# Understanding Online Consumer Review Opinions with Sentiment Analysis using Machine Learning

**Christopher C. Yang**
College of Information Science and Technology
Drexel University, PA, USA

**Xuning Tang**
College of Information Science and Technology
Drexel University, PA, USA

**Y. C. Wong**
Digital Library Laboratory
Chinese University of Hong Kong, Hong Kong

**Chih-Ping Wei**
Department of Information Management
National Taiwan University, Taiwan

## Abstract

*With the advent of Web 2.0 technologies, the Web has evolved to become a popular channel of communication and interaction between Web users and online consumers. Social media, unlike traditional media, have rich but unorganized content contributed by users, often in fragmented and sparse fashion. Users usually spend a lot of their time filtering useless information and yet are not able to capture the essence. In this study, we focus on user-contributed reviews of products, which many online consumers use to support their purchase decisions by identifying products that best fit their preferences. In the recent years, sentiment classification and analysis of online consumer reviews has drawn significant research attention. Most existing techniques rely on natural language processing tools to parse and analyze sentences in a review, yet they offer poor accuracy, because the writing in online reviews tends to be less formal than writing in news or journal articles. Many opinion sentences contain grammatical errors and unknown terms that do not exist in dictionaries. Therefore, this study proposes two supervised learning techniques (class association rules and naïve Bayes classifier) to classify opinion sentences into appropriate product feature classes and produce a summary of consumer reviews. An empirical evaluation that compares the performance of the class association rules technique and the naïve Bayes classifier for sentiment analysis shows that our proposed techniques achieve more than 70% of the macro and micro F-measures.*

**Keywords:** Opinions mining, Web mining, Electronic commerce, Machine learning, Sentiment classification, Sentiment analysis, Text mining, Social media analytics

## Introduction

The massive volume of user-contributed content on the Web gets updated so frequently that it is almost impossible for search engines to index it and offer real-time searching capability. Yet users continue to hunger for up-to-date information to support their tasks. In particular, the Web is an excellent source for gathering consumer review opinions (Blei and McAuliffle, 2008; Dellarocas, 2003; Ding et al., 2008; Liu and Yu, 2008; Forman et al., 2008; Godes and Mayzlin, 2004). On modern business-to-consumer (B2C) electronic commerce platforms, consumers not only browse online catalogs and make purchases but also search for opinions posted by other consumers to support their purchase decision making.

Currently, many Web portals in addition to B2C websites (e.g., amazon.com) provide online consumer review systems that enable users to submit and retrieve consumer reviews. For example, epinion.com, Reteitall.com, and cnet.com are some popular online consumer review websites. They provide a combination of formats for users to submit review opinions, including open text boxes, lists of pros and cons, and ratings on an n-point scale. If a user clicks on a particular product, he or she sees a list of consumer reviews contributed by others. However, it becomes tedious and time consuming to browse through a long list of consumer reviews, especially for comparisons of several products. A review summary for each product that lists the pros and cons of each feature thus is desirable; a general rating is not as sufficient, because potential consumers cannot identify which products that best match their concerned or preferred product features. For example, some potential digital camera consumers might worry about the price and image quality, whereas others are more interested in the battery life and lenses. To address such challenges, extant research has investigated two main problems, sentiment classification and sentiment analysis (Liu et al., 2005). Sentiment classification determines the sentiment orientation, whether positive or negative, of an opinion text. On the other hand, sentiment analysis extracts the product features that an opinion text describes. Because most online consumer review websites provide separate input sections for pros, cons, and ratings, the sentiment orientation is explicit, but the described product features remain hidden in the text. Accordingly, sentiment analysis is the focus of this paper.

Recent research on sentiment analysis relies on natural language processing and linguistic techniques (Scaffidi et al., 2007; Hu and Liu, 2004a; Liu et al., 2005; Popescu, and Etzioni, 2005; Zhang and Varadarajan, 2006), which perform well when the text can be parsed accurately by natural language processing tools. However, text in Web review opinions generally is less rigorous than the wording in formal documents, such as business reports, news documents, or journal articles. The text in Web review opinions often does not conform to linguistic and grammatical rules, and it even might not include complete sentences. In addition, many new terms appear in Web review opinions, including technical terms and name entities such as mp3, 3G, and iPod. In response, we investigate in this study a machine learning approach for classifying the product features described in Web review opinions, even if the language is informal.

The organization of the remainder paper is as follows: In Section 2, we review prior related studies. In Section 3, we present the machine learning approach for classifying consumer review opinions, including class association rules and the naïve Bayes classifier. We next describe the design of our empirical evaluations and discuss important experimental results in Section 4. We conclude in Section 5 with a summary and some potential further research ideas.

## Related Work

As Web 2.0 techniques become increasingly popular, the number of online consumer reviews, forums, and blogs are expanding rapidly. However, it is difficult for users to digest all this information unless an automatic summary is available. Social media summarization takes fragmented messages as inputs and produces an overview of the user-generated content. In particular, sentiment classification and analysis attempt to summarize online consumer review opinions and thereby enable users to compare consumer reviews across multiple products by highlighting the pros and cons of their product features. When Web users submit opinions about a particular product, they write positive and negative comments about varied product features. To provide a detailed summary and comparison across multiple products, we need to compare the ratio of positive and negative comments about each product feature across multiple products. Therefore, potential consumers can easily identify the product that receives positive comments in relation to the product features in which they are most interested. Sentiment classification determines the orientation of a review opinion, regardless of the product features, whereas sentiment analysis extracts review opinions into specific product feature classes (Liu et al., 2005).

### Social Media Summarization

Social media summarization integrates Web opinions through the function of topic modeling to generate a high quality summary. Its critical component is the extraction of aspects of an object that users rate frequently in a set of online reviews (Blei and McAuliffe, 2008; Mei et al., 2007; Titov and McDonald, 2008; Zhai et al., 2004). For example, Titove and McDonald (2008) employ multi-grain latent Dirichlet allocation (LDA) to extract ratable aspects by clustering important terms according to a local topic. Wang et al., (2008) propose a novel model to create a compressed summary, while retaining the main characteristics of the original set of documents. They first calculate sentence-to-sentence similarities using semantic analysis to construct a similarity matrix, then conduct symmetric matrix factorization to group sentences into clusters. Finally, they select the most informative sentences to represent the set of documents. Although this model has performed well with well-structured data, it does not necessarily work effectively with Web 2.0 content, which tends to be less organized. Lu and Zhai (2008) propose a summarization technique based on semi-supervised LDA and probabilistic latent semantic analysis (PLSA) models. They employ a well-written expert review as a template for incorporating other opinions scattered across various sources. Accordingly, this technique generates an opinion summary that consists of expert reviews and supplemental opinions. However, the quality of summary depends on the selected expert review.

