

Sentiment Analysis on UAV-aided Product Comments Based on Machine Learning: From Sentence to Document Level

JINGYANG CAO (✉ jyc5559@163.com)

Chongqing Jiaotong University <https://orcid.org/0000-0001-7027-905X>

Shirong Yin

Chongqing Jiaotong University

Guoxu Zhang

Chongqing Jiaotong University

Research

Keywords: Sentiment analysis, sentiment polarity of the document (SPD), sentiment polarity of a sentence (SPS), the contribution of a sentence to document (CSD)

Posted Date: November 12th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-104009/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Sentiment Analysis on UAV-aided Product Comments based on machine learning: From Sentence to Document Level

Jingyang Cao^{1*}, Shirong Yin¹ and Guoxu Zhang²

*Correspondence:
jyc5559@163.com

¹School of Mechatronics and Vehicle Engineering, Chongqing Jiaotong university, 400074, Chongqing, China
Full list of author information is available at the end of the article

Abstract

This paper presents a novel approach to analyze the sentiment of the product comments from sentence to document level and apply to the customers sentiment analysis of UAV for hotel management. In order to realize the efficient sentiment analysis, a cascaded sentence-to-document sentiment classification method is investigated. Initially, a supervised machine learning method is applied to explore the sentiment polarity of the sentence (SPS). Afterward, the contribution of the sentence to document (CSD) is calculated by using various statistical algorithms. Lastly, the sentiment polarity of the document (SPD) is determined by the SPS as well as its contribution. Comparative experiments have been established on the basis of hotel online comments, and the outcomes indicate that the proposed method not only raises the efficiency in attaining a more accurate result but also assists immensely in regards to the B5G wireless communication supported by the UAV. The findings provide a new perspective that sentence position and its sentiment similarity with document (sentiment condition) dramatically disclose the relationship between sentence and document.

Keywords: Sentiment analysis; sentiment polarity of the document (SPD); sentiment polarity of a sentence (SPS); the contribution of a sentence to document (CSD)

1 Introduction

With the development of wireless communication networks and B5G technology, the application field of UAV is wider. Many people are trying to apply UAV to hotel management, using UAV to guide customers to where they want to go. Through the communication between UAV and customers, we can provide timely and real-time feedback on hotel management and services, and understand the real needs of customers. According to Deloitte's[1] Consumer Products Group's survey, 67% of customers skim through online reviews before shopping and 82% of the total number makes purchasing decisions depending on these reviews. In order to quickly respond to customer feedback, Therefore, It is necessary to analyze the customer's emotion. The organization should examine the unstructured text of online reviews and provide the automatic classification of customers' opinions. Hence, sentiment classification has become a topic of interest for both academia and industry.

Due to the various category of media applications, existing studies deal with the sentiment classification issues at varying levels of granularity, ranging from document to sentence level. A few early attempts of sentiment classification were based

on the occurrence of the words with affective emotions in the entire document, like ‘good’ and ‘bad’. Nevertheless, these document-level analytical techniques rarely transformed the polarity of individual affective words by considering their contexts: making the derived overall polarity assessment less accurate. In order to deal with the issue, the further researches try to classify the sentiment of reviews at a better level, which is the sentence level, and aggregate sentences to forecast the documents’ sentiment polarity. Most studies on the sentence to document classification of a smart environment focus on the reviews written in English, but Chinese holds its unique way of emotional expression so that these research results of English reviews are unable to be directly applied to Chinese feedback. Thus, this paper is dedicated and focused on the enhancement of sentence to document sentiment classification for Chinese mobile cooperation online reviews. It’s believed that the SPD is determined by the sentiment polarity of the sentence and CSD. Thus, the problem is decomposed into the following fundamental subtasks:

- 1) Segment document into several sentences by a punctuation mark (full stop).
- 2) Determine the sentiment polarity of the sentence (SPS).
- 3) Determine the contribution of the sentence to document (CSD).
- 4) Determine the sentiment polarity of the document (SPD).

Following a set of reviews, the sentence-based approach offered in this paper resolves the above-mentioned task and attains a crucial precision of document sentiment classification regarding UAV intelligence scenarios. The papers’ remainder is stated as follows. We review the literature in the ‘Literature review’ section and introduce the method in the ‘Proposed Approach’ section. We conduct experiments on sentiment classification of document and evaluate the outcomes in the ‘Experiments and Evaluations’ section. We conclude the paper with possible further research in the ‘Conclusion’ section.

2 Literature Review

2.1 SENTIMENT CLASSIFICATION OF DOCUMENT LEVEL REVIEW

Most existing researches on sentiment classification were conducted towards the document level reviews, and two approaches have been utilized in these studies: machine learning and semantic orientation.

2.1.1 MACHINE LEARNING APPROACH

The machine learning approach mainly uses text classification algorithms to extract opinion and determine whether the opinion is positive or negative. There are four basic steps of this approach: feature selection, feature extraction, feature weighting and choice of classifier. This approach is efficient and fully automatic, but the machine needs be trained to learn the pattern previously.

FEATURE SELECTION Text features with certain emotional information are selected to help distinguish sentiment polarity of the text. For example, Turney[2] proposed the combinations of adjectives and adverbs as features to identify emotion. Mullen et al[3] extracted value phrases from Turney’s emotional combinations and selected features with semantic orientation based on WordNet. The experiments on movie reviews demonstrated that the hybrid features obtained superior performance. Xu et al[4]. also selected words with semantic orientation as features, and concluded that adjectives and nouns had critical impact on sentiment classification.

FEATURE EXTRACTION Some statistic methods are used to extract features, and the basic idea is to evaluate each feature and remove the features with less value than a threshold. For example, Yang and Pedersen[5] stated that Information Gain (IG) was the best method, as they adopted flat text classification method with the classifier of multiple categories to classify texts and compared the result with several other methods like K-Nearest Neighbor (KNN) and Linear Minimum Variance Match. Tang, et al[6] extracted features with Mutual Information (MI), Information Gain (IG), Chi-square Statistic (CHI) and Document Frequency (DF). The results indicated that IG performed better, by taking the category information and the influence of low-frequent features into consideration. Furthermore, Ng et al[7] compared DF, MI, IG, and CHI, and stated that CHI performed the best, DF and IG ranked higher than MI. Mladenic, et al[8]. and Zhou, et al[9]. argued that MI performed rather badly and IG performed the worst. Qin et al [10]. concluded that the best feature extraction method was MI.

FEATURE WEIGHTING Feature weighting methods mainly are Boolean Weight, Term Frequency (TF), Inverse Document Frequency (IDF), and Term Frequency-Inverse Document Frequency (TF-IDF). For example, Pang, et al [11] applied the Boolean Weight to sentiment classification and the accuracy reached 82.9%. The result indicated that the sentiment of document depends on the appearance of positive or negative features rather than the frequency of them.

CHOICE OF CLASSIFIER A large volume of research showed the dominant power that Support Vector Machine (SVM) based classifier had on sentiment classification, especially in the case of limited training samples. For example, Xia [12] and Peng employed SVM for sentiment classification of hotel reviews, and concluded that the accuracy improved as the number of reviews increased. Jiang [13] adopted SVM to classify sentiment polarity of movie reviews, and obtained the accuracy of 85.4%. Shi, et al [14]. also utilized SVM for sentiment classification of book reviews in both Chinese and English. The results demonstrated that SVM had superior performance on Chinese reviews. Pang et al [11]. compared three classifiers as NB (Naïve Bayes), ME (Maximum Entropy) and SVM, and concluded that SVM was the best. Ye, et al [15] compared SVM with NB based on reviews from travelling blogs, and stated that SVM was better than NB. Ye, et al [16]. Also compared SVM with semantic-based method, and concluded that SVM obtained better performance. Ni, et al [17].

