Sentiment analysis for Chinese reviews of movies in multi-genre based on morpheme-based features and collocations

Heng-Li Yang · August F. Y. Chao

Published online: 23 May 2014

Abstract The application of sentiment analysis, also known as opinion mining, is more difficult in Chinese than in Indo-European languages, due to the compounding nature of Chinese words and phrases, and relatively lack of reliable resources in Chinese. This study used seed words, Chinese morphemes, which are mono-syllabic characters that function as individual words or be combined to create Chinese words and phrases, to classify movie reviews found on Yahoo! Taiwan. We utilized higher Pointwise Mutual Information (PMI) collocations, which consist of selected morpheme-level compounded features to build classifiers. The contributions of this study include the following: (Bird 2006) proposing a method of generating domain-dependent Chinese morphemes directly from large data set without any predefined sentimental resources; (Bradley and Lang 1999) building morpheme-based classifiers applicable in various movie genres, and shown to produce better results than other classifiers based on keywords (NTUSD and HowNet) or feature selection (TFIDF); (Church and Hanks in Computational linguistics, 16(1), 22-29 1990) identifying compounds that have different semantic polarities depending on contexts.

Keywords Sentiment analysis · Opinion mining · Morpheme · Feature · Collocation · Chinese movie

H.-L. Yang (✉) · A. F. Y. Chao
Department Management Information Systems, National Cheng Chi University, 64, Sec.2, Chihnan Road, Wenshan District, Taipei, Taiwan, Republic of China
e-mail: yanh@nccu.edu.tw

A. F. Y. Chao
e-mail: aug.chao@gmail.com

1 Introduction

Emerging computer technologies have made it possible to design mechanisms capable of gathering data from Internet sources such as blogs, filter the data using predefined categories, and identify the opinions of users by distinguishing between positive and negative responses. This type of sentiment mining has received considerable attention due to the exponential growth of online data with the advent of mobile devices and social network sites. The development of sentiment corpora and mechanisms for machine learning are critical in deciphering online postings, which are often unstructured and loosely formatted (Li and Wu 2010).

In addition to manually created wordlists (e.g., ANEW, Affective Norms for English Words, refering to (Bradley and Lang 1999), there are many mature sentiment corpora used for natural language processing (NLP) and the comprehension of word sense in English. One state-of-the-art English sentiment corpus, SentiWordNet (Esuli and Sebastiani 2006), employs machine learning to classify words found in WordNet, which has predefined positive or negative connotations of the words and synsets, or groups of cognitive synonyms (Miller 1995). SentiWordNet can be used to identify the polarity of reviews in many domains. By using pre-tagged wordlists and applying a large corpus to extend the lists of positive and negative words, researchers have produced numerous machine learning algorithms capable of identifying sentiment and thereby enabled the extraction of semantic orientation from reviews.

Methods of sentiment classification, which derived from a combination of text-mining techniques and NLP techniques have been used to identify a given review as positive or negative. Researchers have developed many approaches, particularly in English, to process the sentiments found in opinions from a variety of perspectives. However, sentiment analysis in Chinese is another story. First, Chinese considerably differs from Indo-European languages. Every Chinese character has its own associated meaning, and modern Chinese words consist of one to six characters or ideographic meanings (Wu and Tseng 1993). The absence of word boundaries makes it extremely difficult to assign correct parts-of-speech tags and perform the meaningful segmentation of sub-sentences or phrases in the processing of natural language. Thus, developing methods to disambiguate Chinese word sense poses numerous challenges. Despite the availability of Chinese sentiment corpora, e.g., National Taiwan University Sentiment Dictionary (NTUSD) (Ku et al. 2006) and HowNet (Dong and Dong 2006), the difficulties applying NLP techniques to Chinese would compromise accuracy in the use of extension wordlists. Another approach would adapt a mature English corpus to Chinese sentiment analysis, but it still requires overcoming the inherent differences between the languages as well as the poor performance of machine translation (Wan 2009). In order to prevent the meaning of opinions expressed in Chinese from being misinterpreted by machines, it requires reflection upon the nature of the Chinese language itself to enable the processing of opinions directly from the most basic elements used to represent concepts. This study tried to challenge the above Chinese sentiment analysis problems.

Furthermore, this study sought to overcome two additional problems associated with the analysis of sentiment in Chinese. First, although words in manually established wordlists have well-known positive or negative connotations, a number of neutral words, which are not included in wordlists, can imply either positive or negative senses in their syntactic features (also called aspects), regardless of whether the features are explicit or implicit (Liu 2010). For example, we may say, “the battery life (feature) of this cellular phone is too short,” where “short” is a neutral word, but it has a negative sense in the above context. Second, the dynamic sentiment of word senses in different contexts can express totally different sentiment orientation (Wu and Wen 2010). For example, different from common sense, “悚” (terrifying) and “皮疙瘩” (goose bumps) appearing in opinions for horror movies would express positive polarity.

Zhang et al. (Zhang et al. 2012) found product weakness from Chinese reviews by using morpheme-based sentiment analysis and relying on the similarity calculation in a predefined wordlist, Hownet. This study tried to resolve the problem of the wordlist constraint. We suggested a morpheme-based method of feature selection to search for domain-dependent Chinese compound words directly from the reviews in a large data set without any help of predefined sentimental resources. Because of the availability of data set, we took the sentiment analysis of movie reviews written in Chinese as the example for demonstrating the superiority of our approach. To assemble sentiment orientation wordlists from the data itself, we collected opinions written about movies from Yahoo! Taiwan and compiled them into movie opinion corpus, containing 127,424 opinions in 18 categories of movies with a total of 4,631,482 words. Considering the star-rating as an indicator of either positive or negative sentiment (Turney 2002), and thus negating potential problems with the inference of star-ratings from reviews (Pang and Lee 2005), this study used PMI (point-wise mutual information) to search for co-occurring phrases in the modification of morpheme-level features to be used as signatures of sentiment, from which to build SVM (support vector machine) classifiers. We then compared the effectiveness of the proposed classifier with that of the classifiers built by TF-IDF (term frequency, inverse document frequency), NTUSD and HowNet, with regard to the analysis of opinions of movies of various genres. Finally, we analyzed the sentiment compounds generated by the proposed classifier and compared these sentiments with those generated by NTUSD and HowNet wordlists.

The following paper structure is organized as follows. Previous related works are presented in Section 2. Section 3 introduces the proposed algorithm, dataset collection, and analysis methods. Evaluation and comparison results are presented in Section 4. In Section 5, we draw conclusions and discuss the findings of this study and future work.

2 Related works

2.1 Sentiment analysis

Sentiment analysis can be categorized into phrase-level, sentence-level, and document-level analyses (Pang and Lee 2008). Commonly used Chinese sentiment dictionaries NTUSD (Ku et al. 2006) and HowNet (Dong and Dong 2006) identify polarity as follows. NTUSD uses manually tagged phrases; HowNet determines polarity using its own Chinese common sense knowledge base. Both wordlists can be used to perform phrase-level (Li et al. 2009; Sun et al. 2010) and sentence-level sentiment analysis (Li and Yao 2007; Ku et al. 2008). To expand sentiment wordlists, statistical analysis and pattern matching can be adapted to match words already classified in the wordlists with additional words sharing their sentiment orientation. One statistical method is PMI, which pairs words and compares their co-occurrence. The PMI algorithm is defined as follows:

$$PMI(word_1, word_2) = \log_2\left[\frac{p(word_1\&word_2)}{p(word_1)p(word_2)}\right]$$

PMI can be modeled in both small and large window sizes. Using the smaller window size, PMI searches idioms and common phrases; and using the larger window size, PMI highlights semantic concepts and other larger relationships among words (Church and Hanks 1990). Turney (Turney 2002) developed Sentiment Orientation Point-wise Mutual Information (SO-PMI) to calculate the co-occurrence probability of words and phrases in search engines using the NEAR operation. SO-PMI relies on the assumption that a set of web pages can be considered a large corpus. Although NEAR operation is no longer available in current search engines, a number of researchers have proposed a modified version of SO-PMI (formula presented below) for adaption to Chinese sentiment analysis (Ye et al. 2006; Feng et al. 2012). Although SO-PMI can analyze predefined Chinese keyword lists with ease, using SO-PMI to determine the sentiment of unknown words can be problematic because the corpus used by the algorithm is insufficient for the extension of sentiment words.

$$SO\text{--}PIM(phrase) = \log_2\left[\frac{hits(phrase\,\mathrm{NEAR}\,excellent)hits(poor)}{hits(phrase\,\mathrm{NEAR}\,poor)hits(excellent)}\right] \tag{1}$$

Excellent and poor are two sentiment polarities; NEAR is the search engine operation; hits (token) is the number of tokens returned.

Pattern matching requires Chinese dictionaries, an understanding of grammar details, and natural language processing (NLP) tools to parse sentences into dependency trees capable of isolating sentiment words and their corresponding features. (Tan and Zhang 2008) employed a rule-based approach to Chinese sentiment analysis based on the HowNet lexicon and syntactic structures and analyzed 1,021 documents spanning topics in a variety of domains. Their study analyzed the ranking of books, music, and movie reviews from Amazon China. They reported 79.98 % accuracy using a SVM classifier. These methods can be used in conjunction with a thesaurus to enhance the recognition of sentiment words and improve parsing performance (Xu et al. 2011).

Pre-tagged wordlists are considered essential to Chinese sentiment analysis; however, it is also possible to analyze opinions without the use of seed words. Nasukawa and Yi (Nasukawa and Yi 2003) proposed an NLP-based approach to sentiment analysis for the extraction of sentiment directly from opinions. By applying syntactic parser and self-built sentiment lexicon, their prototype extracts the level of favorability emotion expressed with respect to the topic in opinions. Wan (2009) proposed a pure machine learning approach, known as bilingual co-training, to train both unlabeled product Chinese reviews and its machine-translated English reviews. By leveraging various machine translation services to eliminate the language gap, the bilingual co-training method can outperform both basic and transductive methods.

