efforts to monitor, compare, and analyze user-generated social media content for companies in their specific industries. To date, we have used our proposed approach and tool to collect, monitor, and analyze social media messages in the PC manufacturing industry, the telecommunication industry, the financial service industry, and the retail industry. Feedback from users regarding the utility of the tool in practice needs to be collected to further improve this tool. For future research, we plan to examine the relationship between social media mentions/sentiments and business performance.

References

[1] 20 Newsgroups, The 20 Newsgroups Data Set, 2008 Retrieved from ⟨http://qwone.com/jason/20Newsgroups, at January 2008⟩.
[2] M. Abdous, W. He, Using text mining to uncover students' technology-related problems in live video streaming, Br. J. Educ. Technol. 40 (5), 2011, pp. 40–49.
[3] A.S. Abrahams, W. Fan, J. Jiao, G.A. Wang, Z. Zhang, An integrated text analytic framework for product defect discovery, Prod. Oper. Manage. 2015http://dx.doi.org/10.1111/poms.12303.
[4] S. Abrahams, J. Jiao, W. Fan, G.A. Wang, Z. Zhang, What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings Decis. Support Syst. 55 (4), 2013, pp. 871–882.
[5] S. Abrahams, J. Jiao, G.A. Wang, W. Fan, Vehicle defect discovery from social media, Decis. Support Syst. 54 (1), 2012, pp. 87–97.
[6] G. Barbier, H. Liu, Data mining in social media, Social Network Data Analytics, Springer US, 2011 pp. 327–352.
[7] R. Bellman, R.E. Bellman, R.E. Bellman, R.E. Bellman, Adaptive Control Processes: A Guided Tour, (vol. 4), Princeton University Press, Princeton, NJ, 1961.
[8] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1), 2011, pp. 1–8.
[9] R. Bose, Competitive intelligence process and tools for intelligence analysis, Ind. Manage. Data Syst. 108 (4), 2008, pp. 510–528.
[10] W.B. Cavnar, J.M. Trenkle, *N*-gram-based text categorization, Ann Arbor, MI 48113 (2), 1994, pp. 161–175.
[11] M. Chau, J. J. Xu, Business intelligence in blogs: understanding consumer interactions and communities, MIS Q. 36 (4), 2012, pp. 1189–1216.
[12] H. Chen, M. Chau, D. Zeng, CI Spider: a tool for competitive intelligence on the Web, Decis. Support Syst. 34 (1), 2002, pp. 1–17.
[13] H. Chen, R.H. Chiang, V.C. Storey, Business intelligence and analytics: from big data to big impact, MIS Q. 36 (4), 2012, pp. 1165–1188.
[14] L. Dey, S.M. Haque, A. Khurdiya, G. Shroff, Acquiring competitive intelligence from social media, in: Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, ACM, 2011, p. 3.
[15] W. Duan, Q. Cao, Y. Yu, S. Levy, Mining online user-generated content: using sentiment analysis technique to study hotel service quality, in: Proceedings of the 46th Hawaii International Conference on System Sciences, 2013, pp. 3119–3128.
[16] W. Fan, M.D. Gordon, The power of social media analytics, Commun. ACM 57 (6), 2014, pp. 74–81.
[17] M.U. Fayyad, G. Piatesky-Shapiro, P. Smuth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, Palo Alto, California, 1996.
[18] G. Governatori, R. Iannella, A modeling and reasoning framework for social networks policies, Enterp. Inf. Syst. 5 (1), 2011, pp. 145–167.
[19] W. He, S.H. Zha, L. Li, Social media competitive analysis and text mining: a case study in the pizza industry, Int. J. Inf. Manage. 33 (3), 2013, pp. 464–472.
[20] W. He, Examining students' online interaction in a live video streaming environment using data mining and text mining, Comput. Hum. Behav. 29 (1), 2013, pp. 90–102.
[21] W. He, Improving user experience with case-based reasoning systems using text mining and Web 2.0, Expert Syst. Appl. 40 (2), 2013, pp. 500–507.
[22] J. Hung, Trends of E-Learning Research from 2000 to 2008: use of Text Mining and Bibliometrics, Br. J. Educ. Technol. 43 (1), 2012, pp. 5–16.
[23] J. Hung, K. Zhang, Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching, MERLOT J. Online Learn. Teach. 4 (4), 2008, pp. 426–437.
[24] L. Kahaner, Competitive Intelligence: How to Gather, Analyze and Use Information to Move Your Business to the Top, Touchstone, New York, NY, 1998.
[25] Y. Kim, Weighted order-dependent clustering and visualization of web navigation patterns, Decis. Support Syst. 43 (4), 2007, pp. 1630–1645.
[26] A. Klein, O. Altuntas, M. Riekert, V. Dinev, A combined approach for extracting financial instrument-specific investor sentiment from weblogs, Wirtschaftsinformatik Proceedings, 2013, p. 44.
[27] A. Kloptchenko, T. Eklund, J. Karlsson, B. Back, H. Vanharanta, A. Visa, Combining data and text mining techniques for analyzing financial reports, Intell. Syst. Acc. Finance Manage. 12 (1), 2004, pp. 29–41.
[28] H. Lee, K. Choi, D. Yoo, Y. Suh, G. He, S. S. Lee, The more the worse? Mining valuable ideas with sentiment analysis for idea recommendation in: Proceedings of PACIS, 2013, p. 2013.
[29] J. Leskovec, Social media analytics: tracking, modeling and predicting the flow of information through networks, in: Proceedings of the 20th International Conference Companion on World Wide Web, ACM, 2011.
[30] N. Li, D.D. Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast, Decis. Support Syst. 48 (2), 2010, pp. 354–368.
[31] H. Lin, W. Fan, P. Chau, Determinants of users' continuance of social networking sites: a self-regulation perspective, Inf. Manage. 51 (5), 2014, pp. 595–603.
[32] B. Liu, Sentiment analysis and subjectivity, in: N. Indurkhya, F.J. Damerau (Eds.), Handbook of Natural Language Processing, Taylor and Francis Group, Boca, 2010
[33] B. Liu, S.G. Cao, W. He, Distributed data mining for e-business, Inf. Technol. Manage. 12 (1), 2011, pp. 1–13.
[34] S. Ludwig, K. de Ruyter, M. Friedman, E.C. Brüggen, M. Wetzels, G. Pfann, More than words: the influence of affective content and linguistic style matches in online reviews on conversion rates, J. Marketing 77 (1), 2013, pp. 87–103.
[35] G. Maskeri, S. Sarkar, K. Heafield, Mining business topics in source code using latent dirichlet allocation, in: Proceedings of the First India Software Engineering Conference, 2008, pp. 113–120.
[36] S. Morinaga, K. Yamanishi, K. Tateishi, T. Fukushima, Mining product reputations on the web, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 341–349.
[37] B. Pang, L. Lee, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 271–278.
[38] B. Pang, l. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 10, Association for Computational Linguistics, 2002.
[39] C. Romero, S. Ventura, E. Garcia, Data mining in course management systems: moodle case study and tutorial, Comput. Educ. 51 (1), 2008, pp. 368–384.
[40] P. Ross, C. McGowan, L. Styger, A Comparison of Theory and Practice in Market Intelligence Gathering for Australian Micro-businesses and SMEs, Social Science Electronic Publishing, Inc, 2012 (available at SSRN 2174337).
[41] T.O. Sprenger, I.M. Welpe, Tweets and trades: The information content of stock microblogs, November 1, 2010, (available at SSRN: http://ssrn.com/abstract=1702854) or http://dx.doi.org/10.2139/ssrn.1702854.
[42] S. Stieglitz, L. Dang-Xuan, Social media and political communication: a social media analytics framework, Soc. Netw. Anal. Min. 3 (4), 2013, pp. 1277–1291.
[43] T.S. Teo, W.Y. Choo, Assessing the impact of using the Internet for competitive intelligence, Inf. Manage. 39 (1), 2001, pp. 67–83.
[44] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the socialweb, J. Am. Soc. Inf. Sci. Technol. 63 (1), 2012, pp. 163–173.
[45] K.J. Trainor, T.K. Michael, A. Raj, Effects of relational proclivity and marketing intelligence on new product development, Marketing Intell. Plann. 31 (7), 2013, pp. 788–806.
[46] G.A. Wang, J. Jiao, A.S. Abrahams, W. Fan, Z. Zhang, ExpertRank: a topic-aware expert finding algorithm for online knowledge communities, Decis. Support Syst. 54 (3), 2013, pp. 1442–1451.
[47] K. Xu, S.S. Liao, J. Li, Y. Song, Mining comparative opinions from customer reviews for competitive intelligence, Decis. Support Syst. 50 (4), 2011, pp. 743–754.
[48] Zeng, H. Chen, R. Lusch, S.H. Li, Social media analytics and intelligence, IEEE Intell. Syst. 25 (6), 2010, pp. 13–16.
[49] Z. Zheng, X. Wu, R. Srihari, Feature selection for text categorization on imbalanced data, ACM SIGKDD Explorations Newsl. 6 (1), 2004, pp. 80–89.
[50] N. Zhong, Y. Li, S. Wu, Effective pattern discovery for text mining, IEEE Trans. Knowl. Data Eng. 24 (1), 2012, pp. 30–44.
[51] M. Zhou, L. Lei, J. Wang, A.G.A. Wang, W. Fan, Social media adoption and corporate disclosure, J. Inf. Syst. 2014http://dx.doi.org/10.2308/isys-50961.
[52] P. Zikopoulos, K. Parasuraman, T. Deutsch, J. Giles, D. Corrigan, Harness the Power of Big Data The IBM Big Data Platform, McGraw Hill Professional, New York, NY, 2012.

