Measuring Term Specificity Information for Assessing Sentiment Orientation of Documents in a Bayesian Learning Framework

D. Cai
School of Computing and Engineering
University of Huddersfield, HD1 3DH, UK
Email: d.cai@hud.ac.uk

***Abstract*—The assessment of document sentiment orientation using term specificity information is advocated in this study. An interpretation of the mathematical meaning of term specificity information is given based on Shannon's entropy. A general form of a specificity measure is introduced in terms of the interpretation. Sentiment classification using the specificity measures is proposed within a Bayesian learning framework, and some potential problems are clarified and solutions are suggested when the specificity measures are applied to estimation of posterior probabilities for the NB classifier. A novel method is proposed which allows each document to have multiple representations, each of which corresponds to a sentiment class. Our experimental results show, while both the proposed method and IR techniques can produce high performance for sentiment classification, that our method outperforms the IR techniques.**

***Index Terms*—term specificity information; specificity measure; naive Bayes classifier; sentiment classification.**

I. Introduction

The proliferation of web-centred social interaction has led to increasing quantities of opinion-dense text. The availability of this data, and the range of scientific, commercial, social and political uses to which it may be put, has reinforced interest in opinions as objects of analysis and fuelled the growth of text Sentiment Classification (SC). Sentiment analysis draws from, and contributes to, broad areas of text analytics, natural language processing and computational linguistics. The basic task for the analysis is to classify the polarity of a given text: whether the opinion expressed is positive or negative. Early studies at the whole document level such as [1], [2] used several methods to classify the polarity of product reviews and movie reviews, respectively. Classifying document polarity on n-ary scales, rather than just positive and negative, can be found, for instance, in [3]–[5]. Good reviews of SC methods can be found, for instance, in [5]–[7].

Generally, three main issues need to be considered in statistical methods of SC: i) methodologies to *identify* sentiment-bearing terms; ii) models to *represent* documents with the identified terms; iii) classifiers to *classify* each document by predicting a class that is most likely to generate the document representation. This study focuses on the second issue: design method to represent documents using *Term Specificity Information* (*TSI*) for accurate and reliable SC.

Several classical classifiers, such as Naive Bayes (NB), k-Nearest Neighbours (kNN), Maximum Entropy (ME) and Support Vector Machine (SVM) have been developed further for SC. Studies have shown NB and SVM to be superior methods for SC [8]–[12]. Studies [13], [14] have experimentally shown performance benefits of representing documents using *TSI* along with SVM for SC. Our experimental results (not discussed in this paper) obtained from *TSI* with SVM, also support these conclusions.

In order to develop SC classifiers with a predictive capability, we need to know the explicit representation of the opinionated documents. That is, we have to design a *weighting* function to generate the document representation corresponding to the individual sentiment classes (each class is treated as a sub-collection). The weights of terms may be expected to enhance the likelihood of correctly predicting document sentiment orientation. This stage is crucial for SC, in particular, for estimating the posterior probability required by the NB classifier. There have been extensive studies on document representation in other areas, such as IR, in which a controlled vocabulary is constructed and the weights of carefully selected terms are used to represent the content of documents over the whole collection.

Specificity information measurement can be naturally and conveniently utilized to estimate posterior probabilities required in the NB classifier. Therefore, this study concentrates on SC in a Bayesian learning framework (rather than in SVM), in which, document representation using *TSI* is essential. The NB classifier is surprisingly effective in practice since its classification decision can be correct even if its probability estimates are inaccurate [15], [16], and it often competes well with more sophisticated classifiers [16], [17]. There are theoretical reasons [17] for the apparently unreasonable efficacy of the NB classifier. However, there has been no systematic discussion on how to use *TSI* to represent documents for SC and there exist some potential problems in applying specificity measures to the NB classifier for SC.

It is worth mentioning, rather than considering all terms in documents, that [18] attempts to determine the specificity of nouns. One possible indicator of specificity is how often the noun is modified: a very specific noun is rarely modified, while a very general noun is often modified. There are three categories of the modifiers: (prenominal) adjectives, verbs, or

other nouns. Their study uses the probability that the noun is modified by any of the modifiers and the probability that the noun is modified by each specific category. Their work considers also how much the modifiers vary: a noun with a great variety of modifiers suggests that it is general, whereas a noun rarely modified or modified by only a few different ways is likely specific. Clearly, their work is entirely different from methods presented in this paper. It is evident that the method given in [13] is a special case of one of our methods.

There are three main concerns in this study. First, we interpret the mathematical meaning of a basic concept on specificity information conveyed by a given term based on Shannon's entropy, and introduce a formal definition of a specificity measure in terms of the interpretation. Second, we propose a general method to represent the statistical importance of terms pertaining to individual documents with estimation of posterior probabilities using term weights obtained from *TSI* for the NB classifier. Third, we clarify some potential problems inherent in applying the specificity measures in a Bayesian learning framework and, then suggest solutions that are easy to apply in practice. Our methods allow each document to have multiple representations, each of which corresponds to a specific sentiment class, which we believe is of benefit to SC tasks. In addition, we present some experimental results, evaluating performance against a standard collection, MovieReviews [19], to verify that both *TSI* and the difference of *TSI*s over the individual sentiment classes may be regarded as appropriate measures for SC.

The remainder of the paper is organized as follows. Section 2 focuses on the mathematical interpretation and formal definition of *TSI*. Section 3 proposes a general form of the NB classifier with posterior probability estimation using *TSI*. Section 4 clarifies problems of applying *TSI* and suggests solutions. Some experimental results of our method are presented in Section 5 and conclusions are drawn in Section 6.

II. Term Specificity Information (*TSI*)

This section gives a mathematical interpretation and formal definition of specificity information of terms.

To begin, let us introduce the notation. Let C be a *collection* of documents and $d \in C$ be a document. Let $\mathcal{C}$ be the *classification* of documents over C and $X \in \mathcal{C}$ (or, $X \subseteq C$) be a *class*. Let V be a *vocabulary* of all the terms used to index individual documents. Let $V_X \subseteq V$ be the sub-vocabulary consisting of those terms appearing in at least one document $d \in X$ and $V_d \subseteq V$ be the set of terms appearing in document $d \in C$.

For simplicity, all our discussions are set to the situation where $|\mathcal{C}| = 2$. Such a setting can be easily generalized to any finite number of classes. Thus, we have $\mathcal{C} = \{X, \bar{X}\}$, where $X = C_P$ (or, $X = C_N$) is a possible sentiment class consisting of all positive (or, negative) documents. Generally, $V_X \cap V_{\bar{X}} \neq \emptyset$, as terms often occur in both positive and negative documents.

A. A General Form of a TSI Measure

Intuitively, a term is said to contain *specificity* information if it tends to be capable of isolating the few documents of interest from many others.