Social media summarization generates summaries by topic modeling but does not work at the sentence level to identify specific product features on which a user comments. That is, topics in social media summarization are not necessarily the product features in online consumer reviews. Nor does this technique aim to classify the orientation of the product features that online consumers review. The main objective instead is to extract representative and informative sentences that summarize a vast collection of social media content. As mentioned, we focus on online consumer reviews in this study. The two related research problems are sentiment classification and sentiment analysis.

### Sentiment Classification

Sentiment classification is a document classification task, which involves two classes: positive and negative. Sentiment classification therefore considers sentiment rather than topics in traditional document classification. Several supervised learning approaches for sentiment classification have been investigated. Weibe et al., (1999) and Hatzivassiloglou and Weibe (2000) identify nouns and adjectives, which are indicative of positive or negative opinions; Turney (2002) uses mutual information between term phrases and positive and negative words such as “excellent” and “poor” to identify opinions. Wei et al., (2006) employ two comprehensive lists of positive and negative words from the General Inquirer (available at http://www.wjh.harvar.edu/~inquirer/) to facilitate sentiment classification tasks. To solve the context-dependent and conflicting opinion word problems, Ding et al., (2008) require extensive manual effort and highly depend on natural language processing rules. In general, these sentiment classification techniques rely on sets of positive and negative lexicons to identify other indicative terms that might appear in review opinions. However, positive and negative indicators vary for different kinds of products. The sets of positive and negative indicators are not interchangeable but tailored for specific products. Recent online consumer review systems also provide user interfaces that separate positive and negative inputs, rather than using only one input field in a free text format. Thus, sentiment classification becomes less challenging in practice, because users submit their positive and negative opinions separately.

### *Sentiment Analysis*

Sentiment analysis is more complicated than sentiment classification: It classifies review opinions into several product feature classes, which vary in number across different types of products. Several research works pertaining to sentiment analysis use supervised or unsupervised learning. For example, Hu and Liu (2004a; 2004b) and Liu et al. (2005) use the NLProcessor linguistic processor (see http://www.infogistics.com) to parse review opinions and extract nouns and noun phrases. With the use of an association rule mining algorithm, they then discover all frequent product features in these noun phrases extracted previously. Sentences without any discovered product features and opinion words will not be classified. However, the accuracy of this unsupervised learning method depends on the performance of the parsing by NLProcessor, and extensive manual effort is needed to adjust the tagging result of NLProcessor before the review opinions can effectively be classified. Jindal and Liu (2006a; 2006b) propose another technique that integrates pattern discovery and supervised learning approaches to indentify comparative sentences in text documents. Their technique is useful for sentimental analysis, because it can extract product features and opinion words in comparative sentences. However, it also relies on POS tagging and keywords, and it has trouble dealing with informal Web language existing widely in online reviews and web blogs. Wei et al. (2010) propose a semantic-based approach that exploits lists of positive and negative adjectives defined in General Inquirer to recognize opinion words semantically and thereby extract product features. However, this technique requires the availability of lists of positive and negative lexicons. Popescu and Etzioni (2005) use a set of domain-independent extraction patterns, predefined in a Web information extraction system (KnowItAll) to instantiate specific extraction rules for each product feature class. Kobayashi et al. (2004; 2005) also adopt an information extraction approach to extract product features. However, predefined extraction patterns are required for these information-extraction-based techniques.

In this study, we instead propose the use of two supervised learning techniques, class association rules and the naïve Bayes classifier, to assign product features without using lexicon sets, natural language processing, or predefined extraction patterns. Rather, we use only a training data set.

## Supervised Learning for Classifying Consumer Review Opinions

Supervised learning simulates the way humans learn from their past experiences to acquire knowledge and thus perform practical tasks in decision making and classification. Supervised learning has been widely used for document classification. Specifically, it takes a set of preclassified training documents to develop a classification model, which then can classify any new documents into one or more predefined classes. However, in sentiment analysis for online review opinions, the unit of analysis is a sentence, rather than a document.

A consumer review contributed by a Web user consists of multiple sentences, each of which may refer to one or more product features. For example, the sentence “The digital camera takes good pictures but it is not expensive at all” describes two product features: image quality and price. It is also possible that a sentence does not describe any product feature. For instance, a reviewer may tell a story of using a product while on vacation: "I took the Canon G7 with me in my summer vacation at Yosemite." This sentence only describes the event, nothing about his or her sentiment on any product features. Therefore, the number of product features in an opinion sentence can range from 0 to *n*, where *n* is the number of possible product features of the focal product. In our preliminary study, we have found that more than 50% of consumer review sentences do not describe any product features, yet documents in document classification are always classified into one or more classes of topics. This large amount of noise makes sentiment analysis not trivial. In addition, identifying the product feature(s) from a sentence is more challenging than identifying a document topic, because a document is much longer and contains more term features to support classification. Furthermore, the class of a document can be determined according to multiple representative term features in a document. However, an opinion sentence may contain only one term feature that describes a product feature. Accurately extracting this term feature from a short sentence also is nontrivial.

We model the problem of conducting a sentiment analysis of consumer review opinions as a supervised learning task. For a product *P*, there is a set of consumer reviews $R = \{r_1, r_2, \ldots, r_{|R|}\}$. For each review $r_i$, there is a set of opinion sentences $S_i = \{s_{i1}, s_{i2}, \ldots, s_{i|S_i|}\}$. For each product *P*, there also is a set of product feature classes $F = \{f_1, f_2, \ldots, f_{|F|}\}$. Some term features, $T_{j=} \{t_{j1}, t_{j2}, \ldots\}$, are associated with each $f_j$. For example, "AA" and "lithium" are term features associated with the product feature "battery" for a digital camera, whereas "GB" and "Compact Flash" are term features associated with the product feature "memory." In the following, we detail the two supervised learning techniques that we use for classifying consumer review opinions, including class association rules and the naïve Bayes classifier.

## Class Association Rules

The goal of the class association rule mining is to extract associations between term features in consumer review opinions and product features for a particular product that cooccur frequently. A set of preclassified opinion sentences provides training examples for determining the class association rules. Each opinion sentence can be labeled with one or more product features $f_j$, or no product feature, that is, *none*. The class association rule mining extracts all ruleitems with support equal to or higher than a prespecified *minimum support threshold* and confidences equal to or higher than a prespecified *minimum confidence threshold*. We define a ruleitem as $(t_{jk}, f_j)$, where $t_{jk} \in T_j$ and $f_j \in F$, and we establish the class association rule as:

$$t_{jk} \rightarrow f_j, \text{ where } t_{jk} \in T_j \text{ and } f_j \in F.$$

With a labeled opinion sentence, we first remove stop words that do not bear any semantics. All unigrams and bi-grams then will be extracted from the sentence. In this study, we do not consider *n*-grams that are longer than two terms, because the frequencies of most longer *n*-grams are too low to be included in class association rules. Moreover, in most cases, these *n*-grams are not valid words but combinations of broken words. Using the extracted unigrams and bi-grams from an opinion sentence, we generate a set of candidate ruleitems by associating each unigram or bi-gram with every product feature labeled for the opinion sentence. Subsequently, the set of candidate ruleitems are stored and then accumulated as we process all labeled opinion sentences. After all of the candidate ruleitems have been generated from the set of training opinion sentences, class association rule mining extracts the rules using two parameters, *minimum support threshold* and *minimum confidence threshold*. That is, a class association rule, $t \rightarrow f$, is deduced if:

$$Support(t,f) = \frac{count(t,f)}{N} \geq \text{minimum support threshold}$$

$$Conf(t,f) = \frac{count(t,f)}{count(t)} \geq \text{minimum conference theshold}$$

where *count*($t$, $f$) is the number of opinion sentences with the term feature (unigram or bi-gram) $t$ labeled $f$, *count*($t$) is the number of opinion sentences with the term feature $t$, and $N$ is the total number of opinion sentences for training.