2.1.2 Semantic Orientation Approach

The semantic orientation approach is based on identifying and extracting sentiment words and phrases, as well as semantic rules and patterns contained in the evaluation text. This approach needs no prior training, but is semi-automatic and based on external resources like corpus and lexicon.

Two types of techniques have been used in previous semantic orientation approach: corpus-based and dictionary-based techniques. The corpus-based techniques aim to find co-occurrence patterns of words to determine their sentiments. For example, Turney [19] proposed a semantic method of classification called SO-PMI (Sentiment

Orientation-Pointwise Mutual Information), and calculated a phrase's semantic orientation by the mutual information between the phrase and the word "excellent" (as the positive polarity) minus the mutual information between the phrase and the word "poor" (as the negative polarity). Riloff and Wiebe [20] used a bootstrapping process to learn linguistically rich patterns of subjective expressions, in order to classify subjective expressions from objective expressions. Zhao et al [21]. used automatic selected syntactic paths to recognize appraisal expressions for sentiment classification. Su et al [22]. proposed a novel mutual reinforcement approach to detect the hidden sentiment association of product features and opinion words, in order to improve the finer-grained opinion mining.

Dictionary-based techniques utilize synonyms, antonyms and hierarchies in WordNet (or other lexicons with sentiment information) to determine word sentiments (Andreevskaia and Bergler [23]. Building upon WordNet, SentiWordNet is a lexical resource for sentiment analysis which has more sentiment related features than WordNet (Esuli and Sebastiani [24]. SentiWordNet has been used as the lexicon in recent sentiment classification studies (Devitt and Ahmad [25] ; Denecke [26] ; Fahrni and Klenner [27]). Besides that, a small lexical resource called WordNet Affect has been developed based on WordNet, and it annotates words with six types of emotions like joy, fear, anger, sadness, disgust and surprise. Das [28] and Bandyopadhyay applied WordNet Affect to annotate the affective emotions of words and sentences for sentiment classification of documents. Furthermore, apart from the commonly-used dictionary, Tong [29] manually established an emotional dictionary of movie reviews, and used it to estimate the author's attitude towards the movie. Hu and Liu [30] created a list of seed words which are manually tagged as "positive" or "negative", and then used synonyms of seed words in WorldNet to determine the new word's polarity.

2.2 SENTIMENT CLASSIFICATION OF SENTENCE-TO-DOCUMENT LEVEL REVIEW

There are mainly two approaches to perform sentence-to-document sentiment classification: a cascaded approach and a joint approach. A cascaded approach first computes the sentiment polarity of each sentence and then inputs the result into a document-level classifier. For example, Yu [31] and Hatzivassiloglou developed several sentence-level classifiers to predict the sentiment polarity of sentences based on their likelihood ratio scores. Pang and Lee [32] used a global minimum-cut inference algorithm to identify and extract subjective sentences, and then inputted them into a document-level classifier. Meena [33] and Prabhakar represented sentences as dependency trees and considered the effect of conjunctions to analyze the sentence structures for sentiment classification. Mao [34] and Lebanon used a sequential conditional random fields (CRF) model to measure the polarity of each sentence and determine the "sentiment flow" of document. Zhao [35], Liu and Wang presented a novel method for sentence sentiment classification based on CRF, in order to reduce the "contextual dependency" and "label redundancy" in reviews. Zhang [36] et al. proposed a rule-based approach determining sentiment polarity of each sentence based on word dependency and aggregating sentences to predict the sentiment polarity of document. Das [37] and Bandyopadhyay applied a CRF-based classifier

to perform sentence level sentiment classification based on the average word level emotion scores obtained from SentiWordNet.

In contrast, a joint approach combines sentence-level and document-level classification in a single joint model to determine the overall polarity. For example, McDonald et al [38]. proposed a CRF-based structured model for sentence and document classification that used the document information to classify sentences, used the sentence information to classify documents and repeated until convergence. This integrated model was based on standard sequence classification techniques using constrained Viterbi to ensure its consistency, and it outperformed polarity classifiers trained in isolation.

3 Proposed Approach

The proposed approach in this paper follows the trend of cascaded approach but from a different perspective, which determines the contribution of sentence to document (CSD) by qualifying the relationship between sentence and document.

3.1 SENTIMENT POLARITY OF DOCUMENT (SPD)

A document c can be represented by a finite sequence of sentences, as $c = \langle s_1, s_2, \dots, s_n \rangle$, where s_i denotes a complete sentence. Thus SPD is the sum of weighted SPS. In addition, either SPD or SPS is denoted by C_{pos} (positive) and C_{neg} (negative). Eq. (1) yields the value of SPD.

$$T(c) = \sum_{i=1}^n T(s_i)w_i \quad (1)$$

- 1) Where $T(c)$ denotes the SPD of c , s_i denotes the i_{th} sentence in document c , and $T(s_i)$ denotes the SPS of s_i . If $s_i \in C_{\text{pos}}$, $T(s_i) = 1$; If $s_i \in C_{\text{neg}}$, $T(s_i) = -1$. w_i denotes the CSD of s_i . $T(c) > 0$ means that c is positive, otherwise c is negative.

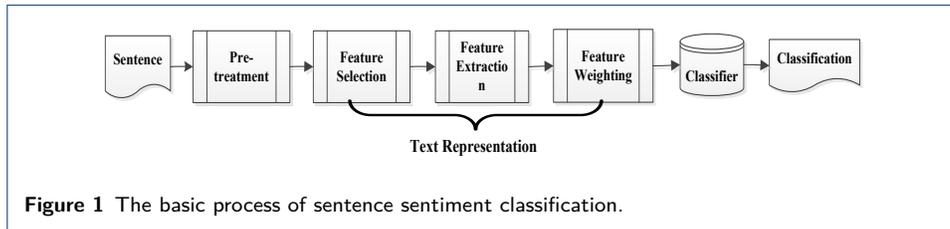
According to the ways of expression in Chinese, opinion is always stated clearly at first, then illustrations are made to support the view, and conclusion is drawn in the end. Take the following piece of hotel online review for example, ‘Fantastic hotel! It’s located in a quiet place. The room is big and clean. The bed is comfortable for sleeping and the receptionists greet me whenever I go in and out. I love it!’. Thus the CSD differs due to its position in the document. Accordingly, Eq. (1) is extended to Eq. (2) as follows:

$$T(c_i) = \begin{cases} T(s_{i1})(w_F + w_E)/2 & n = 1 \\ T(s_{i1})w_F + T(s_{i2})w_E & n = 2 \\ T(s_{i1})w_F + \sum_{j=2}^{n-1} T(s_{ij})w_M + T(s_{in})w_E & n \geq 3 \end{cases} \quad (2)$$

- 2) Where s_{i1} denotes the first sentence, s_{i2}, \dots, s_{in-1} denote the middle sentences, s_{in} denotes the last sentence, w_F denotes the CSD of the first sentence, w_M denotes the CSD of the middle sentence, and w_E denotes the CSD of the last sentence.

3.2 SENTIMENT POLARITY OF SENTENCE (SPS)

Labeling the position of each sentence in a document, and applying a supervised machine learning method to determine the SPS based on statistical natural language processing (SNLP). The basic process is shown in Fig. 1.