2.2 Morpheme in chinese

Chinese text consists of a linear sequence of non-spaced or equally spaced ideographic characters, which are similar to morphemes in English (Wu and Tseng 1993; Wu and Tseng 1999). According to morphological processing, most compound words (compound ideographic characters) represent the form and semantic processing of their constituent morphemes (Zhou et al. 1999), Yuen et al. (2004) conducted a pilot study on strongly-polarized Chinese words, which are composed of positive morphemes (e.g., 獎(gift), 勝(win) 優(good)) or negative morphemes (e.g., 傷(hurt), 貪(greedy), 疑(doubt)). They performed sentiment analysis on the Linguistic Variations in Chinese Speech Communities (LIVAC) corpus. Their research indicated that sentiment analysis can employ the morpheme within each compound to express compound sentiment, and thereby determine the sentiment of the sentence. They claimed that their approach could enhance the effectiveness of sentiment analysis algorithms, even in the absence of a Chinese corpus and without the costs associated with word segmentation. They attributed the efficacy of their method to its focus on morpheme words with sentiment meaning, e.g., 幸(luck), which is a morpheme of 幸運(lucky). From a linguistic point of view, Ku, Huang and Chen (Ku et al. 2009) examined the morphological structures found in Chinese syntax: compounding, affixation, and conversion. They categorized Chinese compound words into eight morphological types in order to perform sentiment analysis. In an experiment, they searched for the sentiment of words according to morphological type and tested those words in both word-/sentence-level polarities. Although morphological information is seldom applied either in Chinese opinion extraction, or in solving the problems of coverage found in opinion dictionaries, Ku, Huang and Chen reported that the adoption of morphological information improves the performance of word polarity detection. Wang, et al. (Wang et al. 2011) separated sentiment words into static sentiment words (SSWs) (i.e., words whose sentiments do not change), and dynamic sentiment words (DSWs) (i.e., words whose sentiments' change would depend on contexts), and then computed the morphological productivity of sentiment words. Furthermore, Zhang et al. (2012) introduced an expert system, called as Weakness Finder, which extracted the features and grouped explicit features by using morpheme based method and HowNet based similarity measurement, and identified and grouped the implicit features with collocation selection method for each aspect.

2.3 Comments on literature review

In general, there are three approaches while doing Chinese sentiment analysis: using and extending pre-defined sentiment keyword lists, using basic natural language processing techniques to extract features and sentiments, and adopting English sentiment analysis resources. Each approach has its problems, which are not yet been fully solved.

The annotated resources for sentiment classification in Chinese are not abundant, so pre-defined sentiment wordlists would not work well in most cases. The co-training method tried to adopt English sentiment resources, but the language gap between English and Chinese is not easily eliminated. Probably we need directly apply basic natural language processing techniques, but there are still some problems to deal with Chinese words.

The morpheme-based method used in text mining might solve the Chinese language processing problem, even in the absence of a Chinese corpus. Zhang et al. (2012) proposed an expert system Weakness Finder by analyzing the customers' reviews on the influential web communities with morpheme based sentiment analysis. However, one should note that they applied morpheme to search for similar concepts in HowNet, which is a predefined word list. It could be claimed in this study that if a dataset is large enough to represent language characteristics in a specified domain, the feature and sentiment compounds would be frequently co-addressed. Therefore, it is our motive to propose morpheme-based sentiment analysis method to extract domain-dependent Chinese morphemes directly from large data set without the help of predefined sentimental resources. We used the data set containing movie reviews written in Chinese as the example for demonstrating the superiority of our method. It is our wish that our method can be applied in situations which proper predefined sentiment resources in Chinese are not available.

3 The proposed approach

3.1 Overview

This study performed sentiment analysis of Chinese without using any resources for sentiment analysis, considering the dearth of annotated resources for sentiment classification. To find sentiment expressions for a given genre and determine the polarity of the sentiment, we adopted the morpheme-based technique of identifying features, to search for corresponding phrases in blogs that express movie review sentiment. We attempted to identify sentence fragments that express the sentiment of the opinions expressed in the text, and to create Chinese dynamic sentiment lexicons that express different sentiment for different contexts.

3.2 Dataset

Assembling a sentiment orientation wordlist from a dataset requires a large dataset containing compound words. We collected movie reviews from Yahoo!Taiwan, and compiled them into a corpus of movie opinions. The corpus contained 127,424 opinions categorized into 18 genres. Each opinion was ranked on a five-star scale. The corpus includes a total of 4,631,482 words (see Appendix I). One movie could belong to one or more genres. The distribution of collected opinions is presented in Fig. 1. The distribution of stars was as follows: one/two stars (22.6 %), three stars (8.6 %), and four/five stars (68.8 %). All movie opinions were first processed according to the part-of-speech (P.O.S) tagger SINICA CKIP,[1] and then stored as a dataset in the web-based hosting service GitHub.[2]

3.3 Morpheme-based features and collocations

3.3.1 Selected morpheme

To extract compound words with meaningful sentiment from movie reviews, Chinese morphemes must be first decided. There were 93,871 distinct words, which appeared at least once in the total 4,631,482 words of the corpus. After excluding those compounds that have no essential meaning for our sentimental analysis (e.g. conjunction, quantifier, etc.), there were 74,366 words. Referring to the properties from movie in http://schema.org/, we first listed those movie description words whose frequencies were high at least as first 10 %. Then, we must decide the morphemes. Unlike in English, features in Chinese are multi-syllabic compounds of morphemes that express specific meanings. There were total possible 4,820 morphemes for these 74,366 words. After consulting with two experts about movie feature compounds, we finally identified eight Chinese morphemes that semantically express different features presented in movie reviews and their frequencies were also high at least as first 5 % in these 4,820 morphemes. Table 1 lists eight morphemes used to search feature compounds, and each morpheme presents a different sense regarding to movie features, such as actors, plots, and special effects. Although we can define morpheme roots for

[1] A Part-Of-Speech Tagger (P.O.S Tagger) is software that reads text and designates each word as a part of speech (and other token), such as noun, verb, adjective. The Part-of-speech tools from SINICA CKIP are available at http://ckipsvr.iis.sinica.edu.tw/

[2] Collected Taiwan Yahoo!Movies Corpus with P.O.S Tags from CKIP, https://github.com/fychao/ChineseMovieReviews

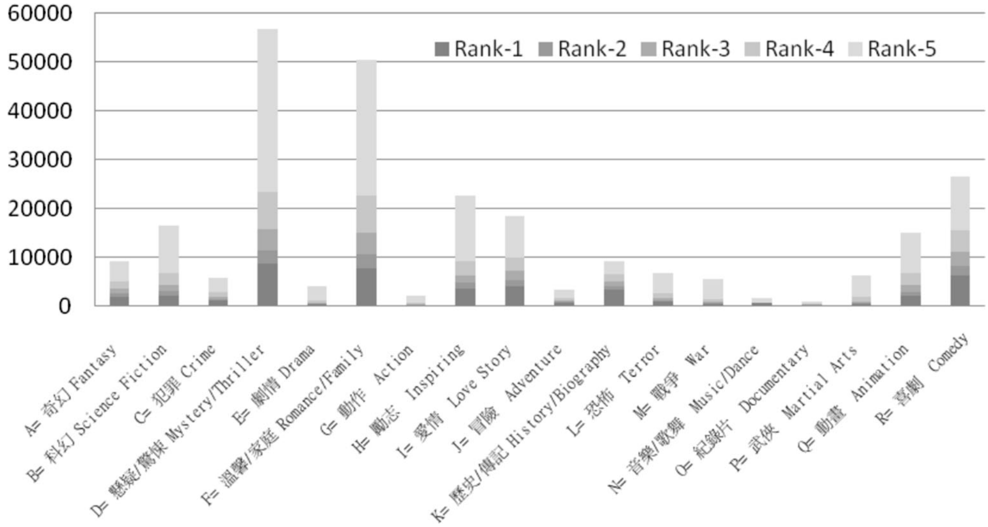


Fig. 1 Distribution of movie opinions collected from Yahoo!Taiwan

feature compounds, different datasets would produce different results.

3.3.2 Using results from NLP tools

Compounding words are rooted in morphemes and can be used to express a variety of movie features, such as actors, plots, and special effects. However, understanding the relationship between the meaning of features and parts-of-speech tags can be difficult without considering the context in which the words are used. For example, 編劇 has two senses in Chinese, screenwriter (noun) and screen-writing (verb). Therefore, more information is required from the sentence to identify its meaning.

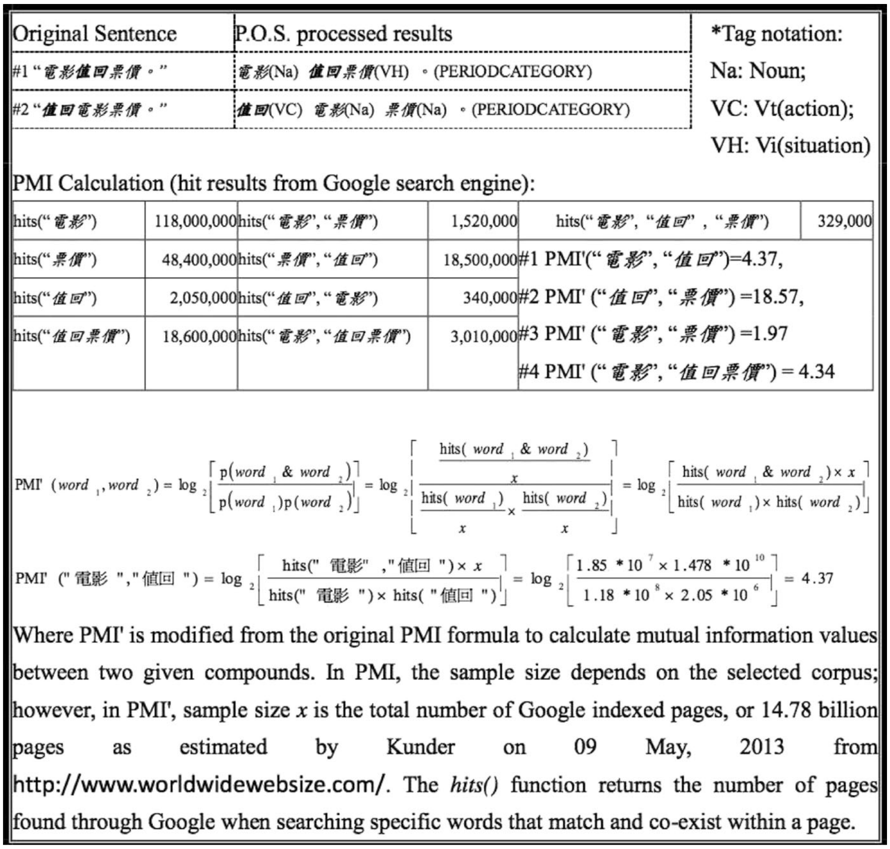
Word segmentation is the necessary first task in NLP in Chinese; however, it is difficult to ensure accurate results when NLP tools are applied to the extraction of movie features from reviews written in Chinese. When using NLP tools to merge different segments of compounds, the PMI value is an important indicator in determining whether two adjacent words (or compounds) should be merged to create a single meaningful phrase. As shown in Table 2, which outlines the problems associated with the segmentation of compounds in Chinese, we consider the example of 值回票價 (get one‘s money’s worth) in two sentences processed by SINICA CKIP .

