Sentence-Level Opinion Analysis for Chinese News Documents Based on Sentiment Information of Social Tags

Jen-Yuan Yeh[1], Shihn-Yuarn Chen[2]
[1]Dept. of Operation, Visitor Service
Collection and Information Management
National Museum of Natural Science
Taichung 40453, Taiwan
[2]OneLab Technology Ltd.
Taipei City 11494, Taiwan
jenyuan@mail.nmns.edu.tw, phanix@gmail.com

Journal of Digital Information Management

***ABSTRACT**: Social tags have been considered to indirectly reflect authorized opinions of taggers. This paper proposes an unsupervised method which derives implicit sentiment information from social tags to decide, in one document, which sentences are opinionated, as well as to annotate them with proper polarity labels. First, for a social tag, its opinion degree is measured by aggregating the opinion degree of related sentiment words, in proportion to the co-occurrence relations between sentiment words and the tag. Second, the opinion degree of a sentence is determined by a combination function of the opinion degree of the tags, in proportion to the similarity between the sentence and each tag. Finally, sentences are sorted in order of their opinion degree, followed by a partition of the ranked list to distinguish sentences into positively opinionated, negatively opinionated, neutral, and non-opinionated ones. The proposed method is examined using the Chinese dataset of the NTCIR Opinion Analysis Task Test Collection and found to perform well. Experimental results testify that social tags are positively conducive to opinion analysis.*

Subject Categories and Descriptors
I.2.7 [Artificial Intelligence]: Natural Language Processing – Text Analysis

General Terms: Algorithms, Experimentation

Keywords: Opinion Mining, Opinion Sentence Extraction, Opinion Polarity Labeling, Social Tags, Unsupervised Method

Received: 1 November 2015, Revised: 27 December 2015, Accepted: 30 December 2015

1. Introduction

Opinion mining[1] deals with the computational treatments of opinions, sentiments, and subjectivity in text [4], [28], [31], [33], [36]. It addresses the problems of (1) identifying passages (e.g., a word/phrase, a sentence, or even a document) conveying attitudinal information, such as attitudes, opinions, appraisals, evaluations, sentiments, and emotions, which the holder expresses towards a given subject; and (2) determining the semantic orientation of these passages. Prior attempts focused mostly on interpretation of belief, metaphor, narrative, point of view, emotion, affect, and evidentiality in text [5], [45]. Since 2001, the field has evolved dramatically following a huge burst of production of opinion texts (i.e., texts containing opinions or sentiments [28]) on the Internet. Among the many noteworthy developments, *sentiment* and *subjectivity classification* are perhaps the two most well-studied. The latter has come to be known as the discrimination between objective and subjective utterances [17], [24], [37], [44], [46], [50], while the former aims to judge the polarity (e.g., positive, negative, and neutral) of an opinion text [16], [26], [35], [42], [47], [50].

Social tagging, referred to as assigning keywords (i.e., tags) to information resources [7], encourages participants to play roles as *tag creators* for contributing keywords for taxonomy, collaborative, and social purposes,

[1] The terms (in alphabetical order) *opinion mining*, *sentiment analysis*, and/or *subjectivity analysis* reflect differences in the connotations but roughly denote the same field of study in the technical literature [33]

and as *tag consumers* for knowing what other people are interested in and how they characterize the same resources. Most studies have utilized tag data to improve performance of website navigation, Web search, and social connection [3], [18]. Researchers have analyzed users' motivations towards tagging as well, and suggest that social tags implicitly convey sentiments that taggers express [41], [51]. However, little research attempts to mine sentiment information from social tags, which motivates this study to investigate ways of measuring the sentiment orientation of social tags and to testify the effects of social tags to opinion analysis.

This paper proposes a novel unsupervised method, OSEPLST (Opinion Sentence Extraction and Polarity Labeling based on Social Tags), to address *sentence-level* opinion extraction and polarity labeling for Chinese news documents. The method attempts to extract implicit sentiment information from social tags and utilizes this kind of information, to decide, in one document, which sentences are opinionated, as well as to annotate them with proper polarity labels. It adopts a broadly used opinion analysis paradigm – *the opinion of the whole is a function of the opinions of the parts* [25]. To be specific, sentiment words are the smallest units to be detected, and their opinion degree and polarities are used to evaluate the tendency of a sentence, and then a document.

The proposed method can be decomposed into three subtasks. The first subtask identifies sentiment words and assigns a graded sentiment value to each sentiment word. For a social tag, its opinion degree is measured by aggregating the opinion degree of related sentiment words, in proportion to the relations between sentiment words and the tag. In the second subtask, the opinion degree of a sentence is determined by a combination function of the opinion degree of the tags, in proportion to the similarity between the sentence and each tag. Finally, in the third subtask, sentences are sorted in order of their opinion degree, followed by a partition of the ranked list to distinguish sentences into positively opinionated, negatively opinionated, neutral, and non-opinionated ones.

The main contributions of this work are twofold. First, it offers a method of deriving implicit sentiment information from social tags, which past researches had made little attempt to do so. Second, based on implicit sentiment information of social tags, an unsupervised sentence-level opinion analysis method is proposed to extract opinion sentences and to determine their polarity labels. The proposed method could be practically beneficial to many information processing and management applications. For example, sentiment lexicons can be automatically constructed or augmented with the help of sentiment analysis of social tags. By sentence-level opinion extraction and polarity labeling, the effectiveness of opinion-oriented applications, including opinion retrieval, opinion question answering and opinion summarization, can be enhanced.

In the following, Section 2 provides a brief review of related work. Section 3 introduces in detail the proposed method.

The experimental results are presented and discussed in Section 4. Finally, Section 5 concludes this paper and outlines possible directions for future work.