As we discussed, the class association rules we extract are not limited to associations between a single term (i.e., unigram) and a product feature. In some cases, phrases (i.e., bi-grams) appear in opinion sentences to describe product features, such as "flash card" in association with the product feature "Memory." The extracted class association rules from these unigrams and bi-grams may conflict though. Let $t_1$ be a unigram, $t_2$ be a bi-gram, and $t_2$ consist of $t_1$. In this case, $t_1 \to f_a$ conflicts with $t_2 \to f_b$ if $f_a \neq f_b$. For example, "flash" → "Flash" and "flash card" → "Memory" conflict. The term feature "flash" is associated with the product feature "Flash," but the term feature "flash card" is associated with "Memory." When such conflict occurs, the rule extracted from a bi-gram (i.e., phrase) overrides the rule extracted from a single term, rather than comparing the support or confidence levels of the two conflicting class association rules, because a phrase describes a product feature more specifically than a single term.

### Naïve Bayes Classifier

The naïve Bayes classifier is a probabilistic classifier, based on the Bayes theorem. It assumes class-conditional independence, such that the chance a term appears in an opinion sentence is independent of the chances that other terms appear in the same opinion sentence for a given product feature. It treats each opinion sentence as a "bag" of terms, so the probability of a term appearing also is independent of its position in the opinion sentence. Accordingly, the naïve Bayes classifier estimates the posterior probability of each product feature $f_i$, given the target opinion sentence $os$, according to the Bayesian rule:

$$p(f_i \mid os) = \frac{p(f_i)p(os \mid f_i)}{p(os)}$$

where $p(f_i)$ is the probability that the product feature $f_i$ appears, $p(os \mid f_i)$ is the conditional probability that the opinion sentence $os$ occurs given $f_i$, and $p(os)$ is the probability that the opinion sentence $os$ occurs.

Evidently, we can ignore $p(os)$ when estimating $p(f_i \mid os)$, because it is identical for all product features, and the relative values of $p(f_i \mid os)$ can determine product feature(s) in the opinion sentence $os$:

$$p(f_i \mid os) \propto p(f_i)p(os \mid f_i)$$

The prior probability of a product feature $f_i$, $p(f_i)$, is computed as:

$$p(f_i) = \frac{n(f_i)}{n}$$

where $n$ represents the total number of opinion sentences in the training data set and $n(f_i)$ denotes the number of opinion sentences in the training data set that are labeled $f_i$.

Assuming that the target opinion sentence $os$ has a set of terms $T$ and a term $t_k$ in $T$ is independent of all other terms in $T$. Accordingly, we can calculate the conditional probability that the opinion sentence $os$ occurs, given $f_i$, or $p(os \mid f_i)$, as:

$$p(os \mid f_i) = \prod_{t_k \in T} p(t_k \mid f_i)$$

Furthermore, the probability of the occurrence of a term $t_k$, given a product feature $f_i$, or $P(t_k|f_i)$, also can be computed as the number of opinion sentences that include $t_k$, given that the opinion sentences are labeled $f_i$, divided by the number of opinion sentences labeled $f_i$ in the training data set. That is,

$$p(t_k \,|f_i) = \frac{n(t_k, f_i)}{n(f_i)}$$

Accordingly, we can derive the posterior probability of the target opinion sentence $os$ being assigned to a product feature $f_i$, $p(f_i \mid os)$, as follows:

$$p(f_i \mid os) \propto p(f_i)p(os \mid f_i) = \frac{n(f_i)}{n} \prod_{t_k \in T} \frac{n(t_k, f_i)}{n(f_i)} .$$

Using this formulation, we classify the target opinion sentence *os* to the product feature with the highest posterior probability. That is,

$$ArgMax_{f_i \in F}\, p(f_i \mid os)$$

The naïve Bayes classifier is a powerful classification tool; however, its performance depends greatly on the term features $t_k$ selected from *T* for sentiment analysis. Besides, its efficiency declines without such feature selection. In an opinion sentence, there are many irrelevant and redundant terms. For example, users may discuss the camera they have purchased and the pictures they have shot. The terms “purchase” and “shoot” appear frequently in opinion sentences about cameras, but these terms do not contribute to the classification of opinion sentences to product features and thus should be regarded as noises. By filtering these terms, we likely improve classification accuracy and efficiency, as well as the interpretability of the classifier. In this study, we investigate two feature selection metrics, information gain and chi-square ($\chi^2$), to select the representative term features for the naïve Bayes classifier.

## Information Gain

The information gain metric assesses the amount of information obtained by a term *t* for class prediction using the absence and presence of *t* in the training data set (Yang and Pedersen, 1997). A term with a high information gain can reduce the information needed to classify the data set; that is, it reduces the impurity or disorder of the data set. The expected information needed to classify a given data set using a term *t* is called the entropy of *t*. To compute the entropy of a term *t*, we must consider the distribution of all product features for the presence and absence of *t*, $p(f_i \mid t)$ and $p(f_i \mid \bar{t})$, respectively:

$$E(t) = -p(t)\sum_{f_i \in F} p(f_i \mid t)\log_2 p(f_i \mid t) - p(\bar{t})\sum_{f_i \in F} p(f_i \mid \bar{t})\log_2 p(f_i \mid \bar{t})$$

The original entropy, based on the distribution of the training data set across all product features, is:

$$E(F) = -\sum_{f_i \in F} p(f_i)\log_2 p(f_i)$$

The information gain for a term *t* is the expected reduction in entropy by considering the term, or *G*(*t*) = *E*(*F*) – *E*(*t*):

$$G(t) = -\sum_{f_i \in F} p(f_i)\log_2 p(f_i) + p(t)\sum_{f_i \in F} p(f_i \mid t)\log_2 p(f_i \mid t)$$
$$+\, p(\bar{t})\sum_{f_i \in F} p(f_i \mid \bar{t})\log_2 p(f_i \mid \bar{t})$$

If the information gain is low, the presence or absence of the term is not important for determining the product feature class. For this study, we select only those terms with information gains equal to or higher than a predefined threshold for the naïve Bayes classifier. All other terms with information gains lower than the threshold are discarded.