In the first sentence analyzed in Table 2, the NPL tool merges compounds into a common Chinese idiom 值回票價, because 值回 (worth) and 票價 (ticket price) can be found in adjacent positions with a PMI’=18.57, which is relatively high. However, in the second sentence, the grammatical analysis performed by the NPL tool suggests merging 電影 (movie) and 票價 as a noun phrase, and it disregards the high value of PMI’ (“值回”, “票價”) because 值回 and票價 are not adjacent. If a Chinese speaker analyzed the sentiment expressed in the compounds in both sentences, the Chinese idiom, 值回票價, would be easily identified as a positive sentiment toward the movie being considered. However, computing software has difficulty identifying the sentiment orientation of the compound 值回, which is seldom used and applied only for the modification of succeeding compounds. Inaccurate results would be the result of attempting to study the sentiment compounds based only on the results of NLP data processing without calibrating suitable information

Table 1 Selected morpheme for movie features

Morpheme	Original meaning	Metaphorical meaning in movies	Selected feature compounds
影(*yǐng*)	Shadow created by object.	The movie itself.	影評(movie reviews),電影(movie), 影帝(award actor)
劇 (*jù*)	One type of performances.	The story told by movie.	編劇(screenwriter or screen-writing), 劇情(plot), 劇本(script)
片 (*piàn*)	Unit for flat things, mind state, area, and scope.	Unit for movie.	影片(movie), 西片(western movie), 片尾(end of movie)
角 (*jiǎo*)	Horn, a prominent object.	Actors or actresses in movies.	主角(protagonist), 角色(actors), 角度(point of view)
效 (*xiào*)	The effect.	Visual or sound effects.	效果(effect), 特效(special effect), 音效(sound effect)
演(*yǎn*)	Perform, evolve.	The performing actions in movies.	演員(actors), 導演(director), 演技(skills of performance)
票(*piào*)	Unit for tickets, tickets.	The movie tickets.	票價(ticket price), 票房(box office), 電影票(movie tickets)
結 (*jiē*)	Knot.	The key or end point of movie story.	結局(ending), 結束(end), 結構(structure)

Table 2 Chinese word segment: comparson for 值回票價



Original Sentence	P.O.S. processed results
#1 "電影值回票價。"	電影(Na) 值回票價(VH) 。(PERIODCATEGORY)
#2 "值回電影票價。"	值回(VC) 電影(Na) 票價(Na) 。(PERIODCATEGORY)

*Tag notation:
Na: Noun;
VC: Vt(action);
VH: Vi(situation)

PMI Calculation (hit results from Google search engine):

hits("電影")	118,000,000	hits("電影", "票價")	1,520,000	hits("電影", "值回" , "票價")	329,000
hits("票價")	48,400,000	hits("票價", "值回")	18,500,000	#1 PMI'("電影", "值回")=4.37,	
hits("值回")	2,050,000	hits("值回", "電影")	340,000	#2 PMI' ("值回", "票價") =18.57,	
hits("值回票價")	18,600,000	hits("電影", "值回票價")	3,010,000	#3 PMI' ("電影", "票價") =1.97 #4 PMI' ("電影", "值回票價") = 4.34	

$$\mathrm{PMI'}\ (word_1, word_2) = \log_2\left[\frac{p(word_1\ \&\ word_2)}{p(word_1)p(word_2)}\right] = \log_2\left[\frac{\frac{\mathrm{hits}(word_1\ \&\ word_2)}{x}}{\frac{\mathrm{hits}(word_1)}{x}\times\frac{\mathrm{hits}(word_2)}{x}}\right] = \log_2\left[\frac{\mathrm{hits}(word_1\ \&\ word_2)\times x}{\mathrm{hits}(word_1)\times \mathrm{hits}(word_2)}\right]$$

$$\mathrm{PMI'}\ (\text{"電影"}, \text{"値回"}) = \log_2\left[\frac{\mathrm{hits}(\text{"電影"}, \text{"値回"})\times x}{\mathrm{hits}(\text{"電影"})\times \mathrm{hits}(\text{"値回"})}\right] = \log_2\left[\frac{1.85 * 10^7 \times 1.478 * 10^{10}}{1.18 * 10^8 \times 2.05 * 10^6}\right] = 4.37$$

Where PMI' is modified from the original PMI formula to calculate mutual information values between two given compounds. In PMI, the sample size depends on the selected corpus; however, in PMI', sample size x is the total number of Google indexed pages, or 14.78 billion pages as estimated by Kunder on 09 May, 2013 from http://www.worldwidewebsize.com/. The *hits()* function returns the number of pages found through Google when searching specific words that match and co-exist within a page.

granularity. On the other hand, it would be unacceptable to divide merged compounds or idioms into smaller information granularity units without considering their grammatical structures, or study sentiment compounds alone without considering their modified features.

3.3.3 Selecting collocations

Turney's method of analyzing sentiment orientation (described in Section 2.1) has been seldom applicable in Chinese. Furthermore, Chinese NLP tools are unable to identify the correct features and their corresponding sentiment compounds. For example, the word 好 (good, well), as a verb modifier, normally has a positive meaning, such as in 好看 (good-looking, or handsome, tagged as 好看 Vi, an intransitive verb, by SINICA CKIP). Nonetheless, 好 cannot be separated from 看 (look) by tagging tools. On the contrary, 好 takes on a negative connotation in some compounds, e.g., 好難看 (quite bad-looking, or quite difficult to look at, tagged as 好 Vi and 難看 Vi by SINICA CKIP). In fact, 好 can even be segmented as a single word with no sense of sentiment at all.

This study adopted Tureny's design in proposing a novel method in which features are combined with collocations (i.e., the corresponding compounds that are used to modify features) to facilitate our understanding of the concepts. Shared concepts can be calculated according to the probability of words co-existing in sentences across a corpus. Turney (Turney 2002) suggested using PMI to determine concepts of sentiment orientation shared between extracted phrases and their representative sentiment polarities, which are "excellent" and "poor". He utilized results from search engines utilizing NEAR operation, which performs a search for co-existing words within a ten word window size in order to identify synonyms (Turney 2001). Given the polarized sentiments "excellent" and "poor", unknown English words can be found its sentiment orientation by calculating the SO-PMI in Formula 1 within a window size. Unfortunately, as in the above described example (好看 and 好難看), Chinese NLP tools do not provide support sufficient to enable sentiment analysis. The orientation of each feature-collocation combination needs to be considered as a joint conceptual unit, which sentiment orientation is judged according to the sentences in context.

In order to find the corresponding collocations of features, we limited window size to ±10 to select feature compounds, if there were no stop words or end punctuation found within this range. PMI values as low as −2 were permitted because we wanted to extract any shared concepts that could provide clues regarding the sentiments expressed in the opinions within the dataset, even if those concepts were not found in frequently co-existing compounds or common phrases.

3.4 SVM classifier and evaluation

3.4.1 Linear SVM

In order to label each feature-collocation combination as positive or negative sentiment, we adopted SVM model (Vapnik 1995) to learn and classify movie opinions. The idea behind SVM is to find a decision surface over a vector space to enable separating the data into the two classes. This study used a linear SVM model, which considers arbitrary data $\overrightarrow{x}$ scattered in a separable space, and learns vector $\overrightarrow{w}$ and constant b from a training set of data, allowing the model to find the decision hyperplane, written as follows:

$$\overrightarrow{w} \bullet \overrightarrow{x} - b = 0$$

Let training data set $D = \left\{ \left(y_1, \overrightarrow{x}_i \right) \right\}$ be the collected movie opinions, and $y_i \in \{\pm 1\}$ be the positive (+1) and negative (−1) classification for $\overrightarrow{x}$. The linear SVM problem involves finding $\overrightarrow{w}$ and b values capable of satisfying the following constraints to minimize the 2-norm of vector $\overrightarrow{w}$.

$$\overrightarrow{w} \bullet \overrightarrow{x} - b \geq +1 \, for \, y_i = +1$$

$$\overrightarrow{w} \bullet \overrightarrow{x} - b \leq -1 \ \ for \, y_i = -1$$

Various researchers (Tan and Zhang 2008; Ku et al. 2009; Sun et al. 2010) have reported that SVM classifiers are more accurate than other classifiers, such as naïve Bayes, conditional random fields, and classifiers based on information gain.