2. Related Work

The field of opinion mining has widely investigated sentiment and subjectivity classification. Sentiment classification evaluates the sentiment orientation of an opinion text at the scales of *words* or *phrases* [2], [16], [22], [47], *sentences* [23], [26], [44], [50], and *documents* [13], [27], [35], [42]. Some examples follow. [16] proposed constraints on the semantic orientation of conjoined adjectives to predict whether two adjectives are of the same or different orientation. [22] constructed measures, using Osgood's semantic differential technique, for affective or emotive aspects of meaning derived from the structure of the WordNet.[2] [23] measured the strength of sentiment polarity of a word and built strategies to assign a sentiment category to a sentence. [26] considered together sentiment words, opinion operators, opinion holders, and negation operators, for Chinese opinion sentence extraction. [42] classified a review as recommended or not recommended according to the average semantic orientation of the phrases in it. [35] examined naïve Bayes classifiers, maximum entropy classification, and support vector machines for sentiment classification based on the bag-of-words document representation model.

A more general form of sentiment classification concerns the prediction of the positivity degree (e.g., 1–5 stars) of the subjective text [19], [34]. Usually, it is regarded as a multi-class text categorization problem. In addition, a fine-grained analysis model, referred to as feature-level sentiment classification, has also been discussed. The model aims to discover the features of entities that have been commented on and to determine the orientation of the opinions [12], [15], [20], [32]. Here, a feature stands for a component or attribute of an object, e.g., the battery life of a cellular phone.

Subjectivity classification, which positively influences sentiment classification [30], concentrates on discriminating objective instances from subjective ones [8], [17], [24], [37], [46], [50]. For instance, [17] studied the effects of dynamic, gradable, and semantically oriented adjectives on sentence subjectivity. [50] trained a naïve Bayes classifier to distinguish between documents with preponderance of opinions and suggested a model for classifying opinion sentences as positive or negative in terms of the main perspective being expressed. [46] employed naïve Bayes to recognize sentence subjectiv

[2] http://wordnet.princeton.edu/.

ity and developed a classifier using only unannotated texts for training. [37] established a bootstrapping process that learns linguistically rich extraction patterns for subjective expressions. See [44] for a comprehensive survey on subjectivity classification.

Two techniques have been widely exploited for opinion mining, namely, *supervised opinion classification* [8], [35], [37], [46], [47], [50] and *unsupervised opinion extraction* [13], [17], [26], [42]. The goal of supervised opinion classification is to train an opinion-oriented classifier by machine learning. Typically used features include *term presence and frequency*, *part-of-speech information*, *syntactic relations*, *negation*, and *opinion words* [10], [28], [33]. Unsupervised opinion extraction identifies opinion indicators, followed by assessing the polarity of a text with a scoring function that combines the strength of polarity of the indicators. Generally, supervised classification provides more accurate results, although the performance is highly dependent on the applied domain. Unsupervised extraction, in contrast, requiresno prior training. The selection of opinion mining techniques tends to be a trade-off between accuracy and generality [6].

Analysis of public information from E-commerce and social media has recently received increasing attention (see [1], [9], [14], [39], [43], [48]). For example, [39] trained a set of multi-class classifiers to distinguish between video and product related opinions. In their work, standard feature vectors are augmented by shallow syntactic trees. [9] enhanced polarity classification by leveraging on linguistic rules and sentic computing resources. [1] discussed a methodology by which it is possible to determine the populatity/opinion/sentiment of a product in different locations across male and female users. [43] introduced two hybrid ensemble based models for opinion classification of product reviews.

This work, as mentioned previously, adopts a broadly used opinion analysis paradigm, i.e., the opinion of the whole is a function of the opinions of the parts, which makes it resemble most sentence-level *unsupervised* opinion extraction technologies. However, this work is quite distinct from the others due to the underlying algorithm for determining the sentiment tendency of a sentence based on sentiment information of social tags and the novel partition strategy for classifying sentences into positively opinionated, negatively opinionated, neutral, and non-opinionated ones. In addition, the proposed method in this work is domain-independent since it considers neither domain-specific knowledge nor deep linguistic analysis of texts, which is required in most related works. Finally, as far as we know, previous studies have not yet investigate the use of social tags for opinion analysis, which distinguishes this work from the literature.

3. Sentence-level Opinion Analysis Based on Sentimental Information of Social Tags

3.1 Overview of the Proposed Method

Figure 1 illustrates an overview of the proposed OSEPLST method. The input is a group of topic-related Chinese news documents3 with social tags. The output is a set of opinion sentences, with polarity labels if requested. The opinion analysis process is composed of three phases: (1) *preprocessing* preprocesses the input documents and social tags; (2) *opinion analysis* accesses the opinion degree of social tags and of sentences; and (3) *opinion extraction* classifies sentences into opinionated or non-opinionated, and annotates opinion sentences with proper polarity labels.

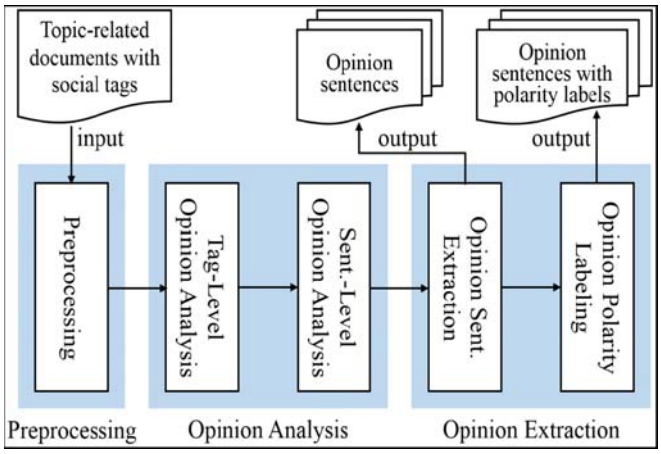

Figure 1. Overview of the proposed OSEPLST method

The entire process can be further divided into several stages:

(**1) Preprocessing:**[4] A tokenizer segments text into words, numbers, and symbols. A sentence splitter identifies the boundaries of sentences. A passage indexer constructs a vector representation for every sentence. Further, tag normalization is performed to group synonymous tags together. For example, 老虎伍茲 (*Tiger Woods*)", "伍茲 (Woods)," and "艾德瑞克·伍茲(*Eldrick Woods*)" are normalized to the full name of the famous golfer, "艾德瑞克·「老虎」·伍茲 (*Eldrick Tiger Woods*)."