Consider a conditional probability distribution $P_x(d|t)$ satisfying: $P_x(d|t) \geq 0$ and $\sum_{d \in X} P_x(d|t) = 1$. The entropy function (Shannon's entropy) of $P_x(d|t)$ is

$$H\big(P_x(d|t)\big) = -\sum_{d \in X} P_x(d|t) \log P_x(d|t)$$

where $P_x(d|t)$ is called the document frequency distribution (over X) of term $t \in V_X$. We here adopt the notational convention: $y \log(y) = 0$ if $y = 0$.

Note that, from the properties of the entropy function, if term t is uniformly distributed over X:

$$P_x(d|t) = \frac{1}{|X|} \quad \text{for every } d \in X$$

where $|X|$ is the cardinality of X, then the entropy of t arrives at the maximum:

$$H\big(P_x(d|t)\big) = -\sum_{d \in X} \frac{1}{|X|} \log \frac{1}{|X|} = \log\big(|X|\big) = H_{max}$$

which is called the maximum entropy of term t. Clearly, we have $H_{max} \geq 0$ as $|X| \geq 1$. $H\big(P_x(d|t)\big)$ can be regarded as a measure of the degree of uncertainty based on what we know about t concerning X. Thus, t is said to be more informative than $t' \in V_X$ if $H\big(P_x(d|t)\big) < H\big(P_x(d|t')\big)$ as t reduces uncertainty. The reduction based on $H\big(P_x(d|t)\big)$ essentially amounts to specificity information of t.

The above statements may already mathematically interpret what it is meant by the basic concept of specificity information conveyed by term t. Thus, we can now introduce a formal definition as follows.

Definition 2.1 For a given class $X \in \mathcal{C}$ and an arbitrary term $t \in V_X$, suppose $P_x(d|t)$ be the conditional probability distribution over X. A general form of a *term specificity information* measure, denoted by $tsi_x(t)$, is defined by

$$tsi_x(t) = \begin{cases} H_{max} - H\big(P_x(d|t)\big) & t \in V_X \\ undefined & t \in V - V_X \end{cases} \quad (1)$$

which measures the extent of uncertainty reduction caused by, or the amount of specificity information of, t concerning X.

Clearly, we have $tsi_x(t) \geq 0$ for every $t \in V_X \subseteq V$. A basic idea for $tsi_x(t)$ is: if term t has a skewed document frequency distribution, $P_x(d|t)$, over X, then t may be expected to be a good discriminator for distinguishing the few documents of interest from many others in X.

If we accept the assumption that the importance of a term in representing each document is dependent significantly, if not completely, on its specificity over the individual classes, the problem is then reduced to choosing a suitable specificity measure. With Definition 2.1, we discuss below two concrete specificity measures to clarify ideas involved in the general form given in Eq.(1).

B. Example TSI Measures

Two well-known specificity measures, $idf_x(t)$ and $int_x(t)$, as examples, are reconsidered to illustrate the general form, and the relationship between the two specificity measures are established based on the general form.

Example 2.1 Perhaps the most well-known measure capturing the specificity information of term t concerning some class X is the *inverse document frequency* [20]:

$$idf_x(t) = \begin{cases} \log \frac{|X|}{n_x(t)} & t \in V_X \\ undefined & t \in V - V_X \end{cases} \quad (2)$$

where $n_x(t)$, called the *document frequency* of t in X, is the number of documents in X in which term t occurs.

In order to interpret $idf_x(t)$ in terms of the entropy function as given in Eq.(1), let us consider documents represented by binary vectors. Note that t appears in at least one document of X, so $n_x(t) \neq 0$, for every $t \in V_X$. Then, for a given $t \in V_X$, the document frequency distribution for the binary representation is:

$$P_x(d|t) = \begin{cases} \frac{1}{n_x(t)} & d \in X_t \\ 0 & d \in X - X_t \\ undefined & d \in D - X \end{cases} \quad (3)$$

where $X_t \subseteq X$ is the set of document(s) in which t appears. Thus, we obtain

$$H\big(P_x(d|t)\big) = -\sum_{d \in X} \frac{1}{n_x(t)} \log \frac{1}{n_x(t)} = \log\big(n_x(t)\big)$$

Hence, from Eq.(2) and Eq.(3), we have

$$idf_x(t) = \log\big(|X|\big) - \log\big(n_x(t)\big) = H_{max} - H\big(P_x(d|t)\big)$$

which is the exact expression given in Eq.(1) when $t \in V_X$.

The measure $idf_x(t)$ states that the specificity of term $t \in V_X$ is inversely proportional to the document frequency over X. Therefore, it assigns higher values to more specific terms that tend to be capable of isolating few documents from the many others. However, $idf_x(t)$ does not take into consideration term frequency within documents, and terms with the same document frequency will be treated equally by assigning the same weights. □

Example 2.2 A more accurate indication of term importance may be obtained by incorporating term frequency information into the document frequency distribution, which is *noise* of a term [21], it may be used to capture the unspecificity information of term $t \in V_X$ concerning some class X:

$$noise_x(t) = H\big(P_x(d|t)\big) = -\sum_{d \in X} \frac{f_d(t)}{f_x(t)} \log \frac{f_d(t)}{f_x(t)}$$

which is the entropy of the document frequency distribution:

$$P_x(d|t) = \begin{cases} \frac{f_d(t)}{f_x(t)} & t \in V_X \\ undefined & t \in V - V_X \end{cases} \quad (4)$$

where $f_d(t)$ is the frequency of t in d, $f_x(t) = \sum_{d \in X} f_d(t)$ is the total frequency of t in X. In other words, $noise_x(t)$ measures the extent of the lack of concentration of occurrence of t; it emphasizes the uselessness of those terms that are in agreement with $P_x(d|t)$ for all the documents in X.

Note that the specificity of term t is in inverse relation to its noise. Thus, the specificity of t may be computed, for instance, by

$$int_x(t) = \begin{cases} H_{max} - noise_x(t) & t \in V_X \\ undefined & t \in V - V_X \end{cases} \quad (5)$$

which, called the *inverse noise* of t, is the same expression given in Eq.(1).

Because the measure $int_x(t)$ assigns low values to those terms that are not concentrated in a few particular documents, but instead are prevalent in X, it should be an appropriate measure of term specificity. □

It is worth mentioning that there are two statistical concepts [22] used widely to test the performance of a binary classification: *sensitivity* of a test is the proportion of actual positives which are correctly predicted; *specificity* of a test is the proportion of negatives which are correctly predicted. Clearly, they are entirely different from our above discussion (i.e., the specificity of a term, rather than a test) and used in different contexts: sensitivity and specificity estimate the ability of the tests to predict positive and negative results, respectively.

III. Sentiment Classification Using *TSI*

So far, we have given a formal account of *TSI*. We are now in a position to see how the NB classifier, along with estimation of posterior probabilities using term weights obtained from *TSI*, can be used for effective SC.

A. The NB Classifier

The NB classifier is a learning method that requires an estimate of the posterior probability that a document belongs to some sentiment class, and then it classifies the document into the class with the highest posterior probability.