3.4.2 Model evaluation

In order to measure the effectiveness, an F1 measure combining recall and precision (Van Rijsbergen 1979) is usually recommended as an SVM measurement. This study evaluated the precision, balanced accuracy and F1 scores of all the classifiers. The formulae are written below. In these formulae, tp represents true positive (correct results), fp represents false positive (unexpected results), fn represents false negative (missing results), and tn represents true negative (correct absence of results). One should note that in these often used formulae, there is one additional component "# without features". Previous researchers (e.g., (Ku et al. 2009)) applying the wordlists of NTUSD and HowNet would report the effectiveness after excluding the sentences without identifiable features since the size of their wordlists are constant. However, our method and the TFIDF method extract feature words according to the given corpus; the number of extracted feature words is used to determine the size of the wordlists. Therefore, for fairly comparing different methods, the number of those without identifiable features should be added back to the denominators of the Formula 2, 3, and 4.

$$Precision(Accuracy) = \frac{\# \, of \, tp}{\# \, of \, tp + \# \, of \, fp + \# \, without \, features} \tag{2}$$

$$Recall(Sensitivity) = \frac{\# \, of \, tp}{\# \, of \, tp + \# \, of \, fn + \# \, without \, features} \tag{3}$$

$$Specificity = \frac{\# \, of \, tn}{\# \, of \, fp + \# \, of \, tn + \# \, without \ features} \tag{4}$$

$$F1 = \frac{2 \times \mathrm{Recall} \times \mathrm{Pr}ecision}{(\mathrm{Re}call + \mathrm{Pr}ecision)} \tag{5}$$

$$Balanced \, Accuracy = \frac{Sensitivity \times Specificity}{2} \tag{6}$$

3.5 Experiment preparation

3.5.1 Define polarity in our data set

This study collected opinions written about movies from Yahoo!Taiwan as experimental data. These user comments are usually short and include a star ranking between one and five stars. These user rankings were utilized as a criterion for the implied sentiment orientation of opinions (a ranking of one to two stars were considered negative; three stars was neutral; four to five stars were considered positive). To eliminate the effect of neutral rankings, we would build three classifiers in our experiment: (1) a positive ("+") classifier to provide a positive sentiment cluster (four to five stars were considered positive; one to three stars were considered non-positive); (2) a negative ("-") classifier provided a negative sentiment cluster (one to two stars were negative; three to five stars were non-negative); (3) a positive–negative ("±") classifier provided both positive and negative sentiment clusters (one to two stars were considered negative; four to five stars were considered positive). In the random selection of opinions from the dataset, we attempted to balance the number of opinions in the positive and negative categories. However, discarding neutral (three-star) opinions proved very difficult, particularly when dealing with a large number of opinions.

3.5.2 Define referencing model

This study utilized morpheme-based feature-collocation combinations to assist in the determination of positive/negative sentiment orientation in opinions. For comparison, we also analyzed the dataset using other feature selection methods: TFIDF, and predefined sentiment keyword lists NTUSD (Ku et al. 2006) and HowNet (actually we used its subset, HowNet Sentiment Dictionary, called as Senti-HowNet) (Dong and Dong 2006).

TFIDF was implemented using the following formula to calculate whether terms should be designated as frequently appearing and thus be used to determine the positive/negative orientation of sentiments expressed in opinions:

$$TFIDF(t, d) = tf(t, d) \times \log(N/n_i)$$

where *tf*(*t*,*d*) is the number of times that term *t* occurs in document *d*, *N* is the total number of training opinions, and n_i is the number of opinions containing the word *t*.

Although NTUSD and HowNet distinguish between positive and negative words when a lexicon is applied to sentiment analysis, a number of words were not chunked or segmented in the same manner for the processing of opinions. For example, the concept element “悲傷” can be found different forms, such as “悲傷的”, “使悲傷” and “極度悲傷”. To further analyze

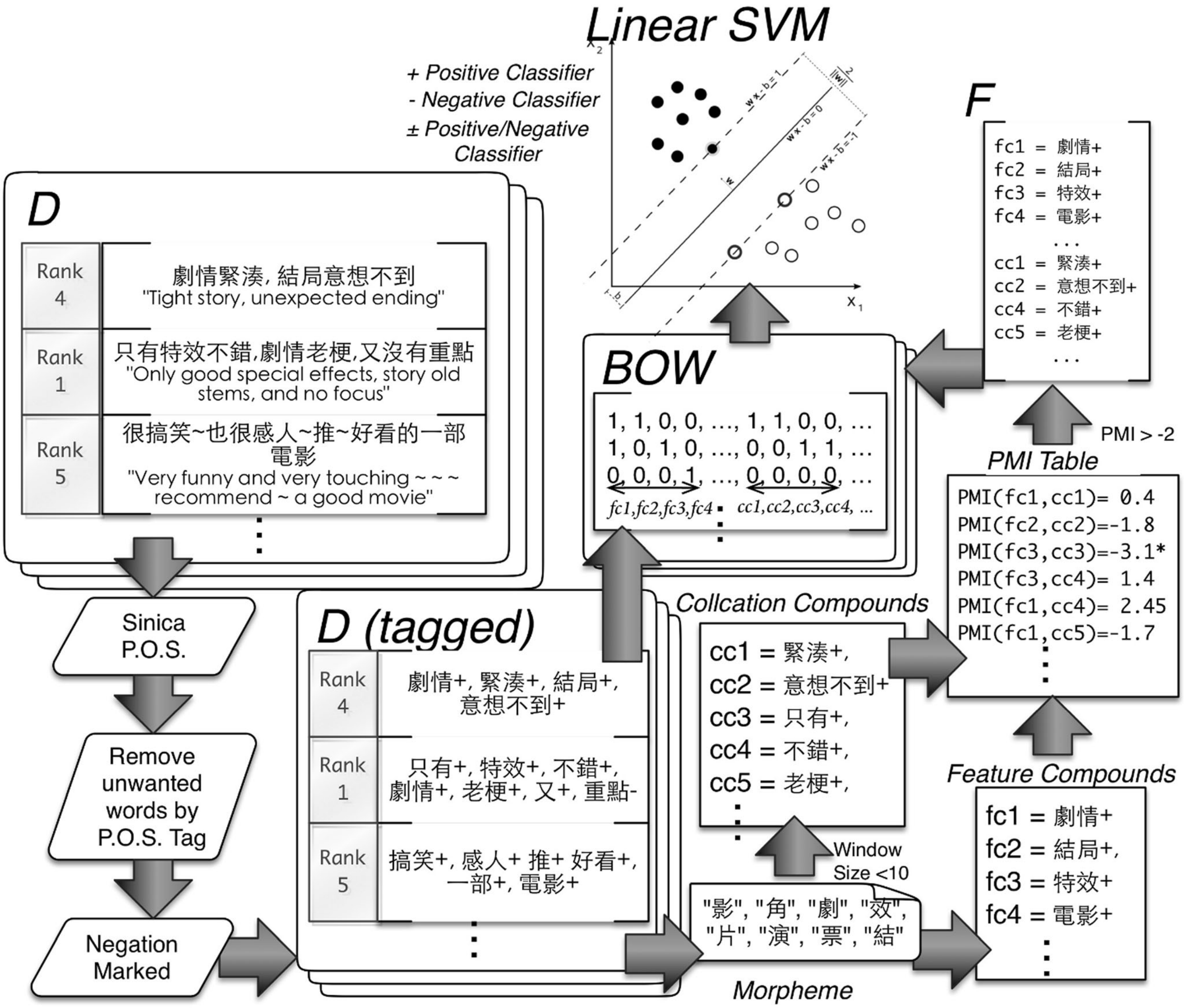


Fig. 2 Processing of data into SVM

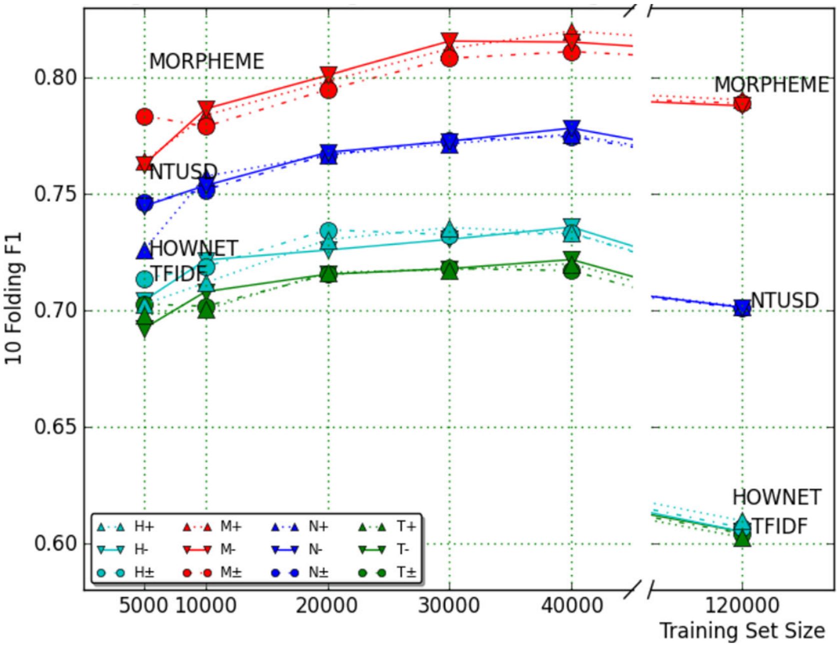


Fig. 3 Average F1 score in the training phase

the results produced by NTUSD and HowNet, we input both sets of results into the SINICA CKIP.

Although NTUSD and HowNet both label the sentiment orientation of each compounds, the P.O.S tool occasionally labels compounds with different orientations. Furthermore, following a pilot experiment using 40,000 random opinions, the performance of both keyword lists were disappointing when only one-sentiment orientations were used (see Appendix II). Thus, we disregarded pre-defined sentiment orientation and ultimately combined both positive and negative compounds to build the classifiers used in our experiment. In addition, the HowNet wordlist was originally encoded in simplified Chinese. Before inputting the lists into the P.O.S tool, these words were translated into traditional Chinese by referencing simplified/traditional Chinese conversion tables[3] from Wikipedia. We eventually employed 6,510 processed features in NTUSD and 7,555 features in HowNet for classifier training in our experiment scenarios. In the application of TFIDF for feature selection, we limited the maximum term size to 8,000 in order to extract only the most meaningful compounds for classification.

3.5.3 Pre-processing

Pre-processing involved sending user opinions to the SINICA CKIP for chunking and segmentation into fundamental concept units. Although P.O.S information is available for each compound segment, not all compounds are meaningful or open to interpretation in terms of sentiment orientation. Therefore, we used a tag-list (see Appendix III) to filter out unwanted compounds and considered only those compounds with essential meanings, which are the most likely to be P.O.S tagged as verbs and nouns in opinions.