(2) Tag-level opinion analysis (see Section 3.2): Sentiment words in the input documents are detected with a sentiment dictionary. For a social tag, its opinion degree is measured as the summation of the opinion degree of related sentiment words, in proportion to the co-occurrence relations between sentiment words and the tag.

[3] The input can be either topic-related news documents (e.g., news documents describing the same event) or one single document. Since the evaluation data (see Section 4.1) comes in form of topic clusters of documents, the input in Figure 1 are topic-related documents.

[4] The CKIP Chinese Word Segmentation System (see http://ckipsvr.iis.sinica.edu.tw/) is employed as the tokenizer. A heuristics is exploited to delimit each sentence by stop marks, such as "。", "！", "；", and "？".

(3) Sentence-level opinion analysis (see Section 3.3): Social tags and sentences are modeled as vectors of index words, where the cosine similarity between a social tag and a sentence is evaluated. The opinion degree of a sentence is determined by a combination function of the opinion degree of the tags that annotate the document which the sentence belongs to, in proportion to the similarity between the sentence and each tag.

(4) Opinion sentence extraction (see Section 3.4): Sentences are sorted in order of their opinion degree, followed by a partition of the ranked list to distinguish opinion and non-opinion sentences. The partition is according to the heuristics that if a sentence has positive opinion degree, it is likely to be positively opinionated; if a sentence has negative opinion degree, it is likely to be negatively opinionated; and a sentence with opinion degree close to 0 is possibly neutral or non-opinionated.

(5) Opinion polarity labeling (see Section 3.4): Extracted opinion sentences are, again, sorted in order of their opinion degree. A partition of the ranked list is determined so that the sentences lying at two extreme regions are classified as positive or negative and the sentences in the mid-part are identified as neutral.

3.2. Tag-Level Opinion Analysis

This study denotes the set of social tags as $\{T\}$ and the set of sentiment terms as $\{t\}$. Given a tag T, its opinion degree, D_T, is aggregated from the opinion degree of related sentiment terms. Here, a sentiment term is recognized as an instance in the National Taiwan University Sentiment Dictionary (NTUSD) [25], which composes of 2,812 positive and 8,276 negative sentiment Chinese words.

The related sentiment terms of a tag T are defined as sentiment terms that have *high* frequencies of co-occurrence with T within a window of text in a data source. Two types of data sources are considered. The first one is a document corpus, e.g., news articles collected in a particular period, and the related sentiment terms are those *strongly* co-occurring with T in the corpus. It probably suffers from one problem that the corpus is not large enough to reflect the real-world distribution of co-occurrence of T and a sentiment term t. In addition, T might only be semantically related to tagged document(s), but does not appear in the corpus.[5] In such a case, the frequency of co-occurrence of T and t becomes zero, implying that the value of D_T will be misestimated. To alleviate this issue, the whole Web is employed as the second type of data source. The related sentiment terms are, hence, those *strongly* co-occurring with T on the Internet. Notably, the advantages of considering webpages on the Internet as a data source are twofold: (1) it reflects much better the real-world distribution of co-occurrence of T and t; and (2) D_T can still be calculated even when T does not appear in the document corpus.

The first type of opinion degree of a tag T, $DEG_{cor,}T$, is named "in-corpus opinion degree of T" and the second one, $DEG_{int,}T$, is named "*on-Internet opinion degree of* T". $DEG_{cor,}T$ takes account of the frequency of co-occurrence of T and a sentiment term t and their frequencies of occurrence, both in a window of sentence, in the document corpus. In contrast, $DEG_{int,}T$ considers the frequency of co-occurrence of T and t and their frequencies of occurrence, both in a window of webpage, on the Internet. Currently, the frequencies of occurrence and co-occurrence on the Internet are estimated as the counts of webpages indexed in the Yahoo! search engine.[6]

It is recalled that a tag T might not appear in the document corpus. If only $DEG_{cor,T}$ is taken into account, the value of D_T will be zero. For a better assessment of DT, this study linearly combines $DEG_{cor,T}$ and $DEG_{int,T}$ with a weight α ($0 \le \alpha \le 1$):

$$D_T = \alpha \times DEG_{cor,T} + (1-\alpha) \times DEG_{int,T} \qquad (1)$$

Note that the weighted linear combination allows D_T to leverage the opinion degree of T, considered both in the document corpus and on the Internet.

The method of computing $DEG_{cor,T}$ (or $DEG_{int,T}$) is provided as follows:

$$DEG_T = \sum_t DEG(t) \times \log \frac{CO(T,t)}{O(T) \times O(t)}. \qquad (2)$$

$DEG(t)$ stands for the opinion degree of a sentiment term t. Since the released NTUSD gives no graded sentiment value to every word in the dictionary, $DEG(t)$ is simply assumed to be +1.0 for a positive sentiment term and -1.0 for a negative sentiment term. $O(T)$ and $O(t)$ are, respectively, the frequencies of occurrence of T and t. $CO\ (T, t)$ represents the frequency of co-occurrence of T and t.

It is worth noting that the opinion degree of a tag T is accumulated by the opinion degree of each sentiment term t, i.e., $DEG\ (t)$, in proportion to the co-occurrence relation between T and t. In Equation. (2), *pointwise mutual information* [11] is adopted to evaluate the relation between T and t.

3.3. Sentence-Level Opinion Analysis

Sentences and tags are represented using the bag-of-words model as vectors. Let $W\ (|W| = n)$ denote the set of terms, $W = \{t_1, t_2, \ldots, t_n\}$. The vector of a sentence S is specified by Equation. (3), in which w_i is the frequency of occurrence of term t_i in S:

[5] It practically happens because people prefer to tag a document using his/her own words based on his/her awareness of the document.