More specifically, the NB classifier is constructed based on Bayes' theorem with a strong conditional independence assumption. That is, for a possible sentiment class $X = C_P$ (or $X = C_N$), it computes the posterior probability, $p(X|d)$, that document $d \in C$ belongs to X:

$$p(X|d) = \frac{p(X)}{p(d)} \cdot p(d|X) \propto\ p(X) \cdot \prod_{t \in V_d} p(t|X) \quad (6)$$

where $p(t|X)$ is the conditional probability of term t occurring in some document of class X, $p(d)$ is the probability that a randomly picked document is d, and $p(X)$ is the probability that a randomly picked document belongs to class X. Note that $p(d)$ in Eq.(6) can be omitted as it is a scaling factor dependent only on terms, and that $p(t|X)$ may be interpreted as a measure of evidence of how much contribution t makes to support class X. Taking logarithms of probabilities on both sides of Eq.(6), we can write the NB Classifier by:

$$\Gamma(d, X) = \log\big(p(X)\big) + \sum_{t \in V_d} \log\big(p(t|X)\big) \quad (7)$$

given $0 < p(X) < 1$ and $p(t|X) > 0$ (where $t \in V_d$). Then document d is classified into class X^* if it has the highest posterior probability or, equivalently, it satisfies:

$$\Gamma(d, X^*) = \max\left\{\Gamma(d, X), \Gamma(d, \bar{X})\right\}$$

The parameters given in Eq.(7), such as, *a priori* probability $p(X)$ and the *posterior* probability $p(t|X)$ may be estimated by

$$p(X) = \frac{|X|}{|C|} \tag{8}$$

$$p(t|X) = \begin{cases} \frac{\varpi_X(t)}{\sum_{t \in V} \varpi_X(t)} & t \in V_X \\ undefined & t \in V - V_X \end{cases} \tag{9}$$

where $\varpi_X(t)$ is a *weighting function* estimating the importance of term $t \in V_X$ in representing class X.

Estimation of $p(X)$ is normally straightforward, it may be, for instance, the ratio of class cardinalities of X and C as given in Eq.(8). Estimation of $p(t|X)$ is however the main concern of studies and our discussion below is based on using term weights obtained from *TSI* as discussed in the last section.

B. Estimation of Posterior Probabilities

It can be seen, from Eq.(9), that estimation of $p(t|X)$ is uniquely determined by its argument $\varpi_X(t)$. Generally, we can express

$$\varpi_X(t) = \begin{cases} \sum_{d \in X} \pi(d) \cdot w_{d|X}(t) & t \in V_X \\ undefined & t \in V - V_X \end{cases} \tag{10}$$

in which, $w_{d|X}(t)$ is a *weighting function* estimating the importance of t in representing document $d \in X$; $\pi(d)$ is a function indicating the presumed importance of d in X. Thus, $\varpi_X(t)$ is the sum of weights, multiplied by the importance of the corresponding d, of term $t \in V_d$ over all documents $d \in X$.

It is now clear, with the general expression given in Eq.(10), that estimation of $p(t|X)$ is reduced to estimation of two components, $w_{d|X}(t)$ and $\pi(d)$, of $\varpi_X(t)$.

1) Estimation of $w_{d|X}(t)$

As we know, document representation, $w_{d|X}(t)$, plays an essential role in determining SC effectiveness. The issue of accuracy and validity of representation has long been a crucial and open problem. It is beyond the scope of this paper to discuss the issue in greater detail. A detailed discussion about representation techniques may be found, for instance, in [23].

Our concern is with applying *TSI* for the estimation of posterior probability required in the NB classifier. Therefore, in order to give a general expression of $w_{d|X}(t)$ incorporating term specificity information, we need to introduce a further piece of notation—we need to define the intuitive concept of specificity strength of terms over the classification.

Definition 3.1 Suppose we have a classification $\mathcal{C} = \{X, \bar{X}\}$. The *specificity strength* of term t in support of X against $\bar{X}$ is defined by

$$\Delta tsi_{(X:\bar{X})}(t) = \begin{cases} tsi_X(t) - tsi_{\bar{X}}(t) & t \in V_X \cap V_{\bar{X}} \\ undefined & t \in V - V_X \cap V_{\bar{X}} \end{cases} \tag{11}$$

where $tsi_X(t)$ is the *TSI* measure given in Eq.(1).

Obviously, the larger the *difference* is, the more specificity information term t conveys in support of X against $\bar{X}$. Thus, $\Delta tsi_{(X:\bar{X})}(t)$ may be regarded as the specificity strength of t over $\mathcal{C}$ and as an appropriate measure for SC. Clearly, unlike $tsi_X(t)$, $\Delta tsi_{(X:\bar{X})}(t) \geq 0$ may or may not hold for every $t \in V_X \cap V_{\bar{X}}$.

Now we are ready to formally write $w_{d|X}(t)$. Suppose each document $d \in X$ can be represented as a $1 \times n$ matrix $M_{d|X} = \left[w_{d|X}(t)\right]$. With Definitions 2.1 and 3.1, a general expression of a weighting function incorporating term specificity information is defined as follows.

Definition 3.2 Suppose we have a classification $\mathcal{C} = \{X, \bar{X}\}$. A general form of the *weight* of term t in representing document $d \in X$ is defined by

$$w_{d|X}(t) = \begin{cases} w_{d|X}\left(f_d(t), \Im_{(X:\bar{X})}(t)\right) & t \in V_X \cap V_{\bar{X}} \\ undefined & t \in V - V_X \cap V_{\bar{X}} \end{cases} \tag{12}$$

where $\Im_{(X:\bar{X})}(t)$ is the *TSI* measure given in either Eq.(1) or Eq.(11).

It is worth emphasizing that we here express the weighting function by $w_{d|X}(t)$ rather than by $w_d(t)$. That is, our method facilitates SC with the NB classifier: it allows each document to have multiple representations, each of which corresponds to a specific sentiment class X. Estimation of term weights has been extensively studied in the area of IR. However, in traditional IR, document d is represented by a single weighting function $w_d(t)$ corresponding to the whole collection C.

Example 3.1 We may write a number of weighting functions. Eight weighting functions, derived immediately from Eq.(2) and Eq.(5), are given in Table I. The eight functions, and their variations, are widely applied in many applications (and they will be used in our experiments presented in Section 5). □

TABLE I
EIGHT WEIGHTING FUNCTIONS $w_{d|X}(t)$

Symbols	Descriptions
idf	$idf_X(t)$
tf·idf	$f_d(t) \cdot idf_X(t)$
int	$int_X(t)$
tf·int	$f_d(t) \cdot int_X(t)$
Δidf	$idf_X(t) - idf_{\bar{X}}(t) = \log \frac{p(X)}{1-p(X)} - \log \frac{n_X(t)}{n_{\bar{X}}(t)}$
tf·Δidf	$f_d(t) \cdot \left[idf_X(t) - idf_{\bar{X}}(t)\right]$
Δint	$int_X(t) - int_{\bar{X}}(t) = \frac{H\left(P_{\bar{X}}(d\|t)\right)}{\log(\|\bar{X}\|)} - \frac{H\left(P_X(d\|t)\right)}{\log(\|X\|)}$
tf·Δint	$f_d(t) \cdot \left[int_X(t) - int_{\bar{X}}(t)\right]$

We point out that [13] showed good performance using measure $\Delta\text{tfidf} = \log \frac{p(X)}{1-p(X)}$, along with SVM, for SC. It

is now clear that their measure is a special case of $\mathsf{tf}\cdot\Delta\mathsf{idf}$ (i.e., when $|X| = |\bar{X}|$).