At first, the sentence is split into words and words are annotated with certain part-of-speech tags (whether the word is a noun, verb, adjective, etc). Then, adjectives, adverbs, and verbs are selected as initial text features. It should be noted that negative words like “not” is considered as emotion changer, thus this paper identifies negations with a manually prepared list of negative words and combines them with the adjectives or adverbs behind as a feature. For example, if there is a “not” before “pretty”, the feature would be “not pretty”. After that, Information Gain (IG) is utilized to calculate the statistic value of each feature, and features with the values above thresholds are extracted. At last, features are weighted by Boolean value (1 or 0) and inputted into a Support Vector Machine (SVM) classifier. An example of sentiment classification of sentence is shown in Table 1, and the adjectives, adverbs and verbs are text features. The words in bold are significant in sentiment classification and are among the best candidate features in feature set.

Table 1 An example of sentence sentiment classification

Document	Sentence Segmentation	Sentence Position	SPS	Adjectives	Adverbs	Verbs
It's been a while, but I still remember that it's super near the train station, rather convenient but inevitably noisy. There had been a lot of Korean-Japanese tourists in this hotel, and the receptionists were inhospitable. We had two people registered for a standard room, but they only gave one room card with the aggressive look in their eyes. I was so upset that lost my mood.	It's been a while, but I still remember that it's super near the train station, rather convenient but inevitably noisy.	The first sentence	Positive	near	super	
				convenient	rather	
				noisy	inevitably	
	There had been a lot of Korean-Japanese tourists in this hotel, and the receptionists were inhospitable. We had two people registered for a standard room, but they only gave one room card with the aggressive look in their eyes.	The middle sentences	Negative	inhospitable	only	register
				standard		
				aggressive		
				cheap		
	I was so upset that lost my mood.	The last sentence	Negative	upset		lost

3.3 CONTRIBUTION OF SENTENCE TO DOCUMENT (CSD)

The contribution of sentence to document (CSD) is determined by the relationship between sentence and document, thus three statistical methods are introduced in this paper to measure the relationship (w_i), which are Equal Weight, Correlation Degree and Sentiment Condition Probability.

3.3.1 Equal Weight

It is assumed that each sentence has the same weight in document, and SPD is simply the sum of SPS. Thus w_i is shown in Eq. (3).

$$w_F = w_M = w_E \quad (3)$$

Where F denotes the first sentence, M denotes the middle sentence and E denotes the last sentence.

3.3.2 Correlation Degree

It is assumed that w_i is determined by the probability that the sentence s_i and the document c share the same sentiment. Thus w_i is shown in Eq. (4).

$$w_i = P((c \in C_{\text{pos}} \cap s_i \in C_{\text{pos}}) \cup (c \in C_{\text{neg}} \cap s_i \in C_{\text{neg}})) \quad (4)$$

3.3.3 Sentiment Condition Probability

The basic intuition of this method is that a sentence positively influences a document with the same sentiment and negatively influences a document with the opposite sentiment. In other words, a positive sentence is more likely to indicate a positive document and a negative sentence is more likely to indicate a negative document. Hence sentiment polarity of a given sentence s_i is used as a precondition, and w_i is a conditional probability that document c has the same sentiment as s_i does. Thus w_i shown in Eq. (5).

$$w_i = \begin{cases} P(c \in C_{\text{pos}} | s_i \in C_{\text{pos}}) \\ P(c \in C_{\text{neg}} | s_i \in C_{\text{neg}}) \end{cases} \quad (5)$$

These methods above estimate the relationship between sentence and document. Nevertheless, the position influence on the relationship has been neglected, making them less reliable. Especially in Chinese, it is assumed that the contributions of the first and the last sentences are greater than that of the middle sentences. Therefore, position information has to be added to the CSD algorithms.

In this paper, the first sentence is denoted by F, the middle sentences are denoted by M, and the last sentence is denoted by E. Besides that, an identifier demonstrating sentence position, SPS and SPD has been employed, where the first letter denotes the sentence position, the second denotes the sentence polarity, and the third denotes the document polarity. For example, MPP indicates a positive sentence located in the middle of a positive document. All identifiers are shown in Table 2.

Either the first sentence or the last sentence is necessary for any document, thus the number of positive documents is denoted as $L_{\text{FPP}} + L_{\text{FNP}}$ (or $L_{\text{EPP}} + L_{\text{ENP}}$), while the number of negative documents is denoted as $L_{\text{FPN}} + L_{\text{FNN}}$ (or $L_{\text{EPN}} + L_{\text{ENN}}$).

By taking position information into consideration, Eq. (4) is extended to Eq. (6) - Eq. (8) and Eq. (5) is extended to Eq. (9) - Eq. (11).

Table 2 Identifier demonstrating sentence position, SPS and SPD

Sentence Position	SPS	Positive Document	Negative Document
First Sentence	Positive	FPP	FPN
	Negative	FNP	FNN
Middle Sentences	Positive	MPP	MPN
	Negative	MNP	MNN
Last Sentence	Positive	EPP	EPN
	Negative	ENP	ENN

$$w_F = P(c \in C_{\text{pos}} \cap s_1 \in C_{\text{pos}}) + P(c \in C_{\text{neg}} \cap s_1 \in C_{\text{neg}}) = \frac{L_{\text{FPP}} + L_{\text{FNN}}}{L_{\text{FPP}} + L_{\text{FNP}} + L_{\text{FPN}} + L_{\text{FNN}}} \quad (6)$$

$$w_M = P(c \in C_{\text{pos}} \cap s_1 \in C_{\text{pos}}) + P(c \in C_{\text{neg}} \cap (s_2, \dots, s_{n-1}) \in C_{\text{neg}}) = \frac{L_{\text{MPP}} + L_{\text{MNN}}}{L_{\text{MPP}} + L_{\text{MNP}} + L_{\text{MPN}} + L_{\text{MNN}}} \quad (7)$$

$$w_E = P(c \in C_{\text{pos}} \cap s_n \in C_{\text{pos}}) + P(c \in C_{\text{neg}} \cap s_n \in C_{\text{neg}}) = \frac{L_{\text{EPP}} + L_{\text{ENN}}}{L_{\text{EPP}} + L_{\text{ENP}} + L_{\text{EPN}} + L_{\text{ENN}}} \quad (8)$$

$$w_F = \begin{cases} \frac{P(c \in C_{\text{pos}} \cap s_1 \in C_{\text{pos}})}{P(s_1 \in C_{\text{pos}})} = \frac{L_{\text{FPP}}}{L_{\text{FPP}} + L_{\text{FPN}}} \\ \frac{P(c \in C_{\text{neg}} \cap s_1 \in C_{\text{neg}})}{P(s_1 \in C_{\text{neg}})} = \frac{L_{\text{FNN}}}{L_{\text{FNP}} + L_{\text{FNN}}} \end{cases} \quad (9)$$

$$w_M = \begin{cases} \frac{P(c \in C_{\text{pos}} \cap (s_2, \dots, s_{n-1}) \in C_{\text{pos}})}{P((s_2, \dots, s_{n-1}) \in C_{\text{pos}})} = \frac{L_{\text{MPP}}}{L_{\text{MPP}} + L_{\text{MPN}}} \\ \frac{P(c \in C_{\text{neg}} \cap (s_2, \dots, s_{n-1}) \in C_{\text{neg}})}{P((s_2, \dots, s_{n-1}) \in C_{\text{neg}})} = \frac{L_{\text{MNN}}}{L_{\text{MNP}} + L_{\text{MNN}}} \end{cases} \quad (10)$$

$$w_E = \begin{cases} \frac{P(c \in C_{\text{pos}} \cap s_n \in C_{\text{pos}})}{P(s_n \in C_{\text{pos}})} = \frac{L_{\text{EPP}}}{L_{\text{EPP}} + L_{\text{EPN}}} \\ \frac{P(c \in C_{\text{neg}} \cap s_n \in C_{\text{neg}})}{P(s_n \in C_{\text{neg}})} = \frac{L_{\text{ENN}}}{L_{\text{ENP}} + L_{\text{ENN}}} \end{cases} \quad (11)$$

4 Experiments and Evaluations

4.1 EXPERIMENT DESIGN

The experiment is designed based on hotel online reviews, and three major steps are included. Firstly, a document is segmented into sentences labeled with their positions. Then a supervised machine learning method, including feature selection, feature extraction, feature weighting and choice of classifier, is employed to determine the SPS. After that, three statistical methods are applied to determine the CSD. And at last, the SPD is calculated based on SPS and CSD in Eq. (2). The process of sentiment classification from sentence to document is shown in Fig. 2.