[6] The Yahoo Boss Search API, available at https://developer.yahoo.com/boss/, is used.

$$\vec{S} = < w_1, w_2, ..., w_n > \quad (3)$$

Similarly, a document d_j is denoted as:

$$\vec{d}_j = < w_{1j}, w_{2j}, ..., w_{nj} > \quad (4)$$

where w_{ij} is the frequency of occurrence of term t_i in d_j.

Since a tag T is used to annotate documents, T can be viewed as the average vector (i.e., the centroid) of the documents annotated by T:

$$\vec{T} = \frac{1}{m}\sum_j^m \vec{d}_j \quad (5)$$

where d_j is a tagged document and m is the number of tagged documents.

Finally, DEG_S, the opinion degree of S, is calculated by:

$$DEG_S = \frac{1}{length(S)}\sum_T (D_T \times \frac{\vec{S}\cdot\vec{T}}{|\vec{S}|\times|\vec{T}|}) \quad (6)$$

where D_T, as indicated in Equation. (1), is the opinion degree of a tag T and $1/length(S)$ serves to normalize the summation. In brief, DEG_S is an aggregated value of the opinion degree of the tags that annotate the document which S belongs to, in proportion to the similarity relation between S and each tag T.

3.4. Opinion Sentence Extraction and Polarity Labeling

Sentences are sorted in decreasing order according to their opinion degree. This study hypothesizes that sentences with large absolute values of opinion degree tend to be opinionated, i.e., most likely opinion sentences lie at two extreme regions of the ranked list. Figure 2 illustrates how opinion sentences are recognized. Since the left-most and the right-most sentences have, respectively, higher and lower opinion degree, they are regarded as opinionated. What falls in the mid-part of the list is the part of non-opinionated. The parameters ε and ϕ are ad justed for the best "*cuts*" that perfectly separate opinion and non-opinion sentences, as indicated by the rule:

(a) S is *opinionated* if it is located in [0%, ε%] or in [ϕ%, 100%]; and (7)

(b) S is *non-opinionated* if it is located in [ε%, ϕ%].

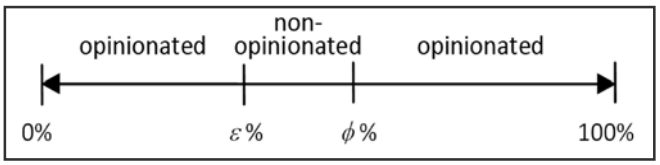

Figure. 2 A sentence list where cuts of ε and ϕ separate opinion and non-opinion sentences

Removing the non-opinionated part in Figure 2 forms a new ranked list, see Figure 3. The sentences in the left-hand side are labeled as positive since they have higher

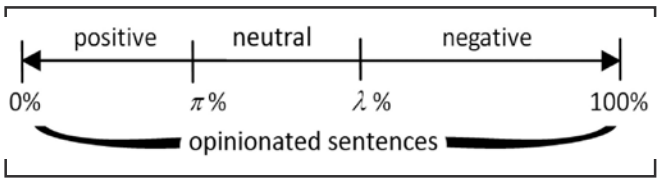

Figure 3. An opinionated sentence list where cuts of π and λ separate positively opinionated, negative opinionated, and neutral sentences

opinion degree; the sentences in the right-hand side are labeled as negative because they have lower opinion degree; and the sentences in the mid-part are labeled as neutral. The partition, as specified in Equation. (8), is decided by the control of parameters π and λ. In the current implementation, the parameters ε, ϕ, π, and λ are set empirically according to the results of preliminary experiments on small randomly built held-out datasets.

(a) S is *positively opinionated* if it is located in [0%, π%];
(b) S is *negatively opinionated* if it is located in [λ%, 100%]; and
(c) S is *neutral* if it is located in [π%, λ%]. (8)

4. Evaluation

4.1. The Dataset

The Chinese dataset of the NTCIR-6 Opinion Analysis Task Test Collection[7] is used. There are 843 annotated documents relevant to 32 topics. Documents are divided into 11, 907 sentences and three NTCIR annotators assigned opinion tags to sentences. An opinion tag indicates whether a sentence is opinionated or non-opinionated, the opinion holder of an opinion sentence, the relevance of a sentence to a topic, and the polarity of an opinion sentence. The polarity label can be positive (POS), negative (NEG) or neutral (NEU). The proposed method is evaluated topic by topic and the average of the evaluation results is reported.

Since the dataset has no social tags, three annotators were trained to annotate documents with tags. The tagging activity came out to have 2,427 distinct tags. The number of distinct tags for one topic ranges from 20 to 288 and the number of distinct tags for one document ranges from 2 to 19. In average, 94 distinct tags for one topic and 6.83 distinct tags for one document.

4.2. Keywords as Tags via Keyword Extraction

A social tagging system may provide automated tagging mechanism for annotating resources in a batch or for recommending tags to users. One simple automated tagging mechanism is keyword (or keyphrase) extraction, which is especially propitious when past tagging records or domain knowledge is unavailable. [29] recognized

[7] http://research.nii.ac.jp/ntcir/permission/ntcir-6/perm-en-OPINION.html.

relatively important keywords in one single document by comparing the distribution of co-occurrence of each term to the frequent terms using x^2 test (i.e., chi-square test).

The keyword extraction method is employed to identify significant keywords as tags for each document. Comparison of the effectiveness between keywords-as-tags and social tags for the proposed method is presented in Section 4.3.2.

The keyword extraction process produced 697 distinct keyword tags. The number of distinct keyword tags ranges from 6 to 58 for one topic and from 0 to 4 for one document. In average, 24.89 distinct keyword tags for one topic and 1.76 distinct keyword tags for one document.