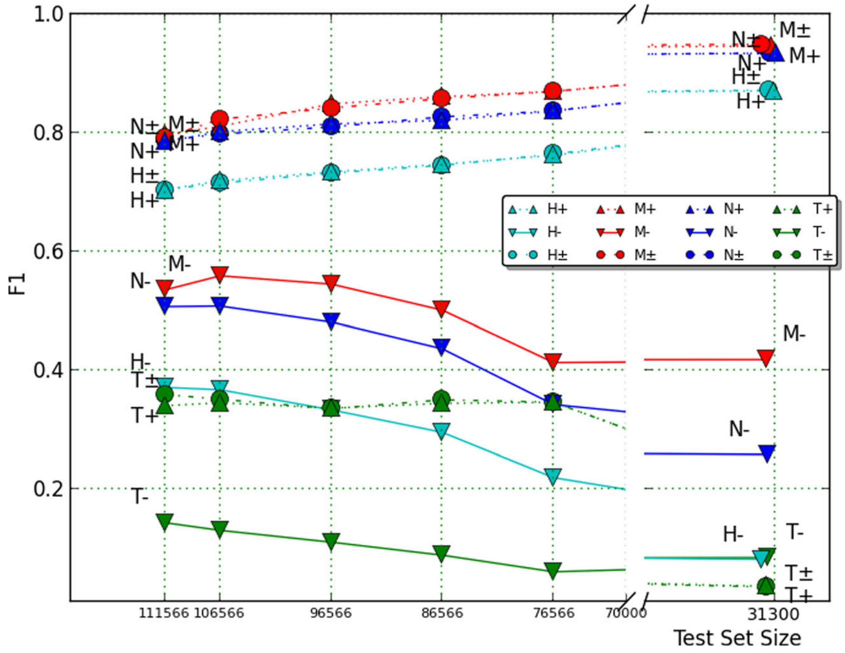
To extract the correct sentimental orientation implied by opinions, we referenced the idea of negation tagging from Das and Chen (Das and Chen 2001). In their work, the words "not," "no," and"never" in English were deemed as negative compounds. In this study, seven compounds were designated as negative compounds: '不', '沒有', '不要', '不能', '沒', '無', '不會' in Chinese. Das and Chen assumed that every word between a negation word and the first punctuation mark following the negation word would be affected. Due to fundamental differences in Chinese grammar, we designated the phrase preceding the sentence boundary tag as the area affected by the negative words. In addition, Das and Chen only marked negation as "-". This study marked each feature compound either with "+" (representing none or even number of negative words within a given range) or "-" (representing odd number of negative words). This resulted in an increase in the total number of compounds.

3.5.4 The whole processing procedures

As depicted in Fig. 2, before building any classifiers, we fed dataset *D* into the SINICA CKIP, removed unwanted words, and added negative/positive markers to obtain tagged-*D*. For example, the second sentence in dataset *D* would become "只有+特效+不錯+劇情+老梗+又+重點-", where "+" is a positive symbol, and "-" is a negative symbol. The negation mark in compound "重點-" (i.e., "focus –") is due to the phrase

[3] Simplified/traditional Chinese conversion tables include parallel translation of common words/phrases in Taiwan, China, Hong Kong, and Singapore, and can be retrieved from following link: http://svn.wikimedia.org/svnroot/mediawiki/trunk/phase3/includes/ZhConversion.php

Fig. 4 Average F1 score in the prediction phase



“沒有” (i.e., “no” or “none”) in the sub-sentence (“又沒有重點”). From the tagged-*D*, we obtained morpheme-based feature compounds and collocation compounds, i.e., the compounds with sentiment orientations waiting to be judged. We then searched for compounds containing selected morphemes and collocations within a window size of 10, filtered out compounds that did not have a PMI value exceeding −2, and constructed the feature set *F* for SVM. For example, in Fig. 2, we can see “只有+”, marked as cc3 in Fig. 2, has been filtered out because the PMI (“特效+”, “只有+”) value is −3.1 and no other PMI containing cc3 has value greater than −2. With the assistance of feature set *F*, we compiled D(tagged) into a bag-of-word (BOW) matrix, in which BOW[d_i, f_j] would be marked as 1 if a feature f_j existed in the sentence d_i; otherwise 0. For example, the first sentence “劇情＋緊湊＋結局＋意想不到+” in D(tagged) would become “1, 1, 0, 0, … 1, 1, 0, 0,”, because the sentence includes fc1=“劇情+”, fc2=“結局+” cc1=“結局+” cc2=“意想不到+”. Finally, we sent the BOW matrix to Linear SVM to classify the sentiment polarity of compounds to obtain our Target results for positive classifiers, negative classifiers, and positive–negative classifiers.

3.5.5 Implementation tools

This study used scikit-learn[4] (Pedregosa et al. 2011), a machine learning library used for Python, to implement linear SVM model and Natural Language Toolkit (NLTK[5]) (Bird 2006) to calculate the PMI of bigrams.

4 Experiment and results

In this section, we report three results of our experiment. First, without considering genres, the proposed method was applied to build sentiment classifiers for various portions of the training data and compared the results of the proposed method with those of TFIDF, NTUSD, and HowNet. Second, we compared the results of applying these methods to the dataset, while taking genre into account. Third, we compared the compounds obtained using the proposed method with the word-lists in NTUSD and HowNet. In each experiment, 10-fold cross-validation[6] was used for training models before making predictions for the test set. In the following, notation “+” represents a positive classifier, “-” represents a negative classifier, “±” represents positive–negative classifier, “M” represents the proposed method (morpheme-based feature-collocation pairs), “N” represents the application of NTUSD, “H” represents the application of HowNet, and “T” represents the application of TFIDF.

4.1 Opinions of movies

As shown in Fig. 1, movie rankings posted on the Yahoo! Taiwan contain more positive star-rankings (four and five stars, totaled to 89,357) than negative rankings (one and two stars, totaled to 27,209). Therefore, the

[4] Scikit-learn v0.12 http://scikit-learn.org/stable/

[5] Natural Language Toolkit 2.0 https://github.com/nltk

[6] Ten folding cross-validation is a process that chunks training dataset into 10 equal-lot of subsets, and then uses one subset for testing and others for training sequentially. Therefore, the validation process involves 10 iterations of training and testing procedures.

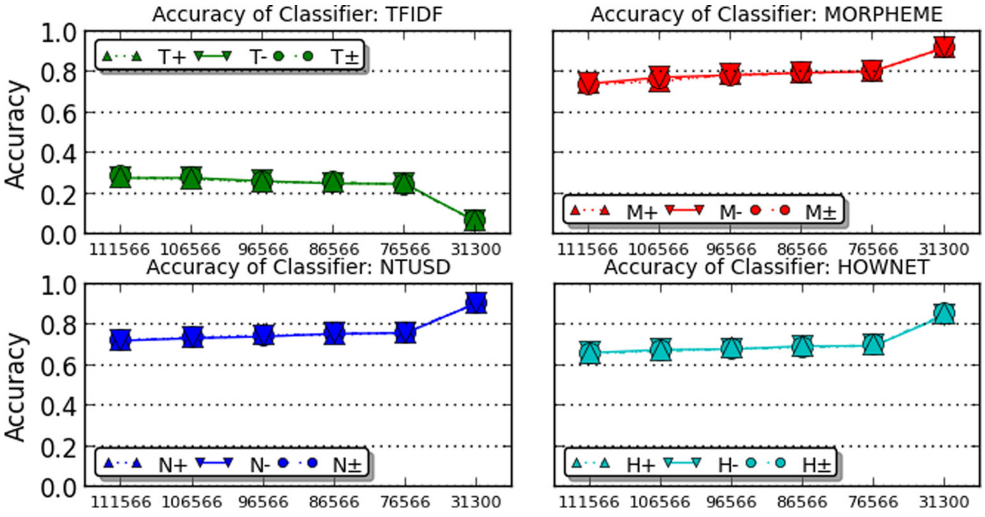


Fig. 5 Accuracy in the prediction phase

dataset is considered inherently unbalanced. Considering that the number of total negative ranking opinions was less than 22,000, the training set would remain unbalanced even if we selected more than 44,000 opinions. In the experiment, the dataset was chunked into smaller segments of 5,000, 10,000, 20,000, 30,000, and 40,000 items, for examining the efficiency of classifiers in different size of balanced training set. To test unbalanced datasets, we used 120,000 training sets to build classifiers.

As illustrated in Fig. 3, in the training phase, our morpheme-based method achieved the highest score among the methods (the average F1 approximately 0.8 in 10-fold training phrase), regardless of whether the dataset was balanced, or whether the classifier was positive, negative, or positive–negative. As shown in Fig. 4, in the prediction phase, our method still obtained a higher score than the other methods (our method archived 0.91 in average F1 score in prediction phrase) regarding all types of classifiers. The TFIDF method performed the worst due to the application of the most frequently-appearing (common) words, rather than words with significant sentiment.

We then compared the Accuracy (see Formula 2) and Balanced Accuracy (see Formula 6) of each of the methods. As shown in Fig. 5 and 6, we were unable to detect significant differences between the results from "+", "-", and "±" type classifiers; however, the proposed approach still outperformed the other methods. The average balanced accuracy rate of classifiers using TFIDF, HowNet, NTUSD, and our Morpheme-based methods were approximately 0.2, 0.5, 0.6 and 0.7, respectively.

Figure 7 presents the ratio of sentences without identifiable features for each method (our method, TFIDF, HowNet, and NTUSD). The sentences without