4.2 CORPUSES

In order to avoid the influence that corpus from different e-commerce websites have on the experiment, we collect hotel online reviews from two famous travel websites in China, *ctrip.com* and *qunar.com*. To keep the balance between positive and negative corpus, 1500 positive documents and 1500 negative documents are collected from *ctrip.com* as the training corpus, 500 positive documents and 500 negative documents are collected from *qunar.com* as the testing corpus. After the corpus being manually preprocessed, the irrelevant and repeated reviews are removed, and about 82% of hotel reviews are kept for experiment.

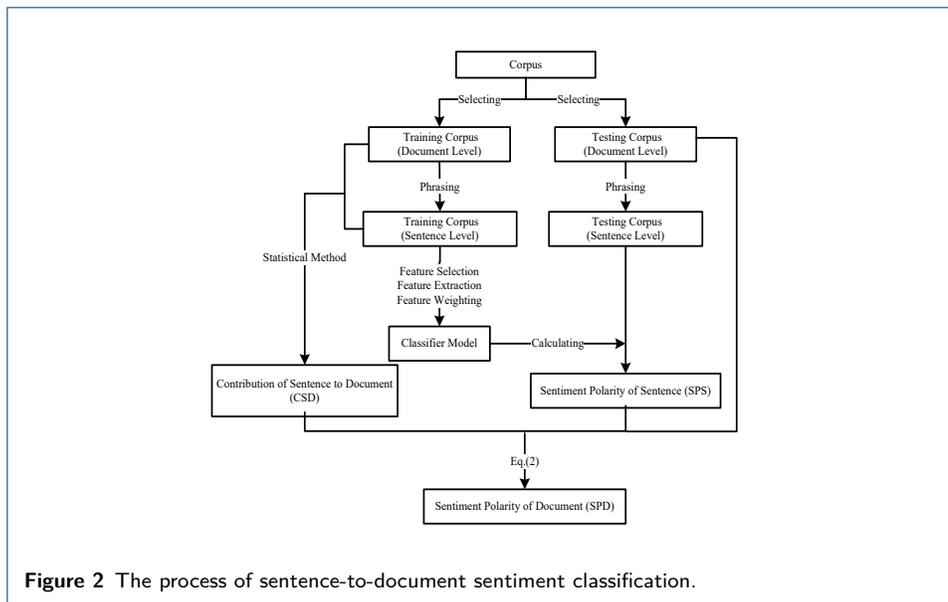


Figure 2 The process of sentence-to-document sentiment classification.

Two researchers manually label the sentiment polarity of each document and the sentiment polarity of each sentence in the training corpus. In order to reduce the subjectivity deviations in the labeling process, 20 pieces of corpus are selected randomly and the value of statistic Kappa are computed to test the consistency of labeling results. The Kappa value is 0.72, higher than 0.7, demonstrating an acceptably stable result. The labeling results of training corpus are shown in Table 3 and Table 4.

Table 3 Labeling results of document-level reviews in training corpus

Item	Result
Amount of reviews	2461
Amount of positive reviews	1195
Amount of negative reviews	1266
Amount of sentences in all documents	11062
Amount of positive sentences in all documents	5349
Amount of negative sentences in all documents	5713

Table 4 Labeling results of sentence-to-document level reviews in training corpus

Sentence Position	SPS	Positive Document	Negative Document	Amount
First Sentence	Positive	1078	80	1158
	Negative	117	1186	1303
Middle Sentences	Positive	2862	314	3176
	Negative	774	2190	2964
Last Sentence	Positive	943	72	1015
	Negative	252	1194	1446

4.3 EXPERIMENT ONE: CALCULATING CSD BY DIFFERENT ALGORITHMS

Inputting the labeling results in Table 4 into Eq. (3) and Eq. (6) – Eq. (11), the results of CSD determined by different algorithms are shown in Table 5.

Where the superscript (*pos*) denotes that a positive sentence is given as the sentiment condition and (*neg*) denotes that a negative sentence is given as the sentiment

Table 5 CSD Calculated by Different Algorithms

Sentence Position	SPS	Equal Weight	Correlation Degree	Sentiment Condition Probability
First Sentence	Positive	WF=33.3%	WF=92.0%	WF(pos)=93.1%
	Negative			WF(neg)=91.0%
Middle Sentences	Positive	WM=33.3%	WM=82.3%	WM(pos)=90.1%
	Negative			WM(neg)=73.9%
Last Sentence	Positive	WE=33.3%	WE=86.8%	WE(pos)=92.9%
	Negative			WE(neg)=82.6%

condition. The result of Correlation Degree demonstrates that sentence position has great impact on CSD, as the first sentence is the highest, which is over 90%, the last sentence follows and the middle sentences rank the last. It is consistent with our previous assumption and further illustrates that the first sentence contributes the most to the sentiment of document compared to the rest. Since review doesn't necessarily end with a conclusion, the contributions of the last sentence and the middle sentences are close to each other. Evidence can be found in hotel online reviews to support the conclusions above, and here is an example.

“The hotel is great! It has large room and new equipment. It is newly decorated. The receptionists are polite and nice. I will definitely try again!”

The result of Sentiment Condition Probability shows that the importance of positive sentence is higher than that of negative sentence in all positions. This indicates that positive sentence is more powerful to distinguish the sentiment polarity of document, especially in the middle position. The reason for this may be that most consumers are critical and they always point out flaws even when praising the object, hence make the negative sentence less capable of predicting the sentiment polarity of document.

4.4 EXPERIMENT TWO: SENTIMENT CLASSIFICATION OF SENTENCE AND DOCUMENT

As mentioned in Section 2, a supervised machine learning method is introduced to classify the sentiment polarity of sentence. At first, the ICTCLAS System developed by the Institute of Computing Technology of Chinese Academy is introduced to do the word segmentation and POS tagging. Then, the threshold value of IG is set at 0.000772297 and 300 text features including adjectives, verbs and adverbs are extracted. After that, text features are transformed into 1-0 format as text vectors, where 1 means the text feature appears in the sentence and 0 means otherwise. At last, text vectors are inputted into a SVM classifier, and 1 or -1 is assigned to each sentence, representing the SPS.

After sentiment classification of sentence, inputting SPS and CSD from Table 5 into Eq. (2), sentiment classification of 10 pieces of documents is shown in Table 6.

Where mark “?” suggests that the SPD is unable to be decided when the value is 0. Accuracy of document sentiment classification with these three CSD algorithms are shown in Table 7.

