Sentiment Analysis based on Specific Dictionary and Sentence Analysis

Xinyue Wang [a,*], Cheng Ding[b], Wenxi Zheng [c] and Min Wu [d]
Center of Modern Educational Technology, University of Science and Technology of China, China
[a]sa514001@mail.ustc.edu.com,[b] sa514004@mail.ustc.edu.cn,[c] wxzheng@ustc.edu.cn,
[d]minwu@ustc.edu.cn
Corresponding author: Xinyue Wang

Keywords: Sentiment Analysis; Knowledge Presentation; Nature Language Processing; Domain Dictionary.

Abstract. Objective or negative. In this work, we apply a simple method to adapt a general purpose polarity lexicon to a specific domain. We propose and evaluate new features to be used in a word polarity based approach to sentiment classification. We consider different aspects of sentences, such as length, purity and position within the opinionated text. This analysis is then used to find sentences that may convey better information about the overall review polarity. We use a subset of hotel reviews from the TripAdvisor database to evaluate the effect of sentence-level features on sentiment classification. Then, we measure the performance of our sentiment analysis engine using the domain-adapted lexicon on a large subset of the TripAdvisor database.

1. Introduction

Emotional analysis aims to assess the subjectivity of the text and the intensity of the idea, while the subjectivity of the text and the strength of the views together indicate the semantic orientation. Such as the use of the knowledge of the polarity of the view, you can analyze the text words, sentences or product reviews to determine its subjectivity and objectivity. Polarity can be divided into positive, objective or negative, and the strength of the polarity of the view can also be expressed by numerically.

This paper mainly solves the following two problems. First, this paper proposes a domain-adaptive approach to adapt domain-independent polarity dictionaries to their specific domain. Second, this paper presents new features of Sentence-based Sentiment Analysis for comment. Although the polar performance can be simple and effective evaluation of the overall evaluation of the polarity in lexical level, but between the lexical level and comment level, there are still a considerable gap. Using the analysis method based on the sentence is in order to make up for the gap.

The following is a brief introduction to the related work. The third part is the main body, which describes the sentence-based emotion analysis algorithm. The fourth part reports experimental results. At the end of the paper made a summary for this study and outline the future research work.

2. Related Work

In previous studies, the document was usually treated as a word bag and the overall polarity of the document was evaluated using the average polarity of the words within the document [5]. Due to the method is too simple, the researchers' later work is focused on the analysis of phrases and sentences. Some of these studies focus on the subjective analysis of phrases / sentences and use the results to determine the subjectivity of the document. In an earlier study, Wiebe found subjective adjectives from the corpus [3]. Subsequently, Hatzivassiloglou and Wiebe [6] also discussed the impact of the adjective point and the subjectivity of the scalability of the sentence. Some recent works have also examined the relationship between word semantic disambiguation and subjectivity, and extracted enough information to more accurately classify emotion [8]. Wiebe et al. also introduced a census method using different features and cues to identify subjectivity [9].

The polarity dictionary refers to the emotional polarity of words and phrases. SentiWordNet [4] and SenticNet [9] are the most commonly used polarity dictionaries in emotional analysis. In [9], the author discussed three main methods of constructing polar dictionary: artificial method, dictionary-based method and corpus-based method. The main drawback of the artificial method (such as [3]) is that the cost (time and effort cost) of hand-picked vocabulary building dictionaries is too high, and there is the possibility of missing important vocabulary. These terms can be captured by automated methods. Method based on dictionary (e.g., [4, 7]) is to building a dictionary by extending the set of polar corpora and using the way the word source (e.g., WordNet) operates. Note that these methods produce domain-independent dictionaries.

The analysis of sentence is not new. Some researchers have solved the emotional problem by removing irrelevant sentences that interfere with the polarity assessment and by identifying the subjective sentences in the commentary [8]. Another approach is to use the sentence structure for analysis, rather than simply as a word bag. Such as in [9], the author analyzes the conjunction to obtain the lexical polarity associated with its domain.

The first and last lines of a comment are usually highly indicative of the polarity of the comment. This observation led to the work of this paper. Based on this simple observation, this paper constructs more complex features in sentence-level affective analysis.

3. Sentiment Analysis Algorithm based on Sentence

For the emotional analysis of specific documents or comments, this paper presents 24 new features. These features can be divided into five groups, are shown in Table 1: (1) Basic Features; (2) Vocabulary Polarity TF-IDF Weight Features; (3) The Characteristics of Atomic Vocabulary Statistics; (4) Punctuation; (5) The Characteristics of The Sentence Level.

In this paper, the method need to use the available with single or multiple polarity dictionary lexical semantic direction information. Specifically, this paper uses SentiWordNet [6] as the base dictionary, and its domain adaptation version as a special field dictionary.

Table 1. Feature Set.