4.3. Experimental Results

4.3.1. The NTCIR Opinion Analysis Evaluation

Four evaluation tasks are defined at the NTCIR-6, including **(1) Task 1:** to decide whether a sentence expresses an opinion; **(2) Task 2:** to identify the opinion holder of an opinion sentence; **(3) Task 3:** to judge whether a sentence is relevant to a particular topic; and **(4) Task 4:** to determine the polarity of an opinion sentence [38]. The guidelines of Task 1 and Task 4 are followed,considering the types of the outputs of the proposed method.The manual annotations are treated as the gold standard. Since sentences are annotated by three annotators, there are the *strict* and the *lenient* annotation sets. While in the previous all annotators must have the same annotation, in the latter at least two annotators have the same annotation. The annotations created by the proposed method are compared with the manual annotations and results for *precision*, *recall*, and *F-measure*, are presented.

$$precision = \frac{\#system\_correct}{\#system\_proposed} \quad (9)$$

$$recall = \frac{\#system\_correct}{\#sentences} \quad (10)$$

$$F-measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (11)$$

where *#system_correct* is the number of correct machine-annotated sentences, *#system_proposed* is the number of machine-annotated sentences, and *#sentences* is the number of human-annotated sentences.

Method	Category	Note
OSEPLST (this study)	Unsupervised	1. Determine the sentiment tendency of a sentence based on sentiment information of social tags and apply a partition strategy for classifying sentences into positively opinionated, negatively opinionated, neutral, and non-opinionated ones; 2. The sentiment dictionary, NTUSD [25], is used.
CUHK	Unsupervised	1. Build an opinion word lexicon based on knowledge acquired from web data; 2. Determine if a sentence is opinionated and its polarity based on heuristic linguistic features that form a scoring function.
Gate-1 / Gate-2	Supervised	1. Apply the binary classification SVM learning directly based on unigram linguistic features; 2. Gate-1 uses the lenient annotations for training, while Gate-2 uses the strict annotations for training.
ISCAS	Unsupervised	1. Consider a Chinese character as the atomic unit and adopt an empirical scoring function to discriminate the subjective sentences from the objective ones and suggest their polarities; 2. The sentiment dictionary, NTUSD [25], is used and is further expanded by a Chinese synonym thesaurus.
NTU	Unsupervised	1. Calculate the opinion degree of a sentence and its polarity based on sentiment words, negation words, opinion operators, and opinion holders; 2. The sentiment dictionary, NTUSD [25], is used.
UMCP-1 / UMCP-2	Unsupervised	1. Detect opinionated sentences if a sentence has at least one word appearing in the POS, NEG, or NEU sentiment lexicons; 2. The sentiment dictionary, NTUSD [25], and the Chinese Positive/Negative Dictionary [40][49] are used; 3. UMCP-1 uses the smaller lexicons, while UMCP-1 uses the bigger lexicons.

Table 1. Comparison between the proposed method, OSEPLST, and the other systems

Table 1 briefs the proposed method, OSEPLST, and the compared systems. Table 2 lists the results of Task 1 evaluation. *L*/*S* indicates which gold standard is compared; *L* for the lenient and *S* for the strict. *SYSID* signifies the peer codes of participants at the NTCIR-6. OSEPLST (ε, ϕ) denotes the proposed method with different cuts defined in Eq. (7). As an example, OSEPLST (50, 57.5) means that ε and ϕ are set, respectively, to 50% and 57.5%. The value of α in Eq. (1) is heuristically assigned to 0.8. In the lenient evaluation, the distinct models of the proposed method have good recalls but have relatively inferior precisions than the other systems, except ISCAS. The best model of the proposed method, i.e., OSEPLST (60, 65), has an F-measure of 0.744, which is higher than that of CUHK (with an increase of 17.17%), Gate-2 (with an increase of 12.90%), and ISCAS (with an increase of 19.04%). When being compared to the best system, i.e., UMCP-1, the proposed method, OSEPLST (60, 65), has a small decrease of 4.30%. Notably, the results also indicate that Gate-2, a supervised system, has inferior F-measure scores than the other unsupervised systems. In the strict evaluation, the distinct models of the proposed method have comparable results in terms of precision, recall, and F-measure. The best model of the proposed method, OSEPLST (50, 57.5), has an F-measure of 0.397, which is higher than that of ISCAS (with an increase of 19.94%), UMCP-1 (with an increase of 1.02%), and UMCP-2 (with an increase of 3.12%). But, OSEPLST (50, 57.5) is inferior to the best system, i.e., Gate-2, with a large decrease of 12.85%. Last, it is seen that a system with higher precision obtains better F-measure. However, the generally low precisions imply that substantial room for improvement remains.

The results of Task 4 evaluation are presented in Table 3. Gate-1 and Gate-2 have no results because they did not participate in Task 4. OSEPLST (ε, ϕ, π, λ) specifies the proposed method with different cuts defined in Eq. (7) and Eq. (8). For instance, OSEPLST (50, 57.5, 57.5, 62.5) indicates that ε, ϕ, π, and λ are set to 50%, 57.5%, 57.5%, and 62.5%, respectively. In the lenient evaluation, the distinct models of the proposed method perform comparably to the other methods. The best model of the proposed method is OSEPLST (60, 65, 62.5, 67.5), with an F-measure of 0.357, which is the third compared to the other systems (for CUHK, with a decrease of 13.45%; for ISCAS, with an increase of 31.09%; for NTU, with a decrease of 7.28%; for UMCP-1, with an increase of 1.68%; for UMCP-2, with an increase of 2.52%). Its precision, 0.295, is the third best and its recall, 0.450, is the best. Similarly, the distinct models of the proposed method have comparable results in the strict evaluation. The best model of the proposed method is OSEPLST (37.5, 60, 47.5, 52.5), whose F-measure is 0.172. Its precision, 0.106, is the second best among compared to other systems, although the recall of the model is the second worst. Again, the generally low precisions suggests the most important work is to improve the precision in the future.