## Chi-Square ($\chi^2$)

The $\chi^2$ metric evaluates the statistically significant difference between proportions for a term and a product feature class. It thus measures whether observations of two variables, expressed in a contingency table, are independent. The $\chi^2$ metric between a term *t* and a product feature $f_i$ is calculated as follows:

$$\chi^2(t, f_i) = \frac{\left(p(t \wedge f_i)p(\bar{t} \wedge \bar{f_i}) - p(t \wedge \bar{f_i})p(\bar{t} \wedge f_i)\right)^2}{p(f_i)p(\bar{f_i})p(t)p(\bar{t})}$$

For each term and each product feature in the training data set, we compute its $\chi^2$ value. To evaluate the goodness of a term, we aggregate cross-class $\chi^2$ values, generally using either the maximum $\chi^2$ or the average $\chi^2$ method, as follows:

$$\chi^2_{\max}(t) = \underset{f_i \in F}{Max}\, \chi^2(t, f_i)$$

$$\chi^2_{avg}(t) = \sum_{f_i \in F} p(f_i)\chi^2(t, f_i)$$

If a term lacks strong classification power, its $\chi^2$ values across the product feature classes will be close. In other words, the differences between the $\chi^2$ values across product features classes are small. In this case, this term would not be useful for the naïve Bayes classifier, because its contributions to different product feature classes are approximately the same, and it does not help identify a correct product feature class. In this study, we therefore filter out terms for which the difference between the maximum $\chi^2$ and the average $\chi^2$ is less than a predefined threshold $\delta$ (i.e., we remove term $t$ if $\chi^2_{\max}(t) - \chi^2_{avg}(t) < \delta$).

## Empirical Evaluation

We have conducted experiments to evaluate empirically the performance of our proposed supervised learning techniques (i.e., class association rules and naïve Bayes classifier) for sentiment analysis. In the first experiment, we evaluated the impact of the minimum support and minimum confidence thresholds on the effectiveness of the class association rules technique. In a second experiment, we investigated the impact of the two term feature selection metrics (i.e., information gain and $\chi^2$) on the effectiveness of the naïve Bayes classifier. Finally, in our third experiment, we compared the performance of the class association rules technique and naïve Bayes classifier for sentiment analysis.

### *Data Set*

To prepare a data set for empirical evaluation purposes, we developed a Web crawler to gather online consumer reviews from amazon.com about six digital camera models. For these six digital camera models, we collected 214 consumer reviews and 3,000 opinion sentences, with an average of 14.02 sentences in each consumer review. The six digital camera models were:

- Canon EOS 20D 8.2MP Digital SLR Camera
- Nikon D70 Digital SLR Camera Kit
- Canon Powershot SD300 4MP Digital Elph Camera with 3x Optical Zoom
- Sony Cybershot DSCP200 7.2MP Digital Camera 3x Optical Zoom
- Canon Powershot S2 IS 5MP Digital Camera with 12x Optical Image Stabilized Zoom
- Canon Powershot A95 5MP Digital Camera with 3x Optical Zoom

After segmenting the collected consumer reviewers into sentences, we labeled each sentence as containing zero or more product features. Across this whole data set, we identified eight product features: (1) battery, (2) flash, (3) image quality, (4) lens, (5) memory, (6) price, (7) usability, and (8) video.

### *Evaluation Metrics*

We employed precision, recall, and F-measure as evaluation metrics; they are common in information retrieval and document classification research. Precision measures the number of correctly classified items out of the total classified by a classification technique, and recall measures the amount of correctly classified items of those manually classified as the gold standard. The F-measure is the harmonic mean of precision and recall, which offers a better measure than the arithmetic mean of precision and recall, because it is not strongly affected by extreme values of either.

For any product feature *f*, the precision ($p_f$), recall ($r_f$), and F-measure (*F measure*$_f$) can be computed as follows:

$$p_f = \frac{TP_f}{TP_f + FP_f}$$

$$r_f = \frac{TP_f}{TP_f + FN_f}$$

$$F\ measure_f = \frac{2 \times p_f \times r_f}{p_f + r_f}$$

where $TP_f$ is the number of opinion sentences correctly labeled with *f* by a classification technique, $FN_f$ is the number of opinion sentences incorrectly labeled with other product features by a classification technique that should be labeled with *f* according to the gold standard, and $FP_f$ is the number of opinion sentences incorrectly labeled with *f* by a classification technique.

We used micro and macro measurements of precision and recall to evaluate the overall performance. The micro and macro measurements are useful to determine if a classification technique under investigation performs better in particular product feature class or performs equally well across all product feature classes because the class sizes may not be uniform. Thus,

$$micro\text{ - }precision = \frac{\sum_{f \in F} TP_f}{\sum_{f \in F} TP_f + \sum_{f \in F} FP_f}$$

$$micro\text{ - }recall = \frac{\sum_{f \in F} TP_f}{\sum_{f \in F} TP_f + \sum_{f \in F} FN_f}$$

$$micro\ F\text{ - }measure = \frac{2 \times micro\text{ - }precision \times micro\text{ - }recall}{micro\text{ - }precision + micro\text{ - }recall}$$

$$macro\text{-}precision = \frac{\sum_{f \in F} p_f}{|F|}$$

$$macro\text{-}recall = \frac{\sum_{f \in F} r_f}{|F|}$$

$$macro\ F\text{ - }measure = \frac{2 \times macro\text{ - }precision \times macro\text{ - }recall}{macro\text{ - }precision + macro\text{ - }recall}$$

### Experiment I

We used a fivefold cross-validation to evaluate the impact of the minimum support and confidence thresholds on the class association rules technique. Specifically, the data set was partitioned into five subsets of approximately equal size. For each fold, a single subset served as the testing set and the remaining four subsets were combined to form the training set. The overall effectiveness was then estimated by averaging the effectiveness obtained from these five folds.

We selected support values ranging from 0.005 to 0.007 and confidence values from 0.35 to 0.65. Figure 1 shows the macro F-measure of the class association rules technique with these support and confidence values; Figure 2 shows the micro F-measure of this technique with the selected support and confidence values. The optimal macro F-measure was 72.26%, and the best micro F-measure was 70.77%. These optimal values emerged when the minimum support value was 0.006 and the minimum confidence value was 0.5. Furthermore, the micro F-measure decreased quickly when the minimum confidence threshold decreased below 0.5 but only slightly when the minimum confidence threshold increased beyond 0.5. In contrast, the effectiveness of the class association rules technique was less sensitive to the minimum support threshold. It can be explained by two observations. The opinion sentences are relatively short, and therefore, there are relatively fewer terms in each opinion sentence. In addition, there are also relatively fewer labeled opinion sentences for some particular product features. That means these product features are not frequently discussed in the online consumer reviews. The impact of the unbalanced data appeared in the form of sensitivity to the minimum confidence threshold.