2) Estimation of $\pi(d)$

There may be many ways to construct function $\pi(d)$. Two functions given in the example below indicate how they can be applied in practice.

Example 3.2 Let $\mathcal{V} = \{\mathcal{V}_X, \mathcal{V}_{\bar{X}}\} \subset V$ be the set of all sentiment-bearing terms selected, in which, $\mathcal{V}_X$ is the subset consisting of all positive (or, negative) terms. Generally, we have $\mathcal{V}_X \cap \mathcal{V}_{\bar{X}} = \emptyset$, but $\mathcal{V}_X \cap V_{\bar{X}} \neq \emptyset$ (or, $\mathcal{V}_{\bar{X}} \cap V_X \neq \emptyset$), that is, a strong positive (or, negative) sentiment-bearing term may also occur in a negative (or, positive) document. Thus, for a given document $d \in C$, we may write a function:

$$\pi_1(d) = \begin{cases} \mu \cdot \left[1 + \frac{|V_d \cap \mathcal{V}_X|}{L_d}\right] & d \in X,\ V_d \cap \mathcal{V}_{\bar{X}} = \emptyset \\ \mu_1 & d \in X,\ V_d \cap \mathcal{V}_{\bar{X}} \neq \emptyset \\ \mu_2 & d \in \bar{X} \end{cases}$$

In particular, when $\mu = 0$, we have

$$\pi_2(d) = \begin{cases} \mu_1 & d \in X \\ \mu_2 & d \in \bar{X} \end{cases}$$

where $\mu, \mu_1, \mu_2 \geq 0$ are constants and $L_d = \sum_{t \in V_d} f_d(t)$ is the *length* of d. □

The function $\pi_1(d)$ may involve SC using a small set of strong *sentiment-bearing* terms. For instance, two lists of strong positive and negative terms may be

$$\mathcal{V}_X = \{admirable, beautiful, creative, delicious, excellent, ...\}$$
$$\mathcal{V}_{\bar{X}} = \{aggravated, bored, confused, depressed, enraged, ...\}$$

respectively. The terms in the lists may be obtained in manual term selection and, hence, they may or may not be relevant to domains of interest or to training topics. Clearly, when taking $\mu \geq \mu_i$ $(i = 1, 2)$, $\pi_1(d)$ assigns a relatively higher value to those documents that contain many strong sentiment-bearing terms in $\mathcal{V}_X$ but contain no strong sentiment-bearing term in $\mathcal{V}_{\bar{X}}$; $\pi_1(d)$ is normally needed for applications where a set of good samples for learning is essential.

The function $\pi_2(d)$ is a special case of $\pi_1(d)$: there is no a set of strong sentiment-bearing terms and, thus it assigns the same value to all documents in X (or, $\bar{X}$). $\pi_2(d)$ is simple and may be the most commonly used function in practice: it indicates that all documents within X (or, $\bar{X}$) are treated as equally important; $\pi_2(d)$ may be needed when one has no particular reason to emphasize any document in X (or $\bar{X}$).

IV. Problems Applying *TSI* for SC

It seems that our method is a straightforward application of *TSI*, but it has some potential pitfalls. This section reveals problems and suggests solutions when applying the *TSI* measures for estimation of posterior probabilities for the NB classifier.

A. Problems

Let us first consider a simple example below, in which, the document frequency distributions are derived by expressions Eq.(3) and Eq.(4) and the values of term specificity information are computed by measures given in Eq.(2) and Eq.(5).

Example 4.1 Suppose we are given $C = \{d_1, ..., d_7\}$, $C_P = \{d_1, ..., d_4\}$, $C_N = \{d_5, d_6, d_7\}$ and $V = \{t_1, ..., t_6\}$. Then we have $V_{C_P} = \{t_1, t_2, t_3, t_4, t_6\}$, $V_{C_N} = \{t_1, t_4, t_5, t_6\}$, and $V_{C_P} \cap V_{C_N} = \{t_1, t_4, t_6\}$. Thus, the term occurrence frequencies and the document frequency distributions are shown in Tables II and III, respectively, and the values of term specificity information computed by $tsi_{_X}(t) = idf_{_X}(t)$ and $tsi_{_X}(t) = int_{_X}(t)$ are given in Table IV. For instance, for $t_1 \in V_{C_P}$, we have

$$\begin{aligned} noise_{_{C_P}}(t_1) &= -\sum_{d \in C_P} \frac{f_d(t_1)}{f_{_{C_P}}(t_1)} \log \frac{f_d(t_1)}{f_{_{C_P}}(t_1)} \\ &= -\Big[\frac{1}{7}\log\frac{1}{7} + \frac{3}{7}\log\frac{3}{7} + \frac{1}{7}\log\frac{1}{7} + \frac{2}{7}\log\frac{2}{7}\Big] \\ &= -\frac{1}{7} \cdot \log\frac{3^3 \times 2^2}{7^7} \end{aligned}$$

with expression Eq.(5) and $H_{max} = \log(|C_P|)$, we obtain

$$\begin{aligned} int_{_{C_P}}(t_1) &= \log(|C_P|) - noise_{_{C_P}}(t_1) \\ &= 4 - [-\frac{1}{7} \cdot \log\frac{108}{7^7}] \end{aligned}$$

Note that, in the above computation, we adopt the notational conventions: $y\log(y) = 0$ if $y = 0$. □

TABLE II
Term occurrence frequencies

	$f_d(t_1)$	$f_d(t_2)$	$f_d(t_3)$	$f_d(t_4)$	$f_d(t_5)$	$f_d(t_6)$
d_1	1	2	0	1	0	0
d_2	3	0	2	0	0	1
d_3	1	0	3	2	0	0
d_4	2	3	1	0	0	0
d_5	0	0	0	1	2	3
d_6	1	0	0	0	2	2
d_7	0	0	0	2	3	2

Some problems arise from the above example. First, for a given class X, the specificity measures $tsi_{_X}(t)$ and $\Delta tsi_{_{(X:\bar{X})}}(t)$ are meaningless for every $t \in V - V_X$ and for every $t \in V - (V_X \cap X_{\bar{X}})$, respectively. That is, some terms may have no *TSI* values. For instance, from Table IV, we can see that there is no specificity information concerning C_P for $t_5 \notin V_{C_P}$, concerning C_N for $t_2, t_3 \notin V_{C_N}$, concerning both C_P and C_N for $t_2, t_3, t_5 \in V - (V_{C_P} \cap V_{C_N})$.