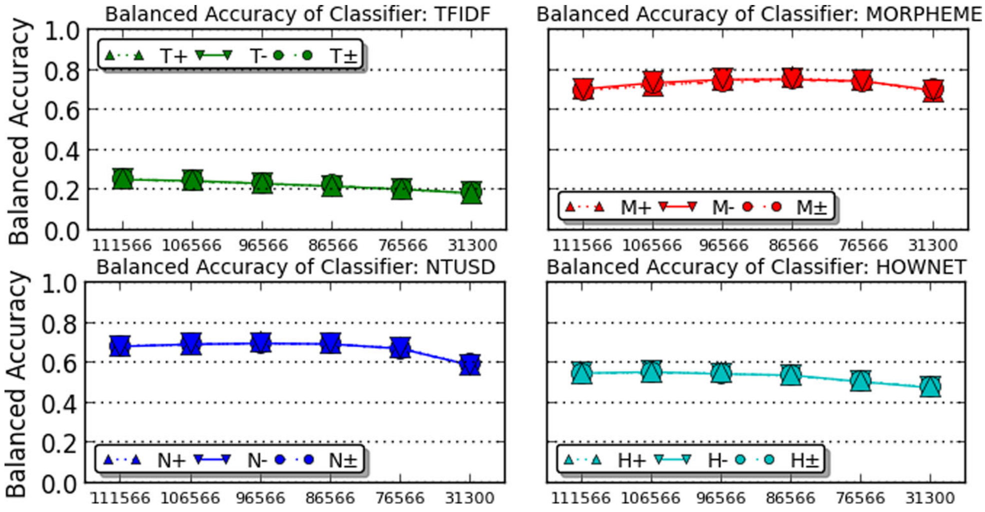


Fig. 6 Balanced accuracy in the prediction phase

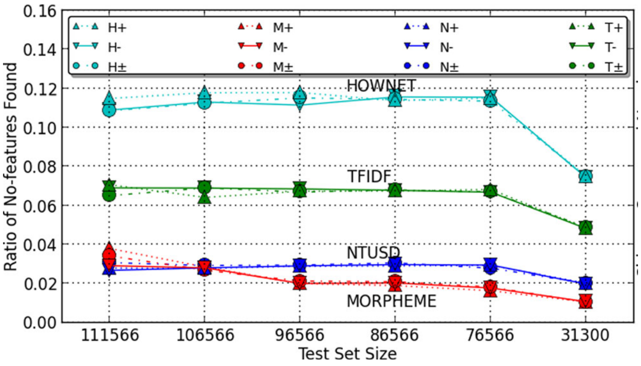


Fig. 7 Sentences without identifiable features within the test set

identifiable features are those in which we cannot find any feature with the given method. As seen in the figure, the no identifiable feature ratios of NTUSD and our morpheme-based method are lower than those of TFIDF and HowNet. The relatively high no identifiable feature ratio in HowNet is perhaps due to the differences between simplified and traditional Chinese, despite of our efforts to translate them. On the other hand, NTUSD is an effective wordlist complied by Taiwanese students (Ku et al. 2006), and it was shown to fit our test dataset well. Nevertheless, our morpheme-based method has an even lower no identifiable feature ratio than the fixed-wordlist versions of HowNet and NTUSD.

It should be noted that the feature size of NTUSD and HowNet wordlists remained constant, whereas the size of wordlists used in the TFIDF method (maximum size 8,000) and the proposed method (no size limit) would depend on the number of extracted feature words from the given training set. Fig. 8 reports the number of extracted compounds, consisting of morpheme-base features and collocations. We observed that in a relatively small training set size of 5,000 (i.e., test set size is 111,566), our morpheme-based method used only 1,200 compounds (Fig. 8) to achieve an F1 score of 0.77 (Fig. 3) and a balanced accuracy score of 0.7 (Fig. 6).

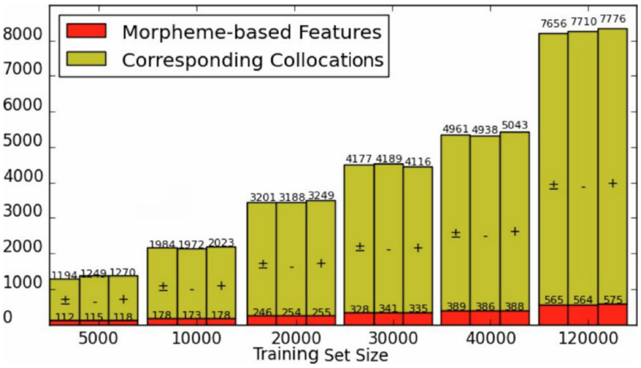


Fig. 8 Ratio of morpheme-based features and collocations in various training sets

In summary, with the size of the training set at 30,000 (i.e., test set size is 86,566) and 40,000 (i.e., test set size is 76,566), the size of compound extraction was approximately 4,100 and 5,000, respectively. All classifiers ("+", "-", "±") reported an F1 of approximately 0.82 (Fig. 4), a balanced accuracy rate of approximately 0.8 (Fig. 6), and a ratio of sentences without identifiable features of approximately 2 % in the test data sets (Fig. 7). Furthermore, in the scenario with an unbalanced training set (size 120,000) (i.e., test set size is 31,300), our morpheme-based method produced approximately 7,700 classifiable compounds and produced the best results among all methods (average F1 score of 0.79, average accuracy of 0.9, average balanced accuracy of 0.7, and average percentage of sentences without identifiable feature of 1.5 %). Compared with fixed size wordlists NTUSD and HowNet, the proposed method extracted only slightly more compounds (features and collocations); however, it suited our data sample very well.

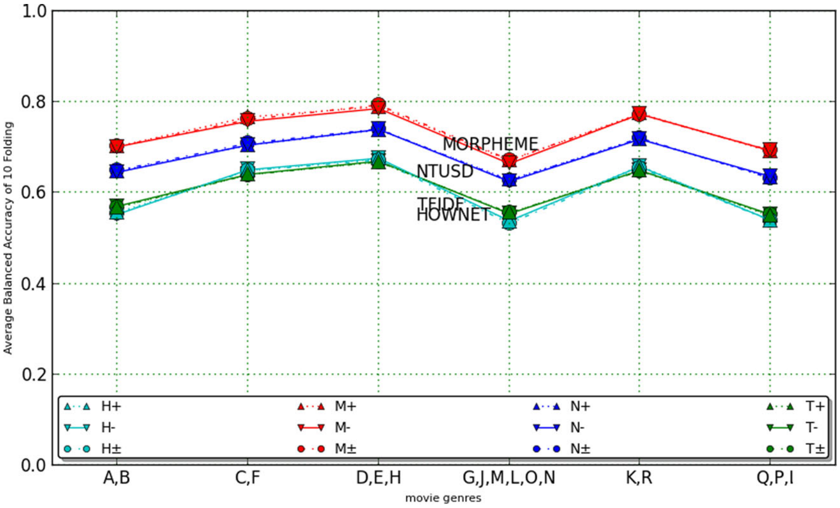
4.2 Multiple movie genres

This study compared the performance of these methods when considering multiple movie genres. We first grouped the movie genres listed in Fig. 1 into six categories, according to similarities among genres. These groups were as follows: (1) "A,B" Group was Fantasy and SciFi; (2) "C,F" Group was Crime and Actions; (3) "D,E,H" Group was Drama, Romance/Family, and Love Story; (4) "P,Q,I" Group was Animation, Comedy and Adventure; (5) "K,R" Group was Terror and Mystery/Thriller; (6) "G,J,L,M,N,O" Group was others.

The training data selected from each genre group contained a maximum of 40,000 opinions, and 10-fold validation was used for each classifier in the constructed phrases. As shown in Fig. 9, TFIDF and HowNet performed virtually the same with regard to average balanced accuracy, while our morpheme-based method outperformed both in each of genre. Furthermore, from Figs. 10, 11, 12, we can see that in spite of classifier types ("+", "-", "±"), the proposed method processed the test data more effectively than the other methods did.

One interesting observation is that group 5 (Terror and Mystery/Thriller) got the best performance; group 6 (Others) performed the worst, in each method except for TFIDF. The further analysis indicated that the sentimental compounds used in genres "K" and "R" have particularly high overlapped (the intersection is about 92 %

Fig. 9 Average balanced accuracy in each genres group in training



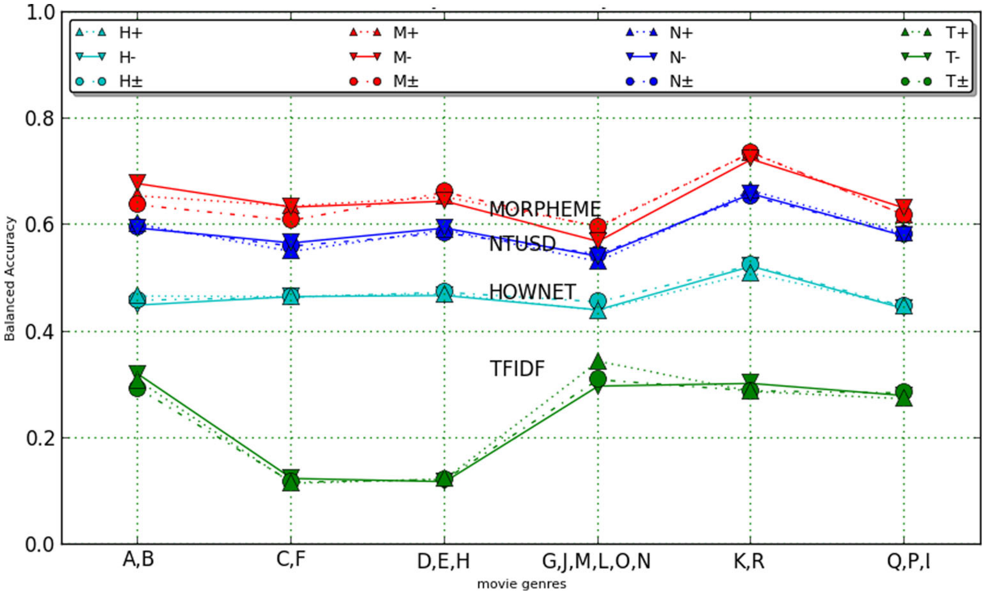
of "K"). It means that users wrote similar compounds to describe movies in group 5. But it was not the case in group 6 which the compounds are too diverse. However, in all of six groups, our method performed better than other three methods as shown in Fig. 10 (the red line).

4.3 Sentiment orientation of compounds

Figure 13 illustrates that only about 30 % of the selected compounds in our morpheme-based method appeared in the NTUSD wordlist and about 10 % appeared in the HowNet wordlist. This is possibly due to the fact that sentiment wordlists NTUSD and HowNet were not specifically designed for the analysis of movie reviews. In addition, in the ratio of overlapping sentiments as determined by NTUSD, the number of negative compounds was significantly higher than that of positive compounds; conversely, in the HowNet overlapping ratio this situation was reversed.