### Experiment II

As we discussed in Section 3.2, term feature selection is likely to improve the performance of the naïve Bayes classifier. In this experiment, we again used a five-fold cross-validation to investigate the impact of the two term feature selection metrics, information gain and 2, on the effectiveness of the naïve Bayes classifier. Before computing the information gain and 2 values, we filtered out rare terms with sentence frequencies of less than 10, because rare terms lack sufficient support and thus would not be useful for our classification.

Figure 3 shows the macro F-measure of the naïve Bayes classifier using information gain for term feature selection. Figure 4 shows the micro F-measure for the same scenario. The optimal macro F-measure was 70.19%, and the optimal micro F-measure was 67.19%, which occurred when the information gain value was 0.012.

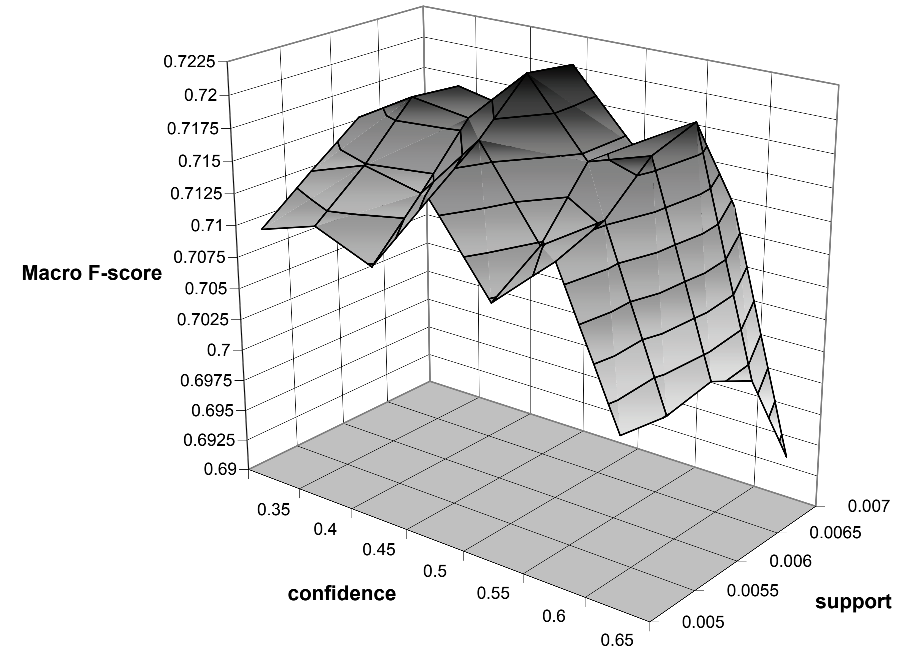

**Figure 1 - Macro F-measure with different minimum support and confidence thresholds**

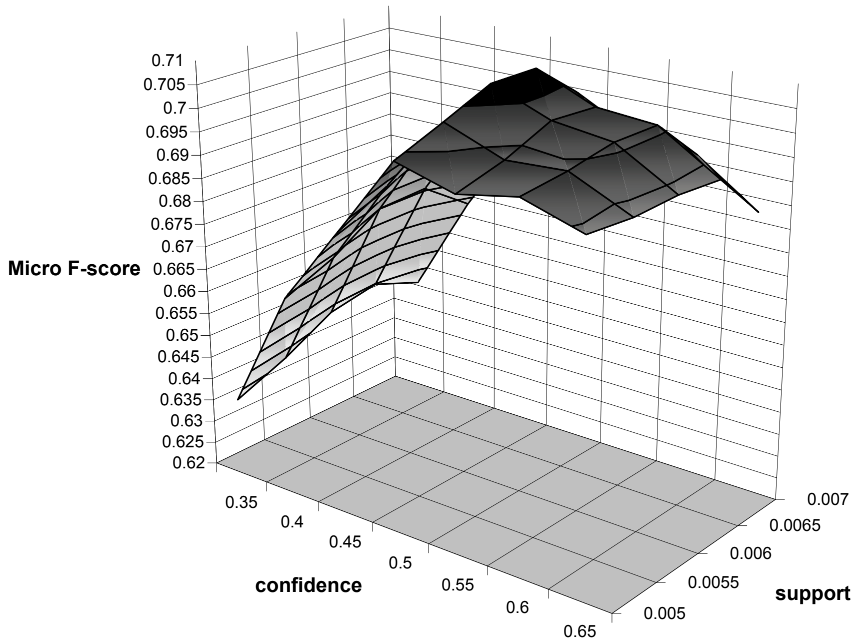

**Figure 2 -Micro F-measure with different minimum support and confidence thresholds**

https://aisel.aisnet.org/pajais/vol2/iss3/7
DOI: 10.17705/1pais.02305

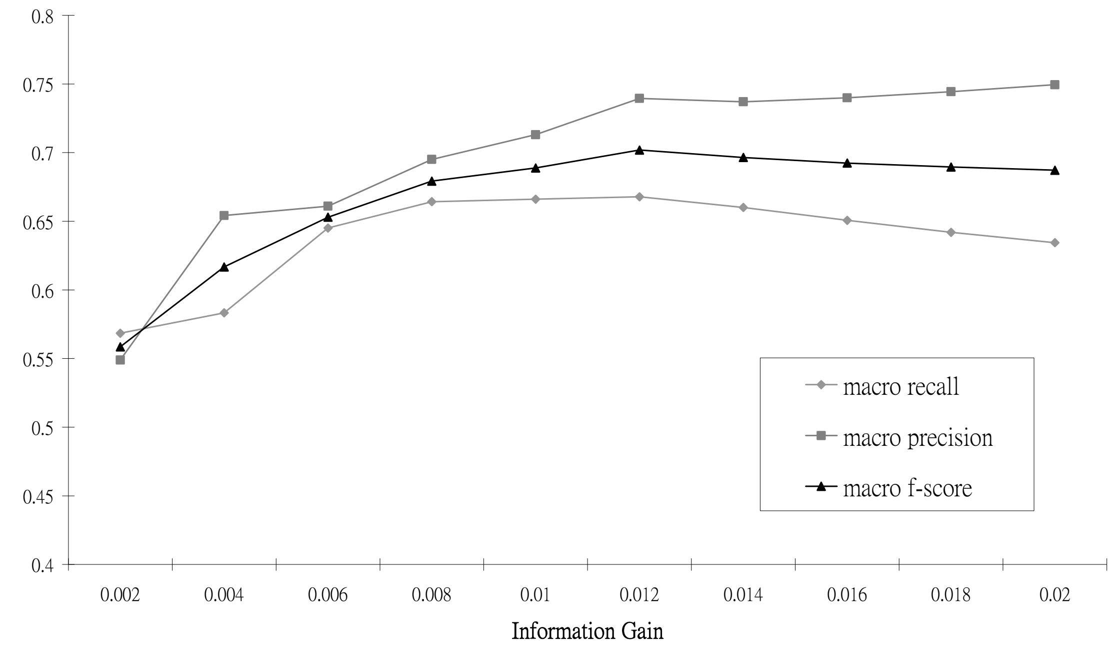

**Figure 3 - Macro F-score values with different information gain thresholds for the naïve Bayes classifier**

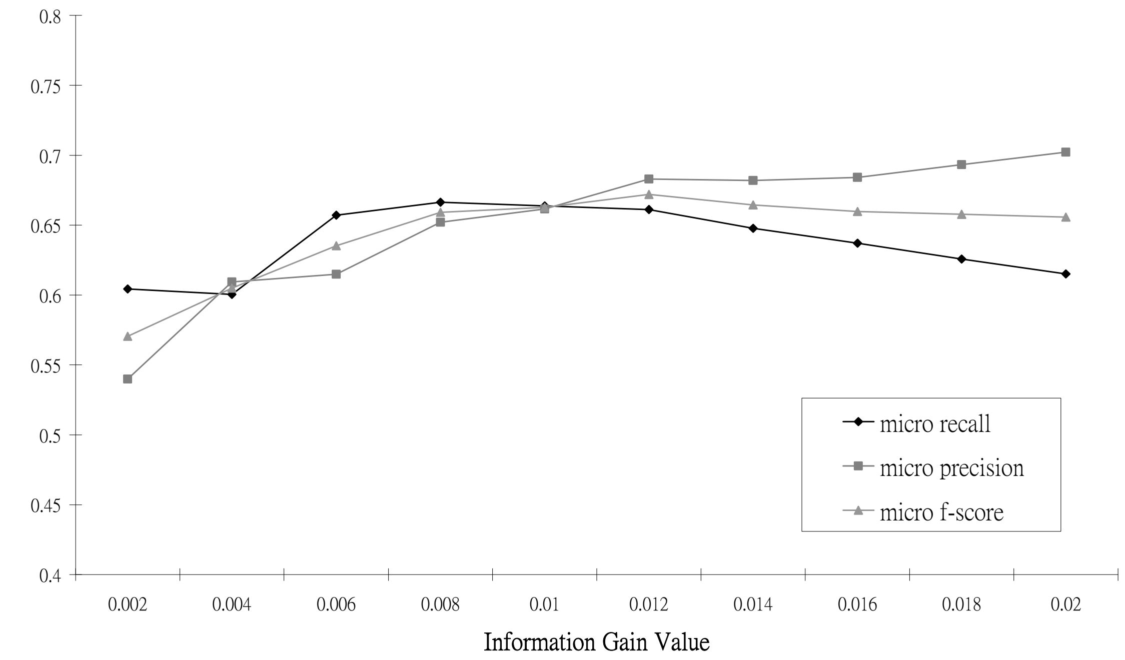

**Figure 4 - Micro F-score values with different information gain thresholds for the naïve Bayes classifier**

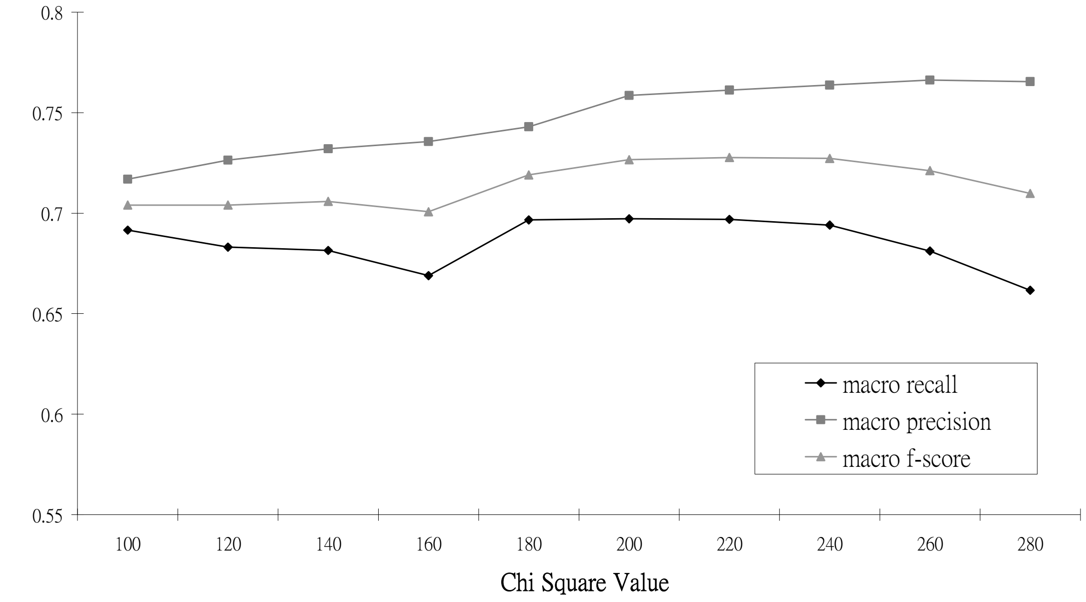

**Figure 5 - Macro F-score values with different $\chi^2$ thresholds for the naïve Bayes classifier**

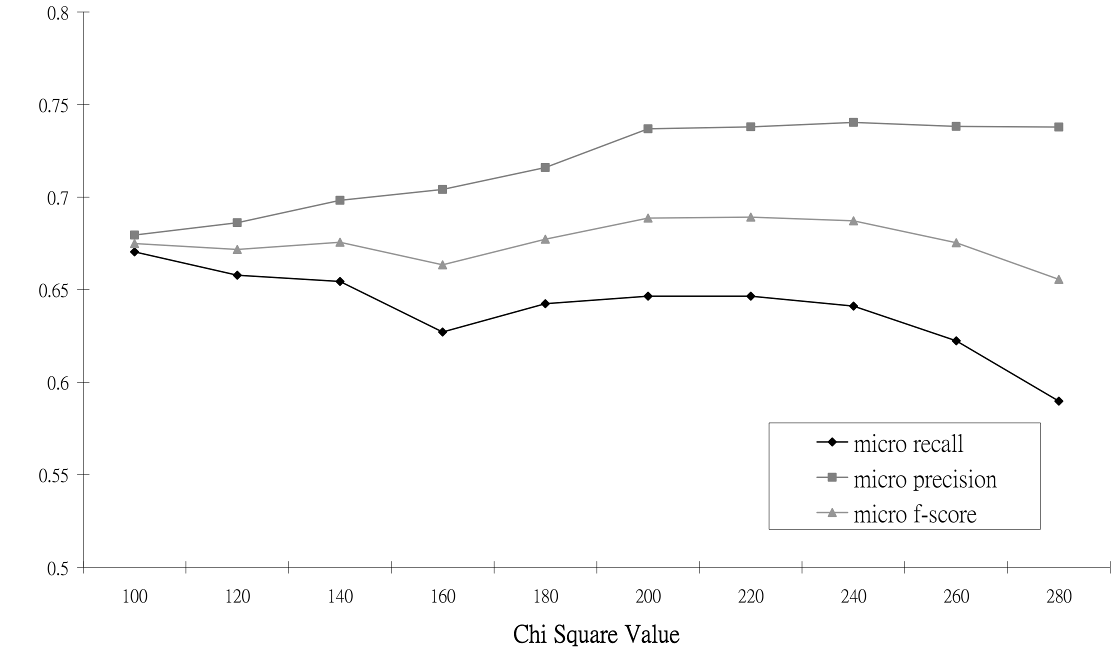

**Figure 6 - Micro F-score values with different $\chi^2$ thresholds for the naïve Bayes classifier**

https://aisel.aisnet.org/pajais/vol2/iss3/7
DOI: 10.17705/1pais.02305

In turn, Figure 5 shows the macro F-measure of the naïve Bayes classier that used the $\chi^2$ metric for term feature selection; Figure 6 contains the concomitant micro F-measure. The optimal macro F-measure of 72.77% and optimal micro F-measure of 68.92% emerged when the $\chi^2$ value was 220. It is also noteworthy that the optimal macro-precision and macro-recall of the naïve Bayes classifier using $\chi^2$ consistently were higher than those of the naïve Bayes classifier using information gain. However, the optimal micro-precision of the naïve Bayes classifier using $\chi^2$ was higher than that of the naïve Bayes classifier using information gain, whereas the optimal micro-recall using $\chi^2$ was lower than the one using information gain. Therefore, the results overall indicated that the $\chi^2$ metric for term feature selection obtained better performance in the naïve Bayes classifier.