Secondly, as mentioned previously, $\Delta tsi_{_{(X:\bar{X})}}(t) \geq 0$ may not hold for every $t \in V_X \cap V_{\bar{X}}$. That is, it may assign negative weights to some terms that occur in both X and $\bar{X}$. For instance, from Table IV, we can see $\Delta tsi_{_{(V_{C_P}:V_{C_N})}}(t) < 0$ for $idf_{_X}(t)$ when $t_1 \in V_{C_P} \cap V_{C_N}$ and for $int_{_X}(t)$ when $t_6 \in V_{C_P} \cap V_{C_N}$. Therefore, when $\Delta tsi_{_{(X:\bar{X})}}(t)$ is applied, function $w_{d|X}(t)$ given in Eq.(12) may assign a negative weight to some terms and $\varpi_{_X}(t)$ given in Eq.(10) cannot be

TABLE III
DOCUMENT FREQUENCY DISTRIBUTIONS $P_{X}(d|t)$

	t_1	t_2	t_3	t_4	t_5	t_6
for calculating $tsi_{X}(t) = idf_{X}(t)$						
$P_{C_P}(d_1\|t)$	$\frac{4}{4}$	$\frac{4}{2}$	$\frac{4}{3}$	$\frac{4}{2}$	-	$\frac{4}{1}$
$P_{C_P}(d_2\|t)$	$\frac{4}{4}$	$\frac{4}{2}$	$\frac{4}{3}$	$\frac{4}{2}$	-	$\frac{4}{1}$
$P_{C_P}(d_3\|t)$	$\frac{4}{4}$	$\frac{4}{2}$	$\frac{4}{3}$	$\frac{4}{2}$		$\frac{4}{1}$
$P_{C_P}(d_4\|t)$	$\frac{4}{4}$	$\frac{4}{2}$	$\frac{4}{3}$	$\frac{4}{2}$	-	$\frac{4}{1}$
$P_{C_N}(d_5\|t)$	$\frac{3}{1}$	-	-	$\frac{3}{2}$	$\frac{3}{3}$	$\frac{3}{3}$
$P_{C_N}(d_6\|t)$	$\frac{3}{1}$	-	-	$\frac{3}{2}$	$\frac{3}{3}$	$\frac{3}{3}$
$P_{C_N}(d_7\|t)$	$\frac{3}{1}$	-	-	$\frac{3}{2}$	$\frac{3}{3}$	$\frac{3}{3}$
for calculating $tsi_{X}(t) = int_{X}(t)$						
$P_{C_P}(d_1\|t)$	$\frac{1}{7}$	$\frac{2}{5}$	$\frac{0}{6}$	$\frac{1}{3}$	-	$\frac{0}{1}$
$P_{C_P}(d_2\|t)$	$\frac{3}{7}$	$\frac{0}{5}$	$\frac{2}{6}$	$\frac{0}{3}$	-	$\frac{1}{1}$
$P_{C_P}(d_3\|t)$	$\frac{1}{7}$	$\frac{0}{5}$	$\frac{3}{6}$	$\frac{2}{3}$	-	$\frac{0}{1}$
$P_{C_P}(d_4\|t)$	$\frac{2}{7}$	$\frac{3}{5}$	$\frac{1}{6}$	$\frac{0}{3}$	-	$\frac{0}{1}$
$P_{C_N}(d_5\|t)$	$\frac{0}{1}$	-	-	$\frac{1}{3}$	$\frac{2}{7}$	$\frac{3}{7}$
$P_{C_N}(d_6\|t)$	$\frac{1}{1}$	-	-	$\frac{0}{3}$	$\frac{2}{7}$	$\frac{2}{7}$
$P_{C_N}(d_7\|t)$	$\frac{0}{1}$	-	-	$\frac{2}{3}$	$\frac{3}{7}$	$\frac{2}{7}$

TABLE IV
TERM SPECIFICITY INFORMATION

	$tsi_{C_P}(t)$	$tsi_{C_N}(t)$	$\Delta tsi_{(C_P:C_N)}(t)$
obtained from $tsi_{X}(t) = idf_{X}(t)$			
t_1	$\log \frac{4}{4}$	$\log \frac{3}{1}$	$-\log 3$
t_2	$\log \frac{4}{2}$	-	-
t_3	$\log \frac{4}{3}$	-	-
t_4	$\log \frac{4}{2}$	$\log \frac{3}{2}$	$\log 2 - \log \frac{3}{2}$
t_5	-	$\log \frac{3}{3}$	-
t_6	$\log \frac{4}{1}$	$\log \frac{3}{3}$	$\log 4$
obtained from $tsi_{X}(t) = int_{X}(t)$			
t_1	$4 + \frac{1}{7} \cdot \log \frac{108}{7^7}$	0	$4 + \frac{1}{7} \cdot \log \frac{108}{7^7}$
t_2	$4 + \frac{1}{5} \cdot \log \frac{108}{5^5}$	-	-
t_3	$4 + \frac{1}{6} \cdot \log \frac{36}{6^5}$	-	-
t_4	$4 + \frac{1}{3} \cdot \log \frac{4}{3^3}$	$3 + \frac{1}{3} \cdot \log \frac{4}{3^3}$	1
t_5	-	$3 + \frac{1}{7} \cdot \log \frac{432}{7^7}$	-
t_6	0	$3 + \frac{1}{7} \cdot \log \frac{432}{7^7}$	$-(3 + \frac{1}{7} \cdot \log \frac{432}{7^7})$

expected to produce non-negative values for every $t \in V_X \cap V_{\bar{X}}$ and, thus $p(t|X)$ given in Eq.(9) may be non-positive. The negative weights may cause a problem in estimating the posterior probability for the NB classifier.

Thirdly, the estimation of the posterior probabilities are normally the maximum likelihood estimate which are given by weights $\varpi_{X}(t)$ and, thus $p(t|X) = 0$ if $\varpi_{X}(t) = 0$. This is problematic: it wipes out all information conveyed by other terms with non-zero probabilities when they are multiplied (see Eq.(6)); it also makes $\Gamma(d, X)$ given in Eq.(7) meaningless.

B. Solutions

There may be many ways to solve the above three problems. We here suggest some simple ways which are easy to apply in practice.

1) Terms Having No TSI Value

To solve the first problem, for each $t \in V$ ($\supseteq V_X$), let us redefine the specificity measure $tsi_{X}(t)$ given in Eq.(1) to

$$tsi'_{X}(t) = \begin{cases} tsi_{X}(t) \geq 0 & t \in V_X \\ \varepsilon_1 & t \in V - V_X \end{cases} \quad (13)$$

where ε_1, called a *pseudo* weight, is assigned to every $t \in V - V_X$ (i.e., to those terms occurring in only $\bar{X}$). A similar discussion can be given to $tsi'_{\bar{X}}(t)$ with a pseudo weight ε_2 assigned to terms occurring in only X. Generally, we have

$$0 \leq \varepsilon_1, \varepsilon_2 \leq \min\left\{tsi'_{X}(t),\, tsi'_{\bar{X}}(t');\ t \in V_X, t' \in V_{\bar{X}}\right\}$$

Note that V can be partitioned into three disjoint sets:

$$V = (V_X - V_{\bar{X}}) \cup (V_X \cap V_{\bar{X}}) \cup (V - V_X)$$

Thus, in the same manner, we may redefine $\Delta tsi_{(X:\bar{X})}(t)$ given in Eq.(11) to

$$\Delta tsi'_{(X:\bar{X})}(t) = \begin{cases} tsi'_{X}(t) - \varepsilon_2 \geq 0 & t \in V_X - V_{\bar{X}} \\ tsi'_{X}(t) - tsi'_{\bar{X}}(t) & t \in V_X \cap V_{\bar{X}} \\ \varepsilon_1 & t \in V - V_X \end{cases} \quad (14)$$

where $tsi'_{X}(t)$ is given in Eq.(13) and ε_1 and ε_2 are the above pseudo weights. A similar discussion can be given to $\Delta tsi'_{(\bar{X}:X)}(t)$.