Table 2 and Table 3 demonstrate that the proposed method performs well with comparable results. The results are not sufficiently satisfactory, compared to systems, e.g., UMCP-1 and Gate-2 in Task 1 evaluation and CUHK in Task 4 evaluation. A preliminary investigation found that, in the real world, co-occurrence of words may evolve during time; but, the on-Internet opinion degree of a social tag T, i.e., $DEG_{int,T}$, in Eq. (1) is measured without taking into account the time factor. For example, it is observed that the news articles tagged with "老虎伍兹 (Tiger Woods)" talk about the brilliant career time of Tiger Woods, so that the tag mostly co-occurs with positive sentiment terms in the dataset. Since his cheating scandal in 2009 and divorce in 2010, the word "老虎伍兹" tended to accompany with negative sentiment terms on the Internet. The rudimentary inspect suggests that, for a tag, the uncertainty caused by the time factor might lead to a large discrepancy between its in-corpus and on-Internet opinion degree and could debase the evaluation of tendency of sentences.

Another reason that causes the unsatisfactory results might be the small amount of the social tags. In the evaluation dataset, there are in average 94 distinct tags for one topic and 6.83 distinct tags for one document (see Section 4.1). This means that averagely only 6.83 social tags are considered for determining the opinion degree of a sentence (see Eq. (6)). Fortunately, this issue could be alleviated in real-world applications as increasingly many websites provide social tagging for their users, implying that a large amount of social tags can be collected and analyzed.

4.3.2 The Top k Opinion Sentence Evaluation

For sentiment-aware applications, e.g., topic-oriented opinion retrieval and opinion question answering, the correctness (or precision) regarding the top k results is usually emphasized. That is, increasing the number of correct opinion passages in the top k results is an essential issue, especially when k is small.

The information retrieval measure, NDCG (Normalized discounted cumulative gain) [21], is utilized with slight modifications to analyze the quality of the top k sentences.

$$NDCG_k = \frac{DCG_k}{IDCG_k}, \text{ where } DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (12)$$

*w*here k is the maximum number of sentences to be examined, *reli* is the graded rating of the *i*-th sentence in the list, and $IDCG_k$ is the maximum possible (ideal) DCG for the perfect list that receives an NDCG of 1.0. The value of *reli* is set to 1.0 while the opinion polarity of the *i*-th sentence is the same as the manual annotation. Since the proposed method returns the top k positive and the top k negative opinion sentences, i.e., a total of $2\times k$ sentences are extracted, the NDCG values for the two lists are assessed individually with the lenient gold standard and the average NDCG is reported.

SYSID	L / S	Precision	Recall	F-measure
OSEPLST (37.5, 60)	L	0.621	0.772	0.688
OSEPLST (50, 57.5)	L	0.619	0.919	0.740
OSEPLST (60, 65)	L	0.616	0.939	0.744
CUHK	L	0.818	0.519	0.635
Gate-1	L	0.643	0.933	0.762
Gate-2	L	0.746	0.591	0.659
ISCAS	L	0.590	0.664	0.625
NTU	L	0.664	0.890	0.761
UMCP-1	L	0.645	0.974	0.776
UMCP-2	L	0.630	0.984	0.768
OSEPLST (37.5, 60)	S	0.250	0.771	0.378
OSEPLST (50, 57.5)	S	0.252	0.927	0.397
OSEPLST (60, 65)	S	0.246	0.931	0.389
CUHK	S	0.341	0.575	0.428
Gate-1	S	0.253	0.979	0.402
Gate-2	S	0.330	0.696	0.448
ISCAS	S	0.221	0.662	0.331
NTU	S	0.258	0.921	0.404
UMC P-1	S	0.245	0.986	0.393
UMCP-2	S	0.239	0.993	0.385

Table 2. Results for precision, recall, and F-measure of the Task 1 evaluation

SYSID	L / S	Precision	Recall	F-measure
OSEPLST(37.5,60,47.5,52.5)	L	0.320	0.398	0.355
OSEPLST(50,57.5,57.5,62.5)	L	0.293	0.435	0.350
OSEPLST (60, 65, 62.5, 67.5)	L	0.295	0.450	0.357
CUHK	L	0.522	0.331	0.405
Gate-1	L	N/A	N/A	N/A
Gate-2	L	N/A	N/A	N/A
ISCAS	L	0.232	0.261	0.246
NTU	L	0.335	0.448	0.383
UMCP-1	L	0.292	0.441	0.351
UMCP-2	L	0.286	0.446	0.348
OSEPLST(37.5,60,47.5,52.5)	S	0.106	0.454	0.172
OSEPLST(50,57.5,57.5,62.5)	S	0.099	0.506	0.166
OSEPLST(60,65,62.5,67.5)	S	0.097	0.508	0.163
CUHK	S	0.197	0.596	0.296
Gate-1	S	N/A	N/A	N/A
Gate-2	S	N/A	N/A	N/A
ISCAS	S	0.059	0.314	0.099
NTU	S	0.104	0.662	0.180
UMCP-1	S	0.085	0.615	0.150
UMCP-2	S	0.081	0.604	0.143

Table 3. Results for precision, recall, and F-measure of the Task 4 evaluation

One simple baseline is developed. For a sentence S, he baseline determines whether it is a positive or negative opinion sentence by its score of opinion tendency:

$$Score_{baseline}(S) = \frac{\sum_{w \in S \,\&\, \mathrm{OP}(w)=\mathrm{POS}} 1.0 \times \mathrm{idf}_w + \sum_{w \in S \,\&\, \mathrm{OP}(w)=\mathrm{NEG}} -1.0 \times \mathrm{idf}_w}{length(S)} \quad (13)$$

where *length* (S) is the length of S, w represents a word in S, OP(w) indicates the opinion polarity of w, i.e., $OP(w)$ = POS means w is a positive sentiment word and otherwise w is a negative sentiment word, and idf_w stands for the inverse document frequency of w. The idea behind the baseline is if a sentence contains more positive sentiment words, it tends to be a positive opinion sentence and otherwise a negative opinion sentence.