From Table 7, it is observed that the accuracy of Correlation Degree and Sentiment Condition Probability are much higher than that of Equal Weight illustrating that sentence position and its sentiment similarity with document greatly disclose

Table 6 Sentiment Classification of Sentence and Document

No.	SPD*	SVM Classifier			Equal Weight		Correlation Degree		Assumption of Sentiment Condition	
		SPS _F	SPS _M	SPS _E	Value	SPD	Value	SPD	Value	SPD
Pos1	1	1			1	1	0.921	1	0.926	1
Pos2	1	1	1		2	1	1.731	1	1.8272	1
Pos3	1	1		1	2	1	1.78	1	1.861	1
Pos4	1	-1		1	0	?	-0.062	-1	0.0183	1
Pos5	1	1	2	1	4	1	3.4	1	3.6634	1
Neg1	-1	1			1	1	0.9	1	0.926	1
Neg2	-1	-1	-1	-1	-3	-1	-2.4	-1	-2.4737	-1
Neg3	-1	1	-1	-1	-1	-1	-0.748	-1	-0.631	-1
Neg4	-1	-1	0	1	0	?	-0.062	-1	0.0183	1
Neg5	-1	-1		-1	-2	-1	-1.7	-1	-1.7337	-1

Table 7 Accuracy of document sentiment classification with different CSD algorithms

Algorithms	Positive	Negative	Accuracy
Equal Weight	81.9%	81.4%	81.6%
Correlation Degree	90.1%	87.3%	88.6%
Sentiment Condition Probability	93.5%	86.0%	90.0%

the relationship between sentence and document. Table 7 also demonstrates that Sentiment Condition Probability performs better than Correlation Degree in general. This is because that Correlation Degree only measures the co-occurrence of sentence and document with the same sentiment, while Sentiment Condition Probability actually estimates how a sentence influence a document with the same sentiment, hence more accurately indicates the importance and significance of sentence to document. Therefore, Sentiment Condition Probability is chosen to calculate CSD. In addition, the accuracy of sentiment classification on positive corpus is higher than that on the negative one in all three cases. The reason is that negative reviews are usually longer than the positive ones and use sentences instead of sentiment words to express opinions, thus the sentiment of negative reviews is implicit and hidden in complicated expressions making text features are fuzzy and difficult to extract and lowering the accuracy of sentiment classification.

4.5 Experiment Three: Comparative Experiment

Comparative experiment has been conducted based on two baselines and the approach proposed in this paper.

- 1) *A joint approach based structured model*: At first, a sequential conditional random fields (CRF) model is utilized to measure the sentiment polarity of each sentence. Then, the sentence information is passed to the document classifier also based on CRF as input. At last, the document information is passed as input into the sentence classifier and repeat until convergence.
- 2) *A cascaded approach based on subjective sentence*: At first, a global minimum-cut inference algorithm is utilized to identify and extract subjective sentences, then the sentiment polarity of the identified subjective sentences are classified, and at last, these sentences are inputted into a document-level polarity classifier based on SVM.

Fig. 3 shows the comparative performances between the two baselines and our approach, which shows a remarkable progress in enhancing the accuracy of document

sentiment classification from joint approach to cascaded approach and a significant improvement in obtaining a more accurate result in particular achieved by our approach.

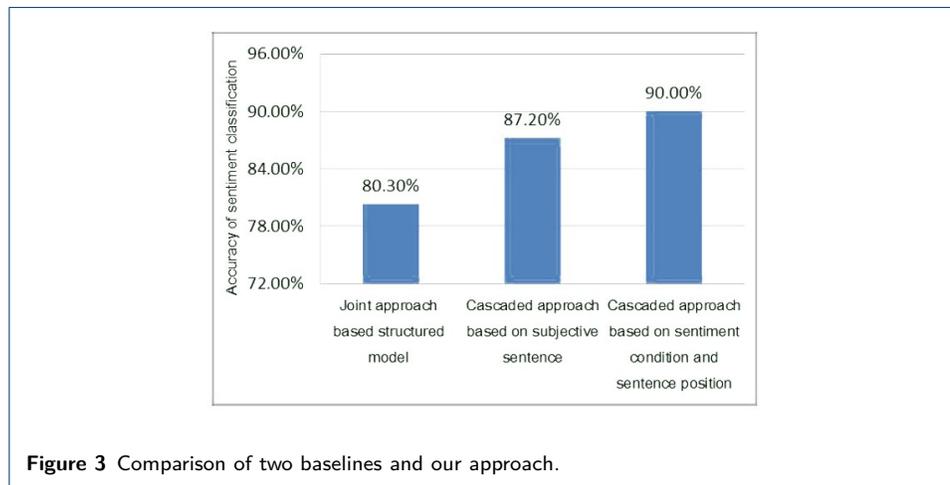


Fig. 3 illustrates that the first baseline has a much lower accuracy, for the training data sets in our experiment are not sufficient enough to train the structured model at both sentence and document level. In contrast, our approach performs much better in the case of limited training sample.

Besides that, Fig. 3 also demonstrates that the baseline cascaded approach is outperformed by our approach. This is because that the second baseline only uses subjective sentence for sentiment classification of document, thus the result will be greatly affected by the accuracy of identifying and extracting subjective sentence. In Chinese online reviews, some expressions being identified as objective sentence by the minimum-cut inference algorithm may actually convey subjective opinions. For example, “*the room is very big*” indicates that the customer is satisfied with the size of room, thus expresses a subjective and positive sentiment. On the contrary, such sentence is classified correctly and contributes to the sentiment classification of document in our approach, leading to a more efficient and effective result.

5 Results and Discussion

The method proposed in this paper firstly computes the sentiment polarity of sentence by a supervised machine learning approach, then sets weight to each sentence based on the relationship between sentence and document, and at last, aggregates the weighted sentiment polarities of sentences to predict the sentiment polarity of document. Comparative experiments have been made to test and verify the effectiveness that the proposed approach has in obtaining a more accurate result. And the experimental results indicate the following conclusions:

- 1) Sentiment polarity of document (SPD) is determined by the sentiment polarity of sentence (SPS) and the contribution of sentence to document (CSD).
- 2) Sentence position and its sentiment similarity with document (sentiment condition) effectively disclose the relationship between sentence and document, and are hereby essential for determining CSD. Specifically, the first sentence

often states the general opinion of document, thus contributes the most to the sentiment polarity of document, while the last sentence follows and the middle sentences rank the last; the contribution of positive sentence to document is higher than negative one in spite of sentence position, and the contribution of negative sentence to document are greatly depended on their positions. In other words, the ways of expression in Chinese has great impact on sentiment classification of Chinese online reviews.

- 3) Cascaded approach performs much better than joint approach with limited training data set.

But the research method in this paper has risks in the sentiment analysis of objective sentences and has limitations in extracting hidden information in sentences.

Discussion:

- 1) Sentences with a neutral sentiment (neither positive nor negative) or without a sentiment (such as objective sentences) and its contribution to the sentiment of document will be discussed.
- 2) The critical point of classification in this paper is 0, which is determined by common sense. In the future, a document-level SVM classifier will be applied by taking the polarity of sentence as input vector.
- 3) The extraction of hidden relations between product features and opinions in reviews will be discussed in the future research, for document should be divided by product features instead of punctuation mark.

Availability of data and materials

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Competing interests

The authors declare that we have no competing interests.

Funding

This work was supported by Chongqing Yongchuan District Social Livelihood Science and Technology Special Project (Ycstc. 20180301).

Authors Contribution

JYC carried out the Algorithms studies, participated in the sequence alignment and drafted the manuscript. GXZ participated in the design of the study and performed the statistical analysis. SRY conceived of the study, and participated in its design and coordination and helped to draft the manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

SPS	Sentiment polarity of the sentence
CSD	Contribution of the sentence to document
SPD	Sentiment polarity of the document
TF	Term frequency
IDF	Inverse Document Frequency
TF-IDF	Term frequency-inverse document frequency
SVM	Support Vector Machine
CRF	Conditional random fields
SNLP	Statistical natural language processing
POS	Positive
NEG	Negative

Author details

¹School of Mechatronics and Vehicle Engineering, Chongqing Jiaotong university, 400074, Chongqing, China.