### *Experiment III*

To compare across the class association rules technique and naïve Bayes classifier with information gain and $\chi^2$ as term feature selection metrics, we created Table 1 to present their macro-precision, macro-recall, macro F-measure, micro-precision, micro-recall, and micro F-measure. The class association rules technique obtained comparable macro F-measures to those of the naïve Bayes classifier with $\chi^2$, with a difference of only 0.51%. However, the class association rules technique achieved a higher micro F-measure than did the naïve Bayes classifier with $\chi^2$, at a difference of 3.34%. In contrast, the naïve Bayes classifier with $\chi^2$ obtained substantially higher macro- and micro-precision than the class association rules technique, whereas the latter obtained substantially higher macro- and micro-recall than the naïve Bayes classifier with $\chi^2$. The differences in macro- and micro-precision were 8.48% and 6.33%, respectively, whereas the differences in macro- and micro-recall were 7.80% and 9.70%, respectively. Thus, the class association rules technique classified more sentences with correct product features but also generated more errors (false positives). The naïve Bayes classifier instead made more accurate assignments but also missed more sentences in the classifications (false negatives).

The class association rules technique obtained higher recall because it classified opinion sentences to product features whenever there was sufficient support and confidence based on a term feature that appeared in the opinion sentence. However, these terms also could appear in other sentences that were not discussing the predicted product feature, which sacrificed precision. The naïve Bayes classifier used the term feature distribution to determine the probability of an opinion sentence being classified to a product feature, such that some term features contributed to a product feature but others lowered the chance. Using the distribution of terms in an opinion sentence, the naïve Bayes classifier achieved higher precision, but it also produced more false negatives by rejecting some opinion sentences that should have been classified to a product feature.

In general, the supervised learning approach using the class association rules technique or the naïve Bayes classifier achieves satisfactory classification effectiveness. The natural language processing approach commonly employed by prior studies requires excessive manual efforts to modify incorrectly parsed sentences, whereas our proposed supervised learning techniques do not require any natural language processing tools for tagging.

## Conclusion

The Web 2.0 has facilitated user interactions on Internet platforms, in which people share their opinions and contribute their content. Electronic commerce thus has extended beyond B2C to include consumer-to-consumer marketplaces. Consumers have a desire to compare products before making a purchase decision, and online consumer reviews provide valuable information that enables them to identify specific products that fit their personal preferences. However, the vast volume of online consumer reviews available on the Web makes browsing through them and achieving a systematic comparison manually not trivial. In response,

**Table 1 - Comparison of best results provided by different term feature selection metrics**

| | | **Naïve Bayes Classifier** | | **Class Association Rules** |
|---|---|---|---|---|
| | | **Information Gain** | **$\chi^2$** | |
| Macro average | Precision | 73.96% | **76.12%*** | 67.64% |