Table 4 presents the NDCG results. Both OSEPLST-*Keyword* and OSEPLST-*Tag* adopt the proposed method but in different ways that the tags are generated. While the former assumes that manual tagging is unavailable and produces tags for documents via keyword extraction (see Section 4.2), the latter applies manual tags, i.e., social tags. The parameters, ε, ϕ, π, and λ, in Eq. (7) and Eq. (8) are not adjusted because only the top k sentences are extracted[8]. The value of α in Eq. (1) is heuristically assigned to 0.6 for OSEPLST-*Keyword* and 0.8 for OSEPLST-*Tag*. What given in parentheses is the relative improvement[9] of the methods in comparison to the baseline. The results show that OSEPLST-*Keyword* and OSEPLST-*Tag* can discover more opinion sentences with right polarity than the baseline. For OSEPLST-*Keyword*, the improvements are 1.56%, 1.54%, and 2.40%, respectively, when k is 10, 20, and 50; and for OSEPLST-*Tag*, the improvements become 10.94%, 7.87%, and 8.49%. As expected, OSEPLST-*Tag* is superior to OSEPLST-*Keyword* with improvements of 9.23%, 6.24%, and 5.95%. It demonstrates that the use of social tags is more effective since manual tags indirectly reflect authorized opinions of taggers [51]. Finally, the downward trend of improvements of OSEPLST-*Tag* over the baseline and over OSEPLST-*Keyword* indicates that OSEPLST-*Tag* works superiorly for a small k.

	$NDCG_{10}$ (k = 10)	$NDCG_{20}$ (k = 20)	$NDCG_{50}$ (k = 50)
Baseline	0.512	0.521	0.542
OSEPLST-*Keyword*	0.520 (+1.56%)	0.529 (+1.54%)	0.555 (+2.40%)
OSEPLST-*Tag*	**0.568** (+10.94%)	**0.562** (+7.87%)	**0.588** (+8.49%)

Table 4. Results for NDCG of the top k opinion sentence evaluation

5. Conclusion

This paper proposes an unsupervised method, OSEPLST (Opinion Sentence Extraction and Polarity Labeling based on Social Tags), towards sentence-level opinion analysis for Chinese news documents. The method attempts to derive implicit sentiment information from social tags and utilizes this kind of information to decide, in one document, which sentences are opinionated, as well as to annotate them with proper polarity labels. The method is examined using the Chinese dataset of the NTCIR Opinion Analysis Task Test Collection and is found to perform well with relatively good results. Experimental results also testify that social tags are positively conducive to opinion analysis.

Future work will continue to verify the effectiveness and the robustness of the proposed method using different datasets. It can be advantageous to the proposed method to build a larger sentiment lexicon dictionary, in which the sentiment orientation of a word is provided and quantitated. Another interest issue is to investigate the influence of the evolution of co-occurrence of words during time for better estimate of the co-occurrence relation between a tag and a sentiment word. Finally, using the output of the proposed method to develop sentiment-aware applications is always desirable for people in processing information and making decisions.

Acknowledgements

This work was supported by the Ministry of Science and Technology, Taiwan (Grant number: MOST 102-2221-E-259-029; MOST 103-2221-E-178-001).

References

[1] Hridoy, Anwar S. A., Ekram, Tahmid, M., Islam, Samiul, M., Ahmed, F., Rahman, R.M. (2015). Localized Twitter opinion mining using sentiment analysis. *Decision Analytics,* 2 (8).

[2] Baccianella, S., Esuli, A., Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *In:* Proc. of the 7th Conference on International Language Resources and Evaluation, 2200-2204.

[3] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z. (2007). Optimizing Web Search Using Social Annotations. *In: Proceedings of the 16th International World Wide Web Conference*, 501-510.

[4] Cambria, E., Hussain, A. (2012). Sentic Computing: Techniques, Tools, and Applications. Dordrecht, Netherlands: Springer.

[5] Carbonell, J.G. (1979). Subjective Understanding: Computer Models of Belief Systems. PhD Thesis, Dept. of Computer Science, Yale University.

[6] Chaovalit, P., Zhou, L. (2005). Movie Review Mining:

[8] For a small k, the adjustment of μ, Õ, À, and » can result in a larger number of extracted sentences than k. In contrast, a smaller number of extracted sentences than k. Thus, the adjustment of parameters is discarded.

[9] The relative improvement is calculated as (b – a) / a * 100 when b is compared to a.

A Comparison between Supervised and Unsupervised Classification Approaches. *In:* Proceedings of the 38th Annual Hawaii International Conference on System Sciences.

[7] Chen, S.-Y., Tseng, T.-T., Ke, H.-R., Sun, C.-T. (2011). Social trend tracking by time series based social tagging clustering. *Expert Systems with Applications* 38 (10), 12807-12817.

[8] Chenlo, J. M., Losada, D. E. (2013). A machine learning approach for subjectivity classification based on positional and discourse features. *Lecture Notes in Computer Science,* (8201) 17-28.

[9] Chikersal, P., Poria, S., Cambria, E., Gelbukh, A., Siong, C. E. (2015). Modelling public sentiment in Twitter: Using linguistic patterns to enhance supervised learning. *Lecture Notes in Computer Science,* (9042), 49-65.

[10] Chinsha, T. C., Shibily, J. (2015). A Syntactic Approach for Aspect Based Opinion Mining. *In:* Proceedings *of the 2015 IEEE Conference on Semantic Computing*, 24-31.

[11] Church, K. W., Hanks P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics,* 16 (1), 22-29.

[12] Dalal, M. K., Zaveri, M. A. (2014). Opinion mining from online user reviews using fuzzy linguistic hedges. *Applied Computational Intelligence and Soft Computing*, 2014. Available at http://dx.doi.org/10.1155/2014/735942.

[13] Dave, K., Lawrence, S., Pennock, D. M. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *In:* Proceedings of the 12th International World Wide Web Conference, p. 519-528.

[14] Du, R., Lu, Z., Pandit, A., Kuang, D., Crittenden, J., Park, H. (2015). Toward Social Media Opinion Mining for Sustainability Research. *In:* Proceedings of the 2015 AAAI Workshop on Computational Sustainability, 21-23.

[15] Ganeshbhai, S. Y., Shah, B. K. (2015). Feature Based Opinion Mining: A Survey. *In:* Proceedings *of the 2015 IEEE International Conference on Advance Computing Conference*, 919-923.