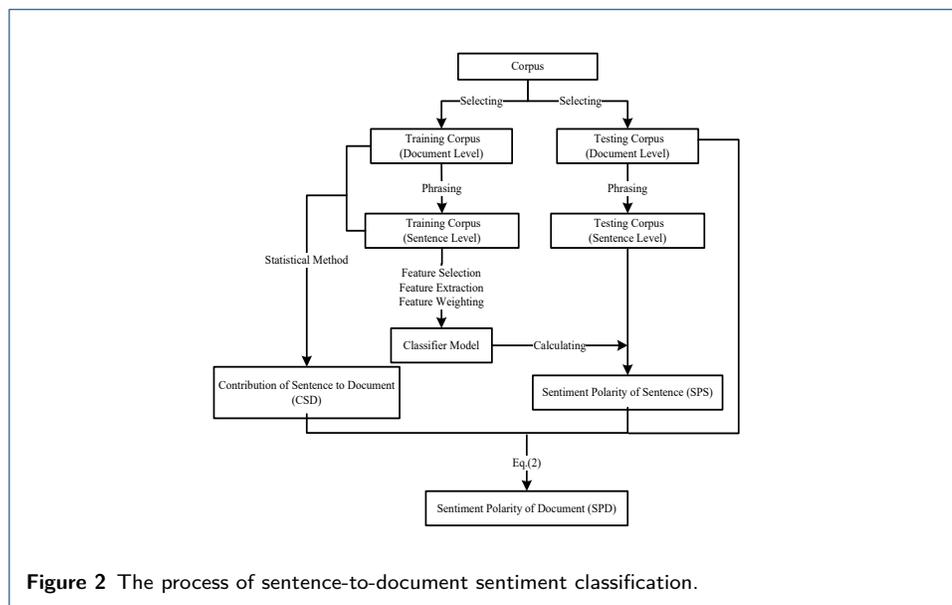
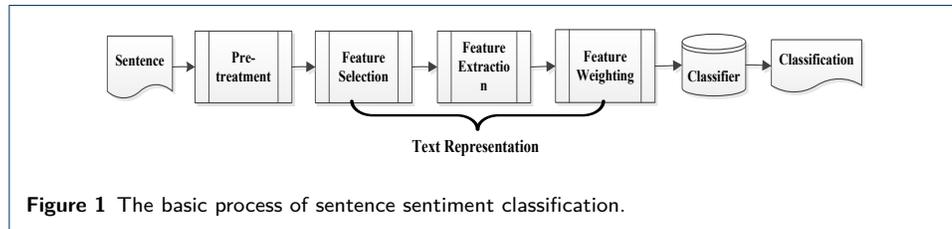
²School of Economics and Management, Chongqing Jiaotong university, 400074, Chongqing, China.

References

1. Deloitte Services LP: New Deloitte Study Shows Inflection Point for Consumer Products Industry; Companies Must Learn to Compete in a More Transparent Age (2007). <https://www.prnewswire.com/news-releases/58235327.html> Accessed 9 Aug 2012
2. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424 (2019)
3. Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources. In: The 2004 Conference on Empirical Methods in Natural Language Processing, pp. 412–418 (2004)
4. Xu, J., Ding, Y.-X., Wang, X.-L.: Sentiment classification for chinese news using machine learning methods. *Journal of Chinese Information Processing* **21**(6), 95–100 (2007)
5. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: 14th International Conference on Machine Learning (ICML 97), vol. 97, p. 35 (1997). Nashville, TN, USA
6. Tang, H.-F., Tan, S.-B., Cheng, X.-Q.: Research on sentiment classification of chinese reviews based on supervised machine learning techniques. *Journal of Chinese information processing* **21**(6), 88–126 (2007)
7. Ng, H.T., Goh, W.B., Low, K.L.: Feature selection, perceptron learning, and a usability case study for text categorization. In: 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 97), pp. 67–73 (1997)
8. Grobelnik, M.: Feature selection for unbalanced class distribution and naive bayes. In: 16th International Conference on Machine Learning (ICML 99), pp. 258–267 (1999). Citeseer
9. Zhou, X., Wang, M., Hu, M.: Research on feature-selection in chinese text classification. *Journal of Chinese Information Processing* **18**(3) (2015)
10. Qin, J., Chen, X., Wang, W., Lu, N.: Feature extraction in text categorization. *Computer Applications* **23**(2), 45–46 (2003)
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: The 2002 Conference on Empirical Methods in Natural Language Processing, pp. 79–86 (2002)
12. Xia, H., Peng, L.: Svm-based comments classification and mining of virtual community: For case of sentiment classification of hotel reviews. In: The 2009 International Symposium on Intelligent Information Systems and Applications (IISA 2009), pp. 507–511 (2009). Citeseer
13. Li, J.: An approach of sentiment classification using svm for chinese texts. In: The 2006 International Conference on Artificial Intelligence (ICAI 06), pp. 759–761 (2006)
14. Shi, W., Qi, G., Meng, F.: Sentiment classification for book reviews based on svm model. In: The 2005 International Conference on Management Science & Engineering, pp. 214–217 (2005)
15. Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications* **36**(3), 6527–6535 (2009)
16. Ye, Q., Lin, B., Li, Y.-J.: Sentiment classification for chinese reviews: A comparison between svm and semantic approaches. In: The 2005 International Conference on Machine Learning and Cybernetics, vol. 4, pp. 2341–2346 (2005). IEEE
17. Ni, X., Xue, G.-R., Ling, X., Yu, Y., Yang, Q.: Exploring in the weblog space by detecting informative and affective articles. In: The 16th International Conference on World Wide Web, pp. 281–290 (2007)
18. Zhang, Z., Ye, Q., Li, Y., Law, R.: Sentiment classification of online cantonese reviews by supervised machine learning approaches. *International Journal of web engineering and technology* **5**(4), 382–397 (2009)
19. Liu, B.: *Web Data Mining*. Springer, ??? (2007)
20. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: The 2003 Conference on Empirical Methods in Natural Language Processing, pp. 105–112 (2003)
21. Zhao, Y., Qin, B., Che, W., Liu, T.: Appraisal expression recognition with syntactic path for sentence sentiment classification. *International Journal of Computer Processing of Languages* **23**(01), 21–37 (2011)
22. Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., Swen, B., Su, Z.: Hidden sentiment association in chinese web opinion mining. In: The 17th International Conference on World Wide Web, pp. 959–968 (2008)
23. Andreevskaia, A., Bergler, S.: Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In: 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)
24. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: 5th Conference on Language Resources and Evaluation, vol. 6, pp. 417–422 (2006). Citeseer
25. Devitt, A., Ahmad, K.: Sentiment polarity identification in financial news: A cohesion-based approach. In: 45th Annual Meeting of the Association of Computational Linguistics, pp. 984–991 (2007)
26. Denecke, K.: Using sentiwordnet for multilingual sentiment analysis. In: IEEE 24th International Conference on Data Engineering Workshop (ICDEW 2008), pp. 507–512 (2008). IEEE
27. Fahrni, A., Klenner, M.: Old wine or warm beer: Target-specific sentiment analysis of adjectives. In: AISB 2008 Symposium on Affective Language in Human and Machine, pp. 60–63. University of Zurich, ??? (2008)
28. Das, D., Bandyopadhyay, S.: Document level emotion tagging: machine learning and resource based approach. *Computación y Sistemas* **15**(2), 221–234 (2011)
29. Tong, R.M.: An operational system for detecting and tracking opinions in on-line discussion. In: Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification, vol. 1 (2001)
30. Hu, M., Liu, B.: Mining and summarizing customer reviews, pp. 168–177 (2004)
31. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: The 2003 Conference on Empirical Methods in Natural Language Processing, pp. 129–136 (2003)
32. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, pp. 271–278 (2004)
33. Meena, A., Prabhakar, T.V.: Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In: European Conference on Information Retrieval (ECIR), pp. 573–580 (2007). Springer
34. Mao, Y., Lebanon, G.: Isotonic conditional random fields and local sentiment flow. In: The 2000 Neural