Table 3 presents examples of compounds, including Chinese idioms, slang, and popular terms, which were identified using the proposed method, but did not appear in NTUSD or HowNet. It demonstrates that those predefined wordlists are not suitable for the analysis of movie reviews. They are for general purpose, not for specific domains. On the contrary, our proposed method can operate in a variety of domains, such as movie

Fig. 10 Average balanced accuracy in each genres group in prediction



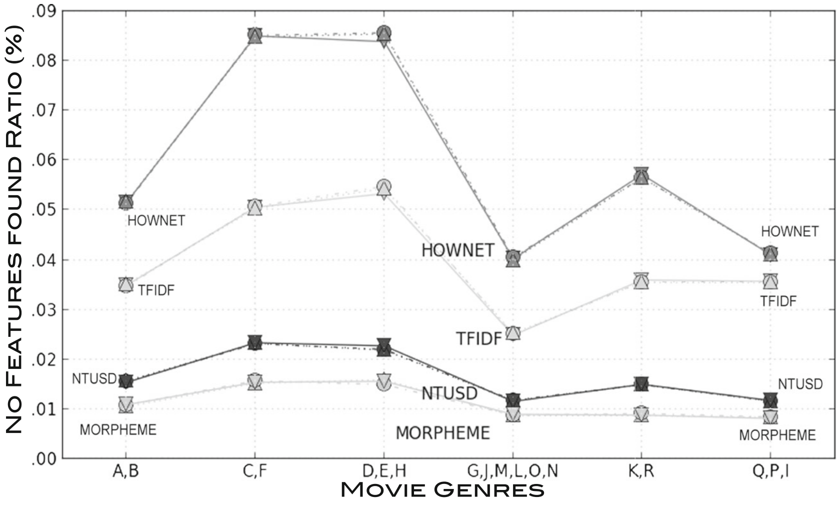


Fig. 11 Ratio of sentences without identifiable features for each genre

reviews, and provide higher accuracy in the provision of sentiment words than other methods.

Our results demonstrated the ability of the proposed method in identifying compounds with differing sentimental orientation in different genres. For example, the compound "驚悚" ("terrifying") in a common sense and in the sentiment wordlists NTUSD and HowNet would have negative sentiment. However, in Group 5 (Terror and Mystery/Thriller), our method reported that "驚悚" possesses a positive connotation of 0.24 and "不驚悚" ("not terrifying") has a negative value of −0.40. Naturally, it is necessary for a horror movie to be terrifying; viewers would be disappointed if that were not the case. Take another example: in common sense and in NTUSD and HowNet, the word "醜"("ugly") is considered to be negative, and used to describe one's appearance as hideous or unsightly. However, the proposed method attributed a positive value of 0.04 in Group 1(Fantasy&SciFi) and a positive value of 0.25 in Group 5 (Terror and Mystery/Thriller). According to the examples in Fig. 14, it is clear that this term would have positive connotations in these genres.

5 Conclusions and future research

If the approach developed for English were adopted directly, sentiment analysis in Chinese would be subject to many forms of bias. This study proposed a morpheme-based method of feature selection to search for domain-dependent Chinese compound words directly from the

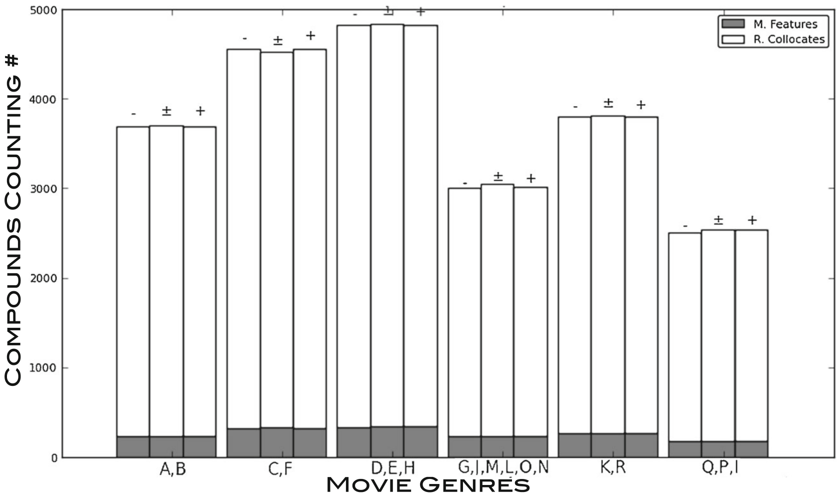


Fig. 12 Ratio of morpheme-based features and collocations in genre group training sets

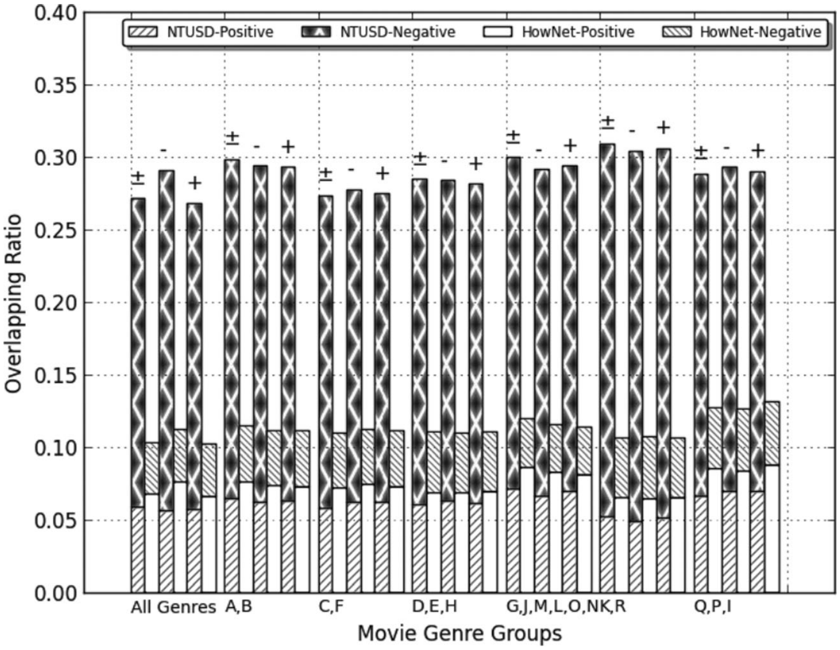
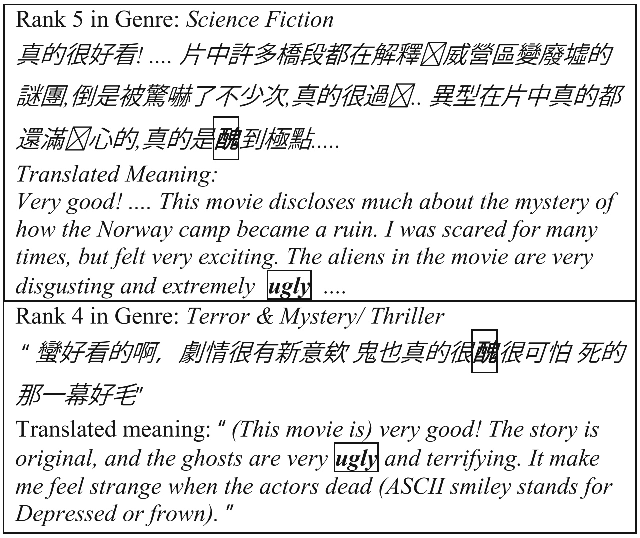


Fig. 13 Selected compounds overlapped with NTUSD and HowNet

reviews in a large data set without any help of predefined sentimental resources. Our method uses a P.O.S tagger tool for the segmentation of texts, filtering morpheme-based features and extracting appropriate collocations using relatively high PMI values to build sentiment classifiers. Results show that the proposed method is capable of achieving a higher level of balanced accuracy with small size of extracted feature and collocation compound set, providing a higher hit rate for features when new opinions are introduced. The proposed method also maintains this good performance across movie genres. Compared with pre-defined wordlists that rely on single polarity, the

Table 3 SVM weights of selected compounds in movie genre groups

Compound	Translated meaning	Over all	Fantasy & SciFi	Crime & actions	Drama, Romance/Family & Romance	Terror & Mystery/ Thriller	Animation, Comedy & Adventure	Others
爆滿	Full of people	0.23	0.26	0.43	0.21	0.35	0.24	0.25
超讚	Wonderful	0.75	0.74	0.79	0.91	0.61	0.92	0.65
目不轉睛	Unable to move avert one's eyes	1.02	0.31	0.19	0.74	0.42	0.29	0.73
感傷	Feel sorry/sad	0.74	0.25	0.52	0.46	0.29	0.14	
意猶未盡	Have not given full expression to one's views	1.17	0.41				0.98	
淋漓盡致	Extreme saturation	0.75		0.87	0.99	–		0.23
負分	Negative score	−0.99		−1.12	−0.6	–	−1.1	
亂編	Making something up	−0.54		−0.67	−0.6			−0.49
爛到爆	Terrible slag	−1.13	−1.24	−0.81	−0.74	−1.47	−0.62	
支離破碎	Scattered and smashed	−1.38	−0.49	−0.46			−0.25	
爛爆	Very bad; slag	−1.05	−1.25	−0.93	−1.14	−1.24	–	
超爛	Very bad; slag	−1.24	−0.96	−1.1	−0.48	−1.19	−1.17	−1.36
無病呻吟	Moan and groan without being ill; to complain without a cause	−1.59	−0.69		−1.17	−0.01		
雜亂無章	Disordered; in a mess	−0.91	−1.21	−1.58	−0.85		−0.45	



Rank 5 in Genre: *Science Fiction*

*真的很好看! .... 片中許多橋段都在解釋☒威營區變廢墟的謎團,倒是被驚嚇了不少次,真的很過☒.. 異型在片中真的都還滿☒心的,真的是**醜**到極點.....*

Translated Meaning:

*Very good! .... This movie discloses much about the mystery of how the Norway camp became a ruin. I was scared for many times, but felt very exciting. The aliens in the movie are very disgusting and extremely **ugly** ....*

Rank 4 in Genre: *Terror & Mystery/ Thriller*

*" 蠻好看的啊, 劇情很有新意欸 鬼也真的很**醜**很可怕 死的那一幕好毛"*

Translated meaning: *" (This movie is) very good! The story is original, and the ghosts are very **ugly** and terrifying. It make me feel strange when the actors dead (ASCII smiley stands for Depressed or frown). "*

Fig. 14 Sample Opinions including "醜" **(i.e. "ugly")**

proposed method is better able to identify the sentiment of words, which can vary in polarity according to the genre of movie. This study only took move review as an example. However, the proposed approach is domain- independent and would extract domain-dependent words from a given data set.

This study was subject to a number of limitations. Our morpheme-based method did not take into account semantics and degree of adverbs (e.g., "very" good). Future research could explore the possibility of introducing semantics and degree of sentiment into our approach. In addition, some products or services may have several aspects to be reviewed. For example, a reviewer may comment on dishes, environment, and waiter service of a restaurant, and give different scores. Future morpheme-based method may explore how to identify these different aspects from commented opinions. Finally, in future research, we could investigate the possibility of applying weights to words according to the distance from the target compounds, when employing PMI for filtering.