| | Recall | 66.79% | 69.70% | **77.50%*** |
| | F measure | 70.19% | **72.77%*** | 72.26% |
| Micro average | Precision | 68.30% | **73.80%*** | 67.47% |
| | Recall | 66.11% | 64.70% | **74.40%*** |
| | F measure | 67.19% | 68.92% | **72.26%*** |

we develop supervised learning techniques for sentiment analysis of online consumer reviews. We have investigated the class association rules technique and the naïve Bayes classifier as methods to classify product features that consumers describe in their reviews, which then should produce a summary of comparisons between products at the product feature level. Rather than comparing general ratings of products, sentiment analysis allows users to compare and identify products according to their preferred product features. In our empirical evaluation, the class association rules technique and the naïve Bayes classifier with $\chi^2$ as the term feature selection metric produce comparable results, in terms of their macro F-measure, though the former performs better on the micro F-measure. However, the two methods achieve different results in terms of precision and recall. The naïve Bayes classifier with $\chi^2$ for term feature selection performs substantially better in both macro- and micro-precision; the class association rules technique performs substantially better for both macro- and micro-recall.

In the future, we shall investigate the impact of the size of training set on the effectiveness of these proposed supervised learning techniques. In addition, we shall explore potential unsupervised learning approaches for sentiment analysis, which is challenging because opinion texts tend to be short and sparse, and classifying opinion texts by the similarities of such short and sparse sentences is difficult. Concept mapping represents a potential solution. However, relying on existing ontology is not an ideal solution, because the terms used in social media are constantly evolving. Effective ontology learning methods and appropriate concept mapping mechanisms are potential solutions to this problem.

## References

C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, “Red Opal: Product-Feature Scoring from Reviews,” *Proceedings of the 8th ACM Conference on Electronic Commerce*, San Diego, CA, June 2007, pp. 182-191.

D. M. Blei and J. D. McAuliffe, “Supervised Topic Models,” *Advances in Neural Information Processing Systems (NIPS)*, 2008.

C. Dellarocas, “The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms,” *Management Science* (49:10), October 2003, pp. 1407-1424.

X. Ding, B. Liu, and P. S. Yu, “A Holistic Lexicon-based Approach to Opinion Mining,” *Proceedings of the International Conference on Web Search and Web Data Mining*, Palo Alto, CA, February 2008, pp. 231-240.

C. Forman, A. Ghose, and B. Wiesenfeld, “Examining the Relationship between Reviews and Sales: The Role of Reviewer Identity Discloser in Electronic Markets,” *Information Sys-*

*tems Research* (19: 3), September 2008, pp. 291-313.

D. Godes, and D. Mayzlin, "Using Online Conversations to Study Word of Mouth Communication," *Marketing Science* (23:4), Fall 2004, pp. 545-560.