- Information Processing Systems Conference (NIPS), pp. 961–968 (2007)
35. Zhao, J., Liu, K., Wang, G.: Adding redundant features for crfs-based sentence sentiment classification. In: The 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 117–126 (2008)
 36. Zhang, C., Zeng, D., Li, J., Wang, F.-Y., Zuo, W.: Sentiment analysis of chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology* **60**(12), 2474–2487 (2009)
 37. Das, D., Bandyopadhyay, S.: Sentence level emotion tagging on blog and news corpora. *Journal of Intelligent Systems* **19**(2), 145–162 (2010)
 38. McDonald, R., Hannan, K., Neylon, T., Wells, M., Reynar, J.: Structured models for fine-to-coarse sentiment analysis. In: The 45th Annual Meeting of the Association of Computational Linguistics, pp. 432–439 (2007)

Figures



Tables

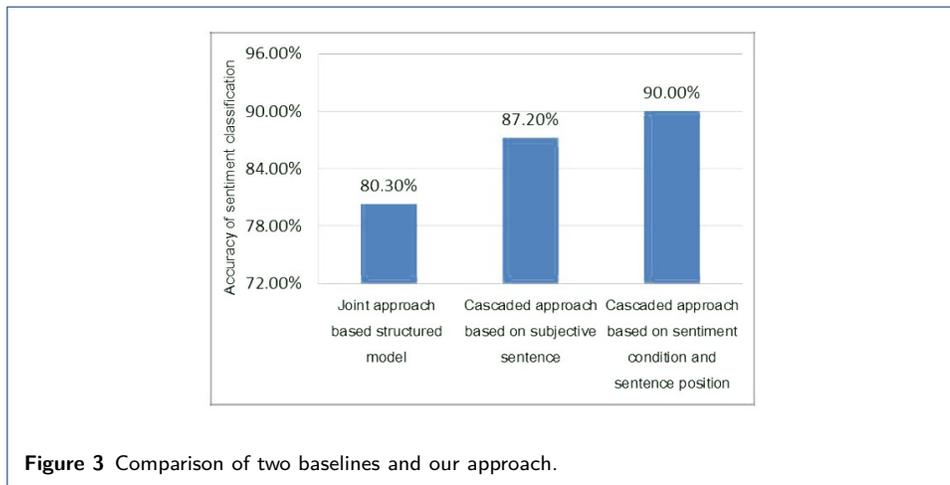


Figure 3 Comparison of two baselines and our approach.

Table 1 An example of sentence sentiment classification

Document	Sentence Segmentation	Sentence Position	SPS	Adjectives	Adverbs	Verbs
It's been a while, but I still remember that it's super near the train station, rather convenient but inevitably noisy. There had been a lot of Korean-Japanese tourists in this hotel, and the receptionists were inhospitable. We had two people registered for a standard room, but they only gave one room card with the aggressive look in their eyes. I was so upset that lost my mood.	It's been a while, but I still remember that it's super near the train station, rather convenient but inevitably noisy.	The first sentence	Positive	near	super	
				convenient	rather	
				noisy	inevitably	
	There had been a lot of Korean-Japanese tourists in this hotel, and the receptionists were inhospitable. We had two people registered for a standard room, but they only gave one room card with the aggressive look in their eyes.	The middle sentences	Negative	inhospitable	only	register
				standard		
				aggressive		
				cheap		
	I was so upset that lost my mood.	The last sentence	Negative	upset		lost

Table 2 Identifier demonstrating sentence position, SPS and SPD

Sentence Position	SPS	Positive Document	Negative Document
First Sentence	Positive	FPP	FPN
	Negative	FNP	FNN
Middle Sentences	Positive	MPP	MPN
	Negative	MNP	MNN
Last Sentence	Positive	EPP	EPN
	Negative	ENP	ENN

Table 3 Labeling results of document-level reviews in training corpus

Item	Result
Amount of reviews	2461
Amount of positive reviews	1195
Amount of negative reviews	1266
Amount of sentences in all documents	11062
Amount of positive sentences in all documents	5349
Amount of negative sentences in all documents	5713

Table 4 Labeling results of sentence-to-document level reviews in training corpus

Sentence Position	SPS	Positive Document	Negative Document	Amount
First Sentence	Positive	1078	80	1158
	Negative	117	1186	1303
Middle Sentences	Positive	2862	314	3176
	Negative	774	2190	2964
Last Sentence	Positive	943	72	1015
	Negative	252	1194	1446

Table 5 CSD Calculated by Different Algorithms

Sentence Position	SPS	Equal Weight	Correlation Degree	Sentiment Condition Probability
First Sentence	Positive	WF=33.3%	WF=92.0%	WF(pos)=93.1%
	Negative			WF(neg)=91.0%
Middle Sentences	Positive	WM=33.3%	WM=82.3%	WM(pos)=90.1%
	Negative			WM(neg)=73.9%
Last Sentence	Positive	WE=33.3%	WE=86.8%	WE(pos)=92.9%
	Negative			WE(neg)=82.6%

Table 6 Sentiment Classification of Sentence and Document

No.	SPD*	SVM Classifier			Equal Weight		Correlation Degree		Assumption of Sentiment Condition	
		SPS _F	SPS _M	SPS _E	Value	SPD	Value	SPD	Value	SPD
Pos1	1	1			1	1	0.921	1	0.926	1
Pos2	1	1	1		2	1	1.731	1	1.8272	1
Pos3	1	1		1	2	1	1.78	1	1.861	1
Pos4	1	-1		1	0	?	-0.062	-1	0.0183	1
Pos5	1	1	2	1	4	1	3.4	1	3.6634	1
Neg1	-1	1			1	1	0.9	1	0.926	1
Neg2	-1	-1	-1	-1	-3	-1	-2.4	-1	-2.4737	-1
Neg3	-1	1	-1	-1	-1	-1	-0.748	-1	-0.631	-1
Neg4	-1	-1	0	1	0	?	-0.062	-1	0.0183	1
Neg5	-1	-1		-1	-2	-1	-1.7	-1	-1.7337	-1

Table 7 Accuracy of document sentiment classification with different CSD algorithms

Algorithms	Positive	Negative	Accuracy
Equal Weight	81.9%	81.4%	81.6%
Correlation Degree	90.1%	87.3%	88.6%
Sentiment Condition Probability	93.5%	86.0%	90.0%

Figures

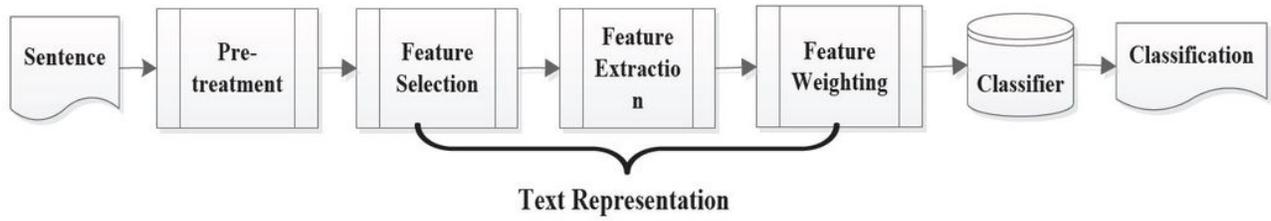


Figure 1

The basic process of sentence sentiment classification.