[16] Hatzivassiloglou, V., McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. *In:* Proceedings *of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, 174-181.

[17] Hatzivassiloglou, V., Wiebe, J. M. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. *In:* Proceedings of the 18th International Conference on Computational Linguistics, 299-305.

[18] Heymann, P., Koutrika, G., Garcia-Molina, H. (2008). Can Social Bookmarking Improve Web Search? *In:* Proceedings of the 1st ACM International Conference on *Web Search and Data Mining*, 195-206.

[19] Hogenboom, A. Boon, F., Frasincar, F. (2012). A statistical approach to star rating classification of sentiment. *Advances in Intelligent Systems and Computing* 171, 251-260.

[20] Hu, M., Liu, B. (2004). Mining and Summarizing Customer Reviews. *In:* Can Social Bookmarking Improve Web Search? *In:* Proceedings of the 1st ACM International Conference on Web Search and Data Mining, p. 195-206. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 168-177.

[21] Jarvelin, K., Kekalainen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems,* 20 (4), 422-446.

[22] Kamps, J., Marx, M. (2002). Words with Attitude. In: Proceedings of the 1st International Conference on Global WordNet, 332-341.

[23] Kim, S. -M., Hovy, E. (2004). Determining the Sentiment of Opinions. *In:* Proceedings of the 20th *International Conference on Computational Linguistics*.

[24] Kim, S.-M., Hovy, E. (2005). Automatic Detection of Opinion Bearing Words and Sentences. *In*: Proceedings of the 2nd International Joint Conference on Natural Language Processing, 61-66.

[25] Ku, L. -W., Liang, Y. -T., Chen, H. -H. (2006). Opinion Extraction, Summarization and Tracking in News and Blog Corpora. *In*: Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, 100- 107.

[26] Ku, L.-W., Lo, Y.-S., Chen, H.-H. (2007). Using Polarity Scores of Words for Sentence-Level Opinion Extraction. *In*: Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, 316-322.

[27] Li, S., Lee, S.Y.M., Chen, Y., Huang, C.-R., Zhou, G. (2010). Sentiment Classification and Polarity Shifting. *In*: Proc. of the 23rd International Conference on Computational Linguistics, p. 635-643.

[28] Liu, B. (2010). Sentiment Analysis and Subjectivity. In: Handbook of Natural Language Processing (2nd ed.), N. Indurkhya and F. J. Damerau, (eds.). New York, NY: Chapman and Hall/CRC, 627-666.

[29] Matsuo, Y., Ishizuka M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools,* 13 (1), 157-169.

[30] Mihalcea, R., Banea, C., Wiebe, J. (2007). Learning Multilingual Subjective Language via Cross-Lingual Projections. *In*: Proceedings of the 45th Annual Meeting

of the Association for Computational Linguistics, 976-983.

[31] Montoyo, A., Martinez-Barco, P., Balahur, A. (2012). Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. *Decision Support Systems*, 53 (4), 675-679.

[32] Mukherjee, S., Bhattacharyya, P. (2012). Feature specific sentiment analysis for product reviews. *Lecture Notes in Computer Science*, (7181), 475-487.

[33] Pang, B., Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2 (1-2), 1-135.

[34] Pang, B., Lee, L. (2005). Seeing Starts: Exploring Class Relationships for Sentiment Categorization with respect to Rating Scales. *In:* Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 115-124.

[35] Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs Up? Sentiment Classification Using Machine Learning Techniques. *In:* Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, 79-86.

[36] Ravi, K., Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems,* 89, 14-46.

[37] Riloff, E., Wiebe, J. (2003). Learning Extraction Patterns for Subjective Expressions. *In:* Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 105-112.

[38] Seki, Y., Evans, D. K., Ku, L. -W., Chen, H. -H., Kando, N., Lin, C.-Y. (2007). Overview of Opinion Analysis Pilot Task at NTCIR-6. In: Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, 265-278.

[39] Severyn, A., Moschitti, A., Uryupina, O., Plank, B., Filippova, K. (2014). Opinion Mining on YouTube. *In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1252-1261.

[40] Shi, J., Zhu, Y. (Ed.) (2005). Bao Yi Ci Ci Dian (Positive Dictionary). Chengdu, Sichuan, China: Sichuan Dictionary Press.

[41] Trant, J. (2009). Studying social tagging and folksonomy: a review and framework. *Journal of Digital Information,* 10 (1).

[42] Turney, P. D. (2002). 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, 417-424.

[43] Vinodhini, G., Chandrasekaran, R. M. (2014). Opinion mining using principal component analysis based ensemble model for e-commerce application. *CSI Transactions on ICT,* 2 (3), 169-179.

[44] Wiebe, J. M., Bruce, R. F., O'Hara, T. P. (1999). Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. *In:* Proceedings of the 37th *Annual Meeting of the Association for Computational Linguistics,* 246-253.

[45] Wiebe, J. M., Rapaport, W. J. (1988). A Computational Theory of Perspective and Reference in Narrative. *In:* Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics, 131-138.

[46] Wiebe, J., Riloff, E. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. *In:* Proceedings *of the 6th International Conference on Intelligent Text Processing and Computational Linguistics*, 486-497.

[47] Wilson, T., Wiebe, J., Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *In:* Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 347-354.

[48] Wlodarczak, P., Ally, M., Soar, J. (2015). Opinion Mining in Social Big Data. *Available at SSRN:* http://dx.doi.org/10.2139/ssrn.2565426.

[49] Yang, L., Zhu, Y. (Ed.). (2005). Bian Yi Ci Ci Dian (Negative Dictionary). Chengdu, Sichuan, China: Sichuan Dictionary Press.

[50] Yu, H., Hatzivassiloglou, V. (2003). Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 129-136.

[51] Zollers, A. (2007). Emerging Motivations for Tagging: Expression, Performance and Activism. *In:* Proceedings of the of the WWW 2007 Workshop on Tagging and Metadata for Social Information Organization.