Wu He is currently an Assistant Professor of Information Technology at Old Dominion University, USA. He earned his PhD in Information Science from the University of Missouri. His research interests include data mining, social media analytics, cyber security, knowledge management, human information behavior, case-based reasoning, and computing education.

Harris Wu is an Associate Professor of Information Technology at the Strome College of Business Administration, the Old Dominion University. His current research interests include social computing, cloud computing, enterprise knowledge management and social media mining.



Dr. Gongjun Yan received his Ph.D. in Computer Science from Old Dominion University. He is currently an Assistant Professor in University of Southern Indiana. His main research areas include algorithms, security, privacy, routing, and healthcare in Internet, Vehicular Ad-Hoc Networks, Sensor Networks and Wireless Communication. In years, Dr. Yan applies mathematical analysis to model behavior of complex systems and integrates existing techniques to provide comprehensive solutions.

Vasudeva Akula is the co-founder and principal business consultant at VOZIQ where he creates and manages strategy for clients in establishing voice of customer based business performance improvement programs. In his 20 year career spanning banking & financial services, insurance, retail, technology, telecom & wireless, utility industries, he played key roles in helping Fortune 500 companies successfully deploy and realize significant ROI out of customer intelligence solutions.

Jiancheng Shen is currently a Ph.D. Candidate in Finance at Old Dominion University, USA. His research interests include neuro-finance, media and financial markets, high frequency financial data analysis and international economics.