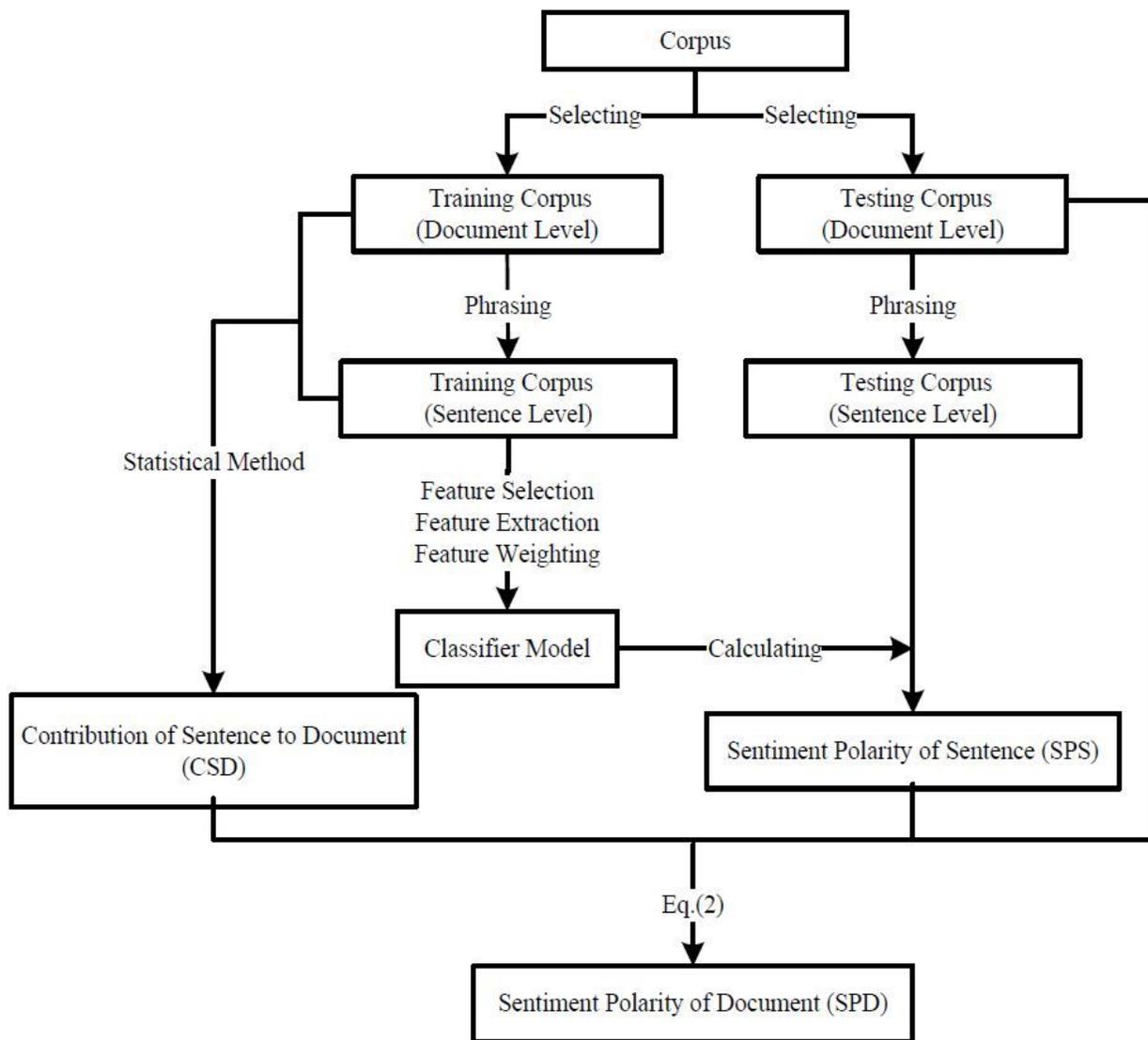


Figure 2

The process of sentence-to-document sentiment classification.

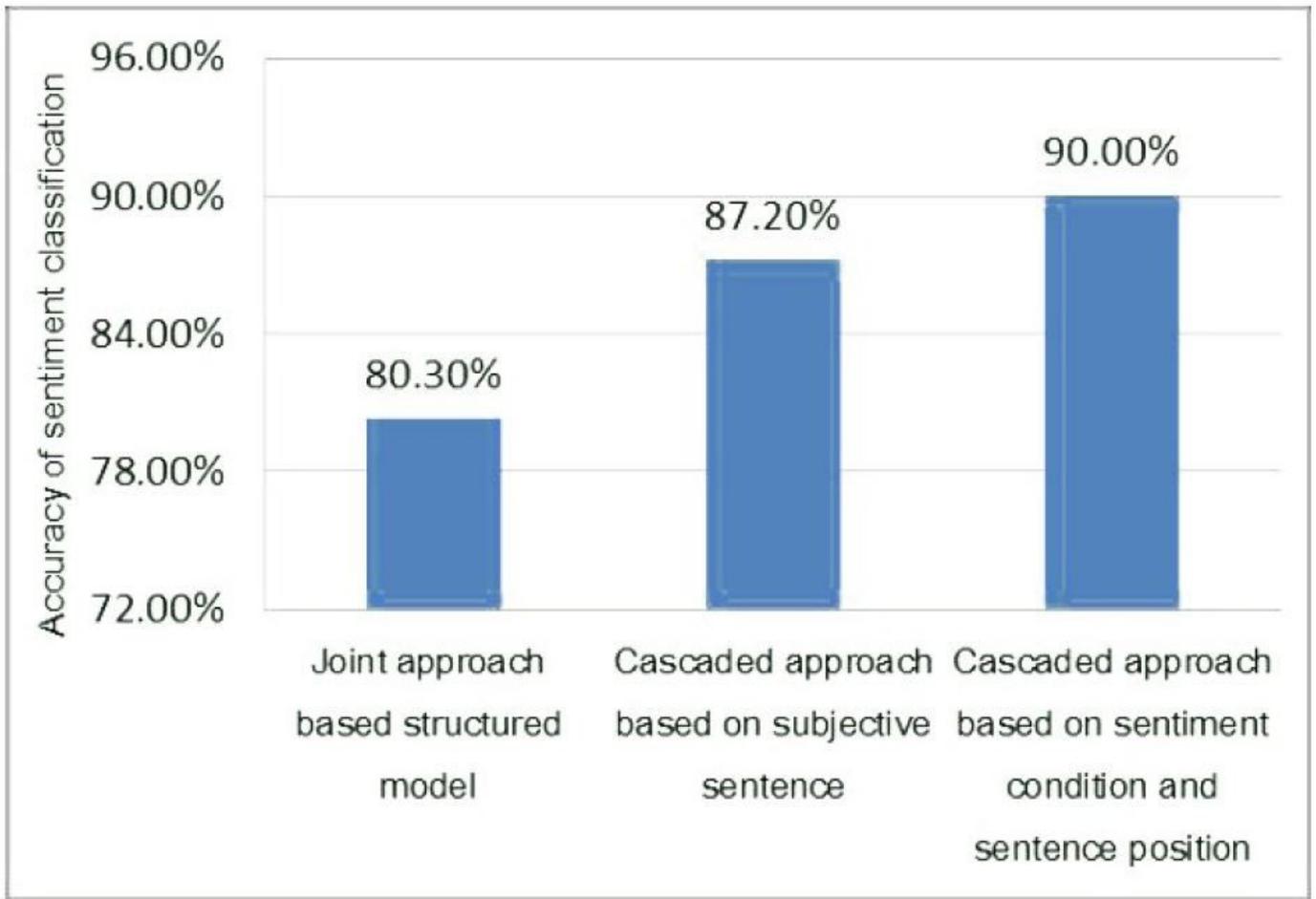


Figure 3

Comparison of two baselines and our approach.