Acknowledgments The authors would like to thank the National Science Council, Taiwan, for financially supporting this research under Contract No. NSC 101-2410-H-004-015-MY3).

Appendixes

Appendix I movie genres

Number of total collected opinions from Yahoo!Movies Taiwan is 127,424 with 5-star-ranked opinions including 4,631,482 words in 18 movie genres. Note that one opinion might belong to one or more movie genres at the same time.

Genres	Rank1	Rank2	Rank3	Rank4	Rank5	Category
A=奇幻 Fantasy	1,870	700	979	1,433	4,137	Group 1
B=科幻 Science Fiction	2,151	857	1,403	2,323	9,657	
C=犯罪 Crime	1,116	326	535	934	2,769	Group 2
F=動作Action	7,741	2,822	4,535	7,550	27,690	
D=劇情 Drama	8,659	2,799	4,334	7,577	33,492	Group 3
E=溫馨/家庭 Romance/ Family	315	134	250	481	2,798	
H=愛情Love Story	3,628	1,078	1,599	2,833	13,455	
P=動畫 Animation	405	193	367	779	4,534	Group 6
Q=喜劇Comedy	2,097	855	1,323	2,463	8,250	
I=冒險 Adventure	4,023	1,301	1,893	2,818	8,414	
K=恐怖 Terror	3,261	712	1,019	1,454	2,734	Group 5
R=懸疑/驚悚 Mystery/ Thriller	6,343	1,881	2,825	4,456	11,056	
G=勵志 Inspiring	148	62	95	217	1,654	Group 4
J=歷史/傳記 History/ Biography	623	203	316	458	1,710	
L=戰爭War	888	306	468	826	4,189	
M=音樂/歌舞 Music/Dance	415	172	291	583	4,094	
N=紀錄片 Documentary	647	17	43	51	878	
O=武俠 Martial Arts	160	60	112	182	436	
Total: 260,720	44,490	14,478	22,387	37,418	141,947	
	17 %	6 %	9 %	14 %	54 %	
Counting	58,968		22,387	179,365		
	22.6 %		8.6 %	68.8 %		

Appendix II the pilot experiment

In a pilot experiment, we used 40,000 randomly selected opinions as training set from those pre-defined wordlists, NTUSD and HowNet, to build SVM classifiers. We kept those words their original sentiment orientation. That is, we applied positive wordlists for positive classifiers, applied negative wordlists for negative classifiers. The results show that this approach is not adequate for general purpose classifiers, because all statistical data are quite low except NTUSD negative wordlist.

Use keyword list to build SVM Model	Result
Use: HOWNET positive Model: positive Classifier # of features: 3,651	F1:0.272, Precision :0.307, Rrecall:0.244
Use: HOWNET negative Model: negative Classifier # of features: 3,036	F1:0.073, Precision:0.079, Rrecall:0.049
Use: NTUSD positive Model: positive Classifier # of features: 1,239	F1:0.067, Precision:0.074, Rrecall:0.061
Use: NTUSD negative Model: negative Classifier # of features: 4,829	F1:0.706, Precision:0.684, Recall :0.729

Appendix III exclusion list and sentence boundary

The following P.O.S tags are exclusion list since they could not have essential meaning for sentimental analysis.

1. ‘Caa’ (tagged as conjunction),
2. ‘D’, ‘DE’, ‘Dfa’ (tagged as adverb),
3. ‘Nh’ (tagged as pronoun),
4. ‘Ndaa’, ‘Ndab’,’Ndc’,’Ndd’ (tagged as time noun),
5. ’Nep’, ‘Neqa’,’Neqb’,’Nes’, ‘Neu’ (tagged as modifier),
6. ‘Nf’, ‘Nfa’, ‘Nfb’, ‘Nfc’, ‘Nfd’, ‘Nfe’, ‘Nfg’, ‘Nfh’, ‘Nfi’, (tagged as quantifier),
7. ‘T’, ‘Ta’, ‘Tb’, ‘Tc’, ‘Td’ (tagged as interjection, auxiliary word),
8. ‘V_2’ (tagged as “有”, i.e. “have” or “has”),
9. ‘SHI’ (tagged as “是”, i.e. “is” or “are”)

To determine the end-of-sentence in opinion, we use following P.O.S tags as sentence boundary:

1. “FW”,
2. “QUESTIONCATEGORY”,
3. “COLONCATEGORY”,
4. “COMMACATEGORY”,
5. “DASHCATEGORY”,
6. “ETCCATEGORY”,
7. “PARENTHESISCATEGORY”,
8. “PAUSECATEGORY”,
9. “PERIODCATEGORY”,
10. “QUESTIONCATEGORY”,
11. “SEMICOLONCATEGORY”,
12. “EXCLANATIONCATEGORY”,
13. “BR”,//HTML mark for end of sentence
14. “SPCHANGECATEGORY”

For more information of Sinica CKIP tagging, please refer to http://ckipsvr.iis.sinica.edu.tw/cat.htm

References

Bird, S. (2006). NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 69–72). Sydney, Australia.

Bradley, M. M. and P. J. Lang (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings, Technical Report C-1, *The Center for Research in Psychophysiology*, University of Florida.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics, 16*(1), 22–29.

Das, S., & Chen, M. (2001). Yahoo! for Amazon: extracting market sentiment from stock message boards. *Management Science, 53*(9), 1375–1388.

Dong, Z., & Dong, Q. (2006). HowNet and the Computation of Meaning. *World Scientific*.

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC* (Vol. 6, pp.417–422). Genoa, Italy.

Feng, S., Wang, L., Xu, W., Wang, D., & Yu, G. (2012). Unsupervised learning Chinese sentiment lexicon from massive microblog data. *Advanced Data Mining and Applications, 7713*, 27–38.

Ku, L. W., Liang, Y. T. & Chen, H. H. (2006). Opinion extraction, summarization and tracking in news and blog Corpora. *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, *AAAI Technical Report*, 100–107. CA, USA.

Ku, L. W., Liu, I. C., Lee, C. Y., Chen, K. H., & Chen, H. H. (2008). Sentence-level opinion analysis by COPEOPI in NTCIR-7. In *Proceeding of NTCIR-7 Workshop* (pp. 260–267). Tokyo, Japan.

Ku, L. W., Huang, T. H., & Chen, H. H. (2009). Using morphological and syntactic structures for Chinese opinion analysis. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Vol. 3, no.3, pp. 1260–1269). Singapore.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems, 48*(2), 354–368.

Li, L., & Yao, T. (2007, August). Kernel-based sentiment classification for Chinese sentence. In *Advanced Language Processing and Web Information Technology, ALPIT 2007. Sixth International Conference* (pp. 27–32). Henan, China.

Li, D., Ma, Y. T., & Guo, J. L. (2009). Words semantic orientation classification based on HowNet. *The Journal of China Universities of Posts and Telecommunications, 16*(1), 106–110.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2nd edition.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture* (pp. 70–77). NY, USA.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Annual Meeting-Association for computational linguistics*, 43(1). Jeju, Korea.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval, 2*(1–2), 1–135.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, É., et al. (2011). Scikit-learn: machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.

Sun, Y. T., Chen, C. L., Liu, C. C., Liu, C. L., & Soo, V. W. (2010). Sentiment classification of short Chinese sentences. *Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010)* (pp. 184–198). San Jose de Buan, Philippines.

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. *Expert Systems with Applications, 34*(4), 2622–2629.

Turney, P. D. (2001, September). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the 12th European Conference on Machine Learning* (pp. 491–502).

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics* (pp. 417–424). Freiburg, Germany.

Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworth.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York: Springer.

Wan, X. J. (2009). Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Vol. 1, pp. 235–243). Singapore.

Wang, X., Zhao, Y. Q., & Fu, G. H. (2011). A Morpheme-based Method to Chinese Sentence-Level Sentiment Classification. *International Journal of Asian Language Processing, 21*(3), 95–106. Penang, Malaysia.

Wu, Z., & Tseng, G. (1993). Chinese text segmentation for text retrieval: achievements and problems. *Journal of the American Society for Information Science, 44*(9), 532–542.

Wu, Z., & Tseng, G. (1999). ACTS: an automatic Chinese text segmentation system for full text retrieval. *Journal of the American Society for Information Science, 46*(2), 83–96.

Wu, Y., & Wen, M. (2010, August). Disambiguating dynamic sentiment ambiguous adjectives. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)* (pp. 1191–1199). Beijing, China.

Xu, H., Zhao, K., Qiu, L., & Hu, C. (2011). Expanding Chinese sentiment dictionaries from large scale unlabeled corpus. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, 3*, 53–57. Sendai, Japan.

Ye, Q., Shi,W., & Li. Y. (2006). Sentiment classification for movie reviews in Chinese by improved semantic oriented approach. *Proceedings of the 39th Hawaii International Conference on System Sciences*, HICSS'06, 3. Hawaii, USA.

Yuen, R. W., Chan, T. Y., Lai, T. B., Kwong, O. Y., & T'sou, B. K. (2004). Morpheme-based derivation of bipolar semantic orientation of Chinese words. In *Proceedings of the 20th international conference on Computational Linguistics* (pp. 1008–1014). PA, USA.

Zhang, W. H., Hua, X., & Wei, W. (2012). Weakness Finder: find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications, 39*(11), 10283–10291.

Zhou, X., Marslen-Wilson, W., Taft, M., & Shu, H. (1999). Morphology, orthography, and phonology reading Chinese compound words. *Language and cognitive processes, 14*(5–6), 525–565.

Heng-Li Yang is a professor in the Department of Management Information Systems, National Chengchi University in Taiwan. His research interests include text mining data & knowledge engineering, software engineering, knowledge management, information management in organizations, technology impacts on organizations, and empirical studies in MIS. His papers also appeared on Computers in Human Behavior, Behaviour & Information Technology, Computers & Education, Information & Management, Data and Knowledge Engineering, Online Information Review, Industrial Management & Data Systems, etc. Contact him at yanh@nccu.edu.tw.

August F. Y. Chao is a Ph.D. Candidate in the Department of Management Information Systems, National Chengchi University in Taiwan (R.O.C.). His research interests include text mining, computational linguist in Chinese, ontology application, system simulation in knowledge economics and aquaponics. Contact him at fychao.tw@gmail.com.