Clearly, both $tsi'_{X}(t)$ and $\Delta tsi'_{(X:\bar{X})}(t)$ are meaningful over V. According the results given in Table IV, we may simply take, for instance, $\varepsilon_1 = \varepsilon_2 = 0$ as

$$\min\left\{tsi'_{X}(t),\, tsi'_{\bar{X}}(t'); t \in V_X, t' \in V_{\bar{X}}\right\} = 0$$

for $tsi_{X}(t) = int_{X}(t)$. Thus, the results given in Table V are examples of term specificity information obtained from the redefined specificity measures.

TABLE V
MODIFIED TERM SPECIFICITY INFORMATION

	$tsi'_{C_P}(t)$	$tsi'_{C_N}(t)$	$\Delta tsi'_{(C_P:C_N)}(t)$
obtained from $tsi_{X}(t) = idf_{X}(t)$			
t_1	$\log \frac{4}{4}$	$\log \frac{3}{1}$	$-\log 3$
t_2	$\log \frac{4}{2}$	$\varepsilon_2 = 0$	$\log 2$
t_3	$\log \frac{4}{3}$	$\varepsilon_2 = 0$	$\log \frac{4}{3}$
t_4	$\log \frac{4}{2}$	$\log \frac{3}{2}$	$\log 2 - \log \frac{3}{2}$
t_5	$\varepsilon_1 = 0$	$\log \frac{3}{3}$	$\varepsilon_1 = 0$
t_6	$\log \frac{4}{1}$	$\log \frac{3}{3}$	$2 \log 2$
obtained from $tsi_{X}(t) = int_{X}(t)$			
t_1	$4 + \frac{1}{7} \cdot \log \frac{108}{7^7}$	0	$4 + \frac{1}{7} \cdot \log \frac{108}{7^7}$
t_2	$4 + \frac{1}{5} \cdot \log \frac{108}{5^5}$	$\varepsilon_2 = 0$	$4 + \frac{1}{5} \cdot \log \frac{108}{5^5}$
t_3	$4 + \frac{1}{6} \cdot \log \frac{36}{6^5}$	$\varepsilon_2 = 0$	$4 + \frac{1}{6} \cdot \log \frac{36}{6^5}$
t_4	$4 + \frac{1}{3} \cdot \log \frac{4}{3^3}$	$3 + \frac{1}{3} \cdot \log \frac{4}{3^3}$	1
t_5	$\varepsilon_1 = 0$	$3 + \frac{1}{7} \cdot \log \frac{432}{7^7}$	$\varepsilon_1 = 0$
t_6	0	$3 + \frac{1}{7} \cdot \log \frac{432}{7^7}$	$-(3 + \frac{1}{7} \cdot \log \frac{432}{7^7})$

2) Terms Assigned a Negative TSI Value

To solve the second problem that $\Delta tsi_{(X:\bar{X})}(t) < 0$ may hold for some $t \in V_X \cap V_{\bar{X}}$, for each $t \in V$ ($\supseteq V_X \cap V_{\bar{X}}$), we need to further redefine $\Delta tsi'_{(X:\bar{X})}(t)$ given in Eq.(14) to:

$$\Delta tsi^*_{(X:\bar{X})}(t) = \begin{cases} \left(tsi'_X(t) - \varepsilon_2\right) + \tau_4 & t \in V_X - V_{\bar{X}} \\ \Delta tsi'_{(X:\bar{X})}(t) + \tau_3 & t \in V_X \cap V_{\bar{X}}, \Delta tsi'_{(X:\bar{X})}(t) > 0 \\ \tau_2 & t \in V_X \cap V_{\bar{X}}, \Delta tsi'_{(X:\bar{X})}(t) = 0 \\ \tau_1 & t \in V_X \cap V_{\bar{X}}, \Delta tsi'_{(X:\bar{X})}(t) < 0 \\ \varepsilon_1 & t \in V - V_X \end{cases}$$

where $0 \leq \varepsilon_1 < \tau_1 < \tau_2 \leq \tau_3 \leq \tau_4$ are called *modifying* parameters (e.g., $\tau_1 = 0.5$, $\tau_2 = 1.0$, $\tau_3 = 1.5$, $\tau_4 = 2.0$, $\varepsilon_1 = \varepsilon_2 = 0$ were used in our experiments).

Clearly, $\Delta tsi^*_{(X:\bar{X})}(t) \geq 0$ for all $t \in V$. The basic idea of taking the above modifying parameters is simple. First, we assign τ_1 to those terms having negative weight and τ_2 to those terms having zero weight; the reason $\tau_2 > \tau_1$ is because we believe that terms having negative weight are worse than terms having zero weight. To avoid losing the importance of terms representing d caused by adding τ_1 and τ_2, τ_3 and τ_4 are also added to those terms having positive weight; the reason $\tau_4 > \tau_3$ is because we regard terms occurring in X alone as being more important in representing $d \in X$ than terms occurring in both X and $\bar{X}$. Finally, $\varepsilon_1 < \tau_1$ for those terms occurring in only $V_{\bar{X}}$.

3) Terms with a Zero Posterior Probability

To solve the third problem that $p(t|X) = 0$ if $\varpi_X(t) = 0$, a smoothing method may be required to assign a non-zero probability mass to those terms with $\varpi_X(t) = 0$. For instance, with the *additive smoothing* method,

$$\varpi'_X(t) = \varpi_X(t) + \theta$$

where $\theta > 0$ is a smoothing parameter (for instance, $\theta = 0.5$ was used in our experiments), the posterior probability can be rewritten by

$$\hat{p}(t|X) = \frac{\varpi'_X(t)}{\Psi'} = \frac{\varpi_X(t)}{\Psi'} + \frac{\theta}{\Psi'}$$

and Ψ' is a *normalization factor* after smoothing:

$$\Psi' = \sum_{t \in V} \varpi'_X(t) = \Psi + \theta \cdot |V|$$

where, according to Eq.(9),

$$\Psi = \sum_{t \in V_X} \varpi_X(t) = \sum_{d \in X} \sum_{t \in V_X} \pi(d) \cdot w_{d|X}(t)$$

is a normalization factor before smoothing. Thus, all the terms with $\varpi_X(t) = 0$ are then assigned to an equal non-zero probability mass $\frac{\theta}{\Psi'}$.