V. Hatzivassiloglou, and J. Wiebe, "Effects of Adjective Orientation and Gradability on Sentence Subjectivity," *Proceedings of 18th International Conference on Computational Linguistics (COLING)*, Saarbrucken, Germany, 2000, pp. 299-305.

M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 2004a, pp. 168-177.

M. Hu, and B. Liu, "Mining Opinion Features in Customer Reviews," *Proceedings of American Association for Artificial Intelligence (AAAI) Conference*, 2004b, pp. 755-760.

N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," *Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 06)*, Seattle, WA, 2006a, pp. 244-251.

N. Jindal and B. Liu, "Mining Comparative Sentences and Relations," *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-2006)*, July 2006b, Boston, MA, pp. 1331-1336.

N. Kobayashi, K. Inui, and Y. Matsumotto, "Collecting Evaluative Expressions for Opinion Extraction," *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, Hainan Island, China, March 2004, pp. 596-605.

N. Kobayashi, R. Iida, K. Inui, and Y. Matsumotto, "Opinion Extraction Using A Learning-based Anaphora Resolution Technique," *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-04)*, Jeju Island, Korea, October 2005, pp. 173-178.

B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *Proceedings of 2005 World Wide Web (WWW) Conference*, Chiba, Japan, May 2005, pp. 342-351.

Y. Lu and C. Zhai, "Opinion Integration through Semi-supervised Topic Modeling," *Proceeding of the 17th International Conference on World Wide Web*, New York, NY: ACM, 2008, pp. 121-130.

Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs," *Proceedings of the 16th International Conference on World Wide Web 2007*, New York, NY: ACM, 2007, pp. 171-180.

A. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, Canada, 2005, pp. 339-346.

I. Titov and R. McDonald, "Modeling Online Reviews with Multi-grain Topic Models," *Proceeding of the 17th International Conference on World Wide Web*, New York, NY: ACM, 2008, pp. 111-120.

P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proceedings of the 40th Conference on Association for Computational Linguistics (ACL)*, Philadelphia, PA, 2002, pp. 417-424.

D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document Summarization via Sentence-level Semantic Analysis and Symmetric Matrix Factorization," *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, 2008, pp. 307-314.

C. Wei, Y. Chen, C. Yang, and C. C. Yang, "Understanding What Concerns Consumers: A Semantic Approach to Product Feature Extraction from Consumer Reviews," *Journal of Information Systems and E-Business Management* (8: 2), March 2010, pp.149-167.

C. Wei, C. S. Yang, and C. N. Huang, "Turning Online Product Reviews to Customer Knowledge: A Semantic-based Sentiment Classification Approach," *Proceedings of the 10th Pacific Asia Conference on Information Systems*, Kuala Lumpur, Malaysia, 2006.

J. Wiebe, R. Bruce, and T. O'Hara, "Development and Use of a Gold Standard Data Set for Subjectivity Classifications," *Proceedings of the 37th Conference on Association for Computational Linguistics (ACL)*, College Park, MD, 1999, pp. 246-253.

Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *Proceedings of the 14th International Conference on Machine Learning*, July 1997, pp. 412-420.

C. Zhai, A. Velivelli, and B. Yu, "A Cross-collection Mixture Model for Comparative Text Mining," *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 2004, pp. 743-748.

Z. Zhang and B. Varadarajan, "Utility Scoring of Product Reviews," *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, Arlington, VA, November 2006, pp. 51-57.

## About the Authors

**Christopher C. Yang** is an associate professor in the College of Information Science and Technology at Drexel University. He has also been an associate professor in the Department of Systems Engineering and Engineering Management and the director of the Digital Library Laboratory at the Chinese University of Hong Kong, an assistant professor in the Department of Computer Science and Information Systems at the University of Hong Kong and a research scientist in the Department of Management Information Systems at the University of Arizona. His recent research interests include social media analytics, Web 2.0, security informatics, health informatics, Web search and mining, knowledge management, and electronic commerce. He has published over 200 referred journal and conference papers in Journal of the American Society for Information Science and Technology (JASIST), Decision Support Systems (DSS), IEEE Transactions on Systems, Man, and Cybernetics, IEEE Transactions on Image Processing, IEEE Transactions on Robotics and Automation, IEEE Computer, IEEE Intelligent Systems, Information Processing and Management (IPM), Journal of Information Science, Graphical Models and Image Processing, Optical Engineering, Pattern Recognition, International Journal of Electronic Commerce, Applied Artificial Intelligence, ISI, WWW, SIGIR, ICIS, CIKM, and more. He has edited several special issues on multilingual information systems, knowledge management, Web mining, social media, and electronic commerce in JASIST, DSS, IPM, and IEEE Transactions. He chaired and served in many international conferences and workshops. He has also frequently served as an invited panelist in the NSF and other government agencies review panels. He can be reached at chris.yang@drexel.edu.

https://aisel.aisnet.org/pajais/vol2/iss3/7
DOI: 10.17705/1pais.02305

**Xuning Tang** is a Ph.D. candidate in the College of Information Science and Technology at Drexel University. His research interests are social computing and Web mining. He has published in IEEE Intelligent Systems, Annals of Information Systems, IEEE International Conference on Intelligence and Security Informatics, ACM SIGKDD Workshops on Intelligence and Security Informatics.

**Y.C. Wong** received a M.Phil. degree from the Chinese University of Hong Kong. She was a research assistant in the Digital Library Laboratory. Her publication has appeared in the Proceedings of the International Conference on Electronic Commerce.

**Chih-Ping Wei** received a BS in Management Science from the National Chiao-Tung University in Taiwan, R.O.C. in 1987 and an MS and a Ph.D. in Management Information Systems from the University of Arizona in 1991 and 1996. He is currently a professor of Department of Information Management at National Taiwan University. Prior to joining National Taiwan University in 2010, he was a professor of Institute of Service Science and Institute of Technology Management at National Tsing Hua University in Taiwan and a professor of Department of Information Management at National Sun Yat-sen University in Taiwan. He was also a visiting scholar at the University of Illinois at Urbana-Champaign in Fall 2001 and the Chinese University of Hong Kong in Summer 2006 and 2007. His papers have appeared in Journal of Management Information Systems (JMIS), European Journal of Information Systems, Decision Support Systems (DSS), IEEE Transactions on Engineering Management, IEEE Software, IEEE Intelligent Systems, IEEE Transactions on Systems, Man, Cybernetics, IEEE Transactions on Information Technology in Biomedicine, Journal of the American Society for Information Science and Technology, Information Processing and Management, Journal of Database Management, and Journal of Organizational Computing and Electronic Commerce, etc. His current research interests include information retrieval and text mining, knowledge discovery and data mining, knowledge management, multidatabase management and integration, and data warehouse design. He has edited special issues of Decision Support Systems, International Journal of Electronic Commerce, Electronic Commerce Research and Applications, and Information Processing and Management. He can be reached at the Department of Information Management, National Taiwan University, Taipei, Taiwan, R.O.C; cpwei@im.ntu.edu.tw.