4) Alternative

An alternative way, which can solve both the second and third problems together and may thus be the simplest one, is:

$$\varpi^*_X(t) = \begin{cases} \varpi_X(t) + \theta_1 & \varpi_X(t) > 0 \\ \theta_2 & \varpi_X(t) = 0 \\ \theta_3 & \varpi_X(t) < 0 \end{cases}$$

where $\theta_1 \geq \theta_2 \geq \theta_3 > 0$ are smoothing parameters. Clearly, $\varpi'_X(t)$ adds an equal value θ to all terms regardless of whether $\varpi_X(t)$ is zero or negative or not. That is, $\varpi'_X(t) = \varpi^*_X(t)$ when $\theta_1 = \theta_2 = \theta_3 = \theta$ and, therefore, it is a special case of $\varpi^*_X(t)$.

V. Experiments

This section presents some results from three sets of experiments carried out in order to verify SC effectiveness of our methods. As this study focuses on introducing a general form of a specificity measure and clarifying some potential problems of applications and suggesting solutions, rather than an extensive experimental investigation into the measure, the readers interested in empirical evidence drawn from a number of performance experiments and comparisons are referred to those papers referenced.

Our experiments used a collection from the movie review domain [19], first used in [12] and widely used in SC research. There are 2000 labelled documents in the full collection, consisting of 1000 positive and 1000 negative documents. Before using our formulae, we removed stop words and very high frequency terms (occurring in more than 60% of documents), and used a stemming algorithm [24]. We disregarded the position of terms in documents. Each document was treated as a 'bag-of-words'. Only term frequencies were considered. In our experiments, 10-fold cross-validation and the standard measures *recall* and *precision* were used for evaluation.

The first set of experiments compared the performance obtained from eleven weighting functions: eight are listed in Table I (in Example 3.1) and another three below were used as benchmarks:

$$w_d^{(F)}(t) = f_d(t)$$

$$w_d^{(O)}(t) = \frac{(a+1) \cdot f_d(t)}{a \cdot \left[(1-b) + b \cdot \beta(d,C)\right] + f_d(t)}$$

$$w_d^{(S)}(t) = \frac{\left[1 + \ln\left(1 + \ln(f_d(t))\right)\right] \cdot \log\left(\frac{|C|+1}{L_d}\right)}{(1-c) + c \cdot \beta(d,C)}$$

where parameters $a = 1.2$, $b = 0.75$ and $c = 0.2$, and $\beta(d,C)$ is given in Eq.(15) (see the last set of experiments below). Past experimental studies emphasised that a weighting function using just term frequency information can produce good performance for SC [1], and the Okapi (BM25) [25] and Smart [26] weighting functions have widely been recognized to produce excellent retrieval performances in IR. Table VI displays our experimental results using the eleven weighting functions, and the best results are given in square brackets in bold face.

From the results in Table VI it can be seen: Classifications obtained from (i) all the eleven weighting functions achieved good performance (above 90% recall/precision) at most evaluation points; (ii) idf, tf·idf, int, tf·int achieved consistently better performance than from tf, Okapi and Smart functions; the improvements were shown at all the evaluation points, which verifies *TSI*s are appropriate measures for SC; (iii) △idf, tf·△idf, △int and tf·△int showed a bias towards C_P,

which resulted in a relatively low precision for C_P but a very high precision for C_N; the cause of the bias is an interesting question, and extensive experiments may need to be carried out to train the parameters of $\Delta tsi^*_{(X:\bar{X})}(t)$; (iv) int, tf·int, △int and tf·△int were consistently better than from idf, tf·idf, △idf and tf·△idf, respectively; the improvements were shown at all the evaluation points (the reason for the improvements was explained at the beginning of Section 2.2); (v) tf·idf, tf·△idf, tf·int and tf·△int seem not to achieve the anticipated performance improvements compared with from idf, △idf, int and △int, respectively; this indicates that term specificity information may dominate the classifier performance. In addition, our experimental results bear out past experimental studies that tf can produce good performance for SC.

TABLE VI
PERFORMANCE WITH 11 WEIGHTING FUNCTIONS

	Negative Class C_N		Positive Class C_P	
$w_d(t)$	recall	precision	recall	precision
tf	0.9280	0.9460	0.9470	0.9297
Okapi	0.9350	0.9482	0.9490	0.9363
Smart	0.9290	0.9411	0.9420	0.9303
$w_{d\|X}(t)$	recall	precision	recall	precision
idf	0.9420	0.9684	0.9690	0.9433
tf·idf	0.9350	0.9672	0.9680	0.9367
int	**[0.9560]**	**[0.9747]**	**[0.9750]**	**[0.9566]**
tf·int	0.9440	0.9705	0.9710	0.9452
△idf	0.8790	0.9921	0.9949	0.8927
tf·△idf	0.8420	0.9903	0.9900	0.8618
△int	0.8860	0.9988	0.9990	0.9190
tf·△int	0.8630	0.9947	0.9978	0.8954

The second set of experiments considered the issue of dimension reduction of term space. Dimension reduction is an important issue in document classification, IR, NLP, and many related areas. It is generally the process of reducing the number of random variables under consideration. In our case, it is the process of identification of *informative* terms and, then, documents are represented by all the identified terms. The identified informative terms pertaining to the positive (or, negative) class are regarded as positive (or, negative) *sentiment-bearing* terms. The directed divergence measure [27] was used for the identification:

$$I\big(P_{x}(t);P_{c}(t)\big)=P_{x}(t)\log\frac{P_{x}(t)}{P_{c}(t)}$$

in which, $P_{x}(t)=P(t|X)$ (where $t\in V_X$) may be estimated using expressions given in Eq.(9), Eq.(10) and Eq.(12). Dimension reduction enables sentiment analysis to be performed in the reduced space more accurately and reliably than in the original space. A detailed discussion on informative term identification can be found in [28].

We experimentally studied classification performance using the identified informative terms to represent documents. There were 25259 distinct terms in V after stop word removal. The top δ terms of a ranked list were selected as the informative terms. We iteratively evaluated the eleven weighting functions using the δ terms, with $\delta=14000$ to $\delta=4000$ stepping -2000. The best results with $\delta=10000$ are given in Table VII.

From the results in Table VII it can be seen: Classifications obtained from (i) all the eleven weighting functions achieved consistently good performance at all the evaluation points; (ii) idf and tf·idf showed better performance than using tf, Okapi or Smart at most evaluation points; (iii) tf, Okapi and Smart showed better performance compared with the corresponding performance without using the informative terms at most evaluation points (see Table VI). In addition, our experimental results (not given in this paper) showed that if the number of identified terms is reduced to less than 40% of the original size of the vocabulary, it would not be possible to improve classification performance.

TABLE VII
PERFORMANCE USING 10000 INFORMATIVE TERMS

	Negative Class C_N		Positive Class C_P	
$w_d(t)$	recall	precision	recall	precision
tf	0.9340	0.9459	0.9460	0.9351
Okapi	0.9420	0.9529	0.9530	0.9430
Smart	0.9420	0.9536	0.9540	0.9435
$w_{d\|X}(t)$	recall	precision	recall	precision
idf	**[0.9550]**	0.9539	0.9530	**[0.9641]**
tf·idf	0.9520	**[0.9553]**	**[0.9550]**	0.9517
int	0.9960	0.9233	0.9010	0.9952
tf·int	0.9910	0.9300	0.9170	0.9904
△idf	0.9210	0.9472	0.9490	0.9243
tf·△idf	0.9150	0.9488	0.9510	0.9192
△int	0.9240	0.9455	0.9470	0.9268
tf·△int	0.9210	0.9472	0.9490	0.9243

The last set of experiments involved the construction of the normalization factor, denoted by $\psi(d,X)$, according to the individual documents. There are many ways to construct ψ. One way is to consider a linear combination (with a parameter $\lambda>0$):

$$\psi(d,X)=(1-\lambda)+\lambda\cdot\beta(d,X) \tag{15}$$

where $\beta(d,X)=\frac{L_d}{ave(X)}$ is a *length moderation factor* and $ave(X)=\frac{1}{|X|}\sum_{d\in X}L_d$ is the average length of all $d\in X$. The $\beta(d,X)$ may be used to further moderate the effect of the document length across the individual classes and thus be used to construct ψ. The $\psi(d,X)$ given in Eq.(16) is used in both Okapi and Smart weighting functions.

It is thus interesting to test the usefulness of a combination with the form:

$$w^*_{d|X}(t)=\frac{w_{d|X}(t)}{\psi(d,X)}$$

where $w_{d|X}(t)$ is one of the eight weighting functions listed in Table 1, and λ is set to 0.2, 0.5, 0.8 and 1.0.

The results (not given in this paper) showed that $w^*_{d|X}(t)$ did not provide performance improvement compared with $w_{d|X}(t)$ itself whether using the informative terms or not. The reason for the worse performance is not yet clear. However, we conjecture that it may be because tf, Okapi and Smart are basically term frequency-based weighting functions, and are therefore sensitive to the document length normalization. In contrast, our methods provide term specificity-based weighting functions, thus a skewed document frequency distribution over a class plays a key role in determining SC performance.

Note that the normalization factor Ψ given in Section 4.2 serves for $\varpi_X(t)$, whereas the normalization factor ψ here serves for $w_{d|X}(t)$. That is, the former is used for the weighting function regarding classes, the latter is used for the weighting function according to the individual documents.

VI. Conclusions

This study has advocated the use of *TSI* to assess document sentiment orientation. We discussed the mathematical concept of specificity information conveyed by a given term based on Shannon's entropy, and then introduced a general form of a specificity measure in terms of the concept. Two well-known specificity measures were considered, as examples, to illustrate the general form and their relationship was established based on the general form. We introduced an intuitive concept on specificity strength of terms over the classification and, then proposed a general method to represent the statistical importance of terms pertaining to individual documents with estimation of posterior probabilities using term weights obtained from *TSI* for the NB classifier. We clarified some potential problems inherent in applying the *TSI* measures in a Bayesian learning framework and, then suggest solutions that are easy to apply in practice. We proposed a novel multiple representation method, where each term is assigned multiple weights against individual sentiment classes, and explored a method of applying existing advanced single representation IR techniques to SC. We presented some experimental results and showed that the proposed method outperforms existing advanced IR single representation techniques. We attributed this to the capacity of the proposed method to capture aspects of term behaviour beyond a single representation. Our experimental results also verified that *TSI* may be regarded as an appropriate measure for effective SC. In ongoing work we are exploring reasons why using specificity information may result in a classification bias. Due to its generality, our method can be expected to be a useful tool for a variety of tasks of document classification, IR, NLP, and many related areas.

References

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, 2002, pp. 79–86.

[2] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 2002, pp. 417–424.

[3] B. Snyder and R. Barzilay, "Multiple aspect ranking using the Good Grief algorithm," in *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, 2007, pp. 300–307.

[4] T. Wilson, J. Wiebe, and R. Hwa, "Just how mad are you? Finding strong and weak opinion clauses," in *Proceedings of the 21st Conference of the American Association for Artificial Intelligence (AAAI'04)*, 2004.

[5] M. Thelwall, K. Buckley, G. Paltoglou, C. D., and A. Kappas, "An information theoretic foundation for the measurement of discrimination information," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[6] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, pp. 121–144, 2008.

[7] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 1, no. 1-2, pp. 1–135, 2008.

[8] A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: A case study," in *Proceedings of RANLP'05*, 2005.

[9] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational Intelligence*, vol. 22, no. 2, pp. 110–125, 2006.

[10] S. Matsumoto, H. Takamura, and M. Okumura, "Sentiment classification using word sub-sequences and dependency sub-trees," in *Proceedings of PAKDD'05*, 2005, pp. 301–311.

[11] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *Proceedings of EMNLP'04*, 2004, pp. 412–418.

[12] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of 42nd ACL*, 2004, pp. 271–278.

[13] J. Martineau and T. Finin, "Delta tfidf: An improved feature space for sentiment analysis," in *Proceedings of the Third AAAI International Conference on Weblogs and Social Media*, 2009.

[14] J. Martineau, T. Finin, A. Joshi, and S. Patel, "Improving binary classification on text problems using differential word features," in *Proceedings of 18th ACM Conference on Information and Knowledge Management (CIKM'09)*, 2009, pp. 2019–2023.

[15] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[16] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 2001.

[17] H. Zhang, "The optimality of naive Bayes," in *The 17th International FLAIRS Conference*, 2004.

[18] S. Caraballo and E. Charniak, "Determining the specificity of nouns from text," in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, pp. 63–70.

[19] "Movie Review Data: http://www.cs.cornell.edu/people/pabo/movie-review-data/ (2004)."

[20] K. Sparck Jones, "A statistical interpretation of term specificity and its application to retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[21] G. Salton and M. H. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[22] A. Akobeng, "Understanding diagnostic tests 1: sensitivity, specificity and predictive values," *Foundation Acta Pædiatrica/Acta Pædiatrica*, vol. 96, pp. 338–341, 2006.

[23] D. Cai, "Determining semantic relatedness through the measurement of discrimination information using Jensen difference," *International Journal of Intelligent Systems*, vol. 24, no. 5, pp. 477–503, 2009.

[24] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[25] S. E. Robertson and S. Walker, "Okapi/Keenbow at TREC-8," in *The 8th Text REtrieval Conference (TREC-8)*. NIST Special Publication, 1999, pp. 151–161.

[26] A. Singhal, J. Choi, D. Hindle, D. Lewis, and F. Pereira, "AT&T at TREC-7," in *The 7th Text REtrieval Conference (TREC-7)*. NIST Special Publication, 1999, pp. 239–252.

[27] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[28] D. Cai, "An information theoretic foundation for the measurement of discrimination information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 9, pp. 1262–1273, 2010.