Group	Feature	Feature Name
Base Features	F1	Average Review Polarity
	F2	Comment Purity
	F3	Comment Subjectivity
TF-IDF	F4	TF-IDF value of the words
	F5	Average Comment Polarity with TF-IDF Weights
Atomic Words Statistics	F6	Frequency of Occurrence of Atomic Words
	F7	Average Polarity of Atomic Words
	F8	Standard Deviation of Polarity of Atomic Words
Punctuation Features	F9	Number of Exclamation Mark
	F10	Number of Question Mark
	F11	Number of Smiling Symbol
	F12	Number of Crying Symbol
Sentence Level Features	F13	Average Polarity of First Sentence
	F14	Average Polarity of Last Sentence
	F15	Purity of First Sentence
	F16	Purity of Last Sentence
	F17	TF-IDF Value of First Sentence
	F18	Polarity with the TF-IDF Weight in First Sentence
	F19	TF-IDF Value of Last Sentence
	F20	Polarity with the TF-IDF Weight in Last Sentence
	F21	Number of Sentences in Comment
	F22	Average Polarity of Subjective Sentences
	F23	Average Polarity of Pure Sentences
	F24	Average Polarity of Objective Sentences

3.1 Base Features

This paper uses the polarities, purity, and subjectivity that are commonly used in emotional analysis as the basic features. This paper contains only words that contain "NN", "JJ", "RB" and "VB" POS tags, because these words may be the comments express emotional words.

3.2 TF-IDF Features

This paper calculates the TF-IDF scores in the SentiWordNet vocabulary to capture the domain specificity. If the TF-IDF score is positive, it indicates that the term is related to the positive classification; if negative, then the other hand. This paper calculates the score on the training set, and it is known that the positive comments in the training set are equal to the number of negative comments.

3.3 Atomic Word Statistics

As with other researchers, this paper also selected a small subset of polar dictionaries for the established field, including 20 positive vocabulary and 20 negative vocabularies, so that they can more accurately indicate the polarity of the hotel reviews. Note, however, that while positive sentences may contain words "good" (such as "not good"), negative sentences are unlikely to contain vocabulary "excellent" (e.g. "food is not excellent"). Negative sentences containing positive vocabularies are more common, and negative sentences that contain positive vocabulary are not.

In order to determine the atomic words of the given field, this paper first calculates the TF-IDF scores for all the special vocabularies in the corpus. Then, use the TF-IDF scores to classify and select the list of the top 20 positive and top 20 negative words. These words make up the atom vocabulary, or called SeedW.

3.4 Punctuation Features

This paper has four features related to punctuation. Between the exclamation mark and question mark, the exclamation mark usually makes the emotion more intense (e.g., "food good!!"). But it can also be used to express incredible reactions (e.g., "room without windows!"). The question mark can be used to detect objective or neutral sentences that may be classified as emotional (e.g., "what is the room?").

Emoticons are also important symbols with emotional polarity. They may also have some significant emotional significances. For example, smiley symbols can be used to indicate happy (e.g., "room view is good^_^"), can also be used to ridicule or agree with a joke.

3.5 Sentence Level Features

Usually, the first or last sentence of the comment summarizes the overall feelings of the comment. This point in the hotel's long comments is no doubt. For example, no matter what the details, the comment with the title contains “excellent hotel” undoubtedly clearly shows the enthusiasm of the over commentary emotion.

Sentence level features are extracted from sentences with special orientations or from specific types of sentences (such as subjective sentences).In particular, this paper considers the subjectivity and purity of the sentence and uses the features extracted from it to detect the overall commentary emotion.

To identify subjective sentences, this article examines whether a sentence contains at least one subjective word or a smiling face symbol. In this paper, the method proposed in [8] is used to determine the subjectivity of vocabulary.

Likewise, if the sentence purity is greater than the fixed threshold, then the sentence Si is pure. The sentence purity can only be calculated by the vocabulary in the sentence. In this paper, different thresholds are experimented, and finally determine the most suitable threshold is 0.8.

In order to reach the goal, this paper has tried three different methods.

Method 1: Only keep sentences that may be useful in a comment (e.g., subjective sentences). Each sentence level feature sets a separate threshold. Sentences are defined as pure sentences when their absolute purity no less than 0.8.

Method 2: When calculating the average polarity of the comment, the polarity of the special sentence is given a higher threshold. In fact, this is because other sentences are given a lower weight, not a weight of zero.

Method 3: The information extracted from the sentence level analysis is used as an additional property (e.g., the average polarity of subjective sentences is taken as a new feature), as shown in Table 2. The results of this method are the best and are applied in the final system.

Table 2. Sentence Level Features on Comment R

F13	Average Polarity of First Sentence
F14	Average Polarity of Last Sentence
F15	Purity of First Sentence
F16	Purity of Last Sentence
F17	TF-IDF Weight Polarity of First Sentence
F18	TF-IDF Score of First Sentence
F19	TF-IDF Weight Polarity of Last Sentence
F20	TF-IDF Score of Last Sentence
F21	Number of Sentences in Comment
F22	Average Polarity of Subjective Sentences
F23	Average Polarity of Pure Sentences
F24	Average Polarity of Objective Sentences

4. Experimental Results

Table 3. The Feature Classification Effect on the TripAdvisor Dataset

Feature Subset	**Accuracy Rate (SMO) (%)**	**Accuracy Rate (Logistic)(%)**
Base Features (F1-F3)	59.62	59.66
+ TF-IDF (F4-F5)	59.97	59.48
+Atomic Words Features(F6-F8)	59.97	59.48
+Punctuation Features(F9-F12)	60.47	60.18
+Average Polarity and Purity of First and Last Sentences(F13-F16)	60.60	60.62
+TF-IDF Value of First and Last Sentences(F17-F20)	60.74	60.67
+Number of Sentences(F21)	60.70	60.78
+Average Polarity of Subjective Sentences(F22)	63.76	64.27
+Average Polarity of Pure Sentences(F23)	63.21	62.89
+Average Polarity of Objective Sentences(F24)	63.76	64.27
Base Features + Atomic Words Features(F6-F8)	59.62	59.66
Base Features + Punctuation Features(F9-F12)	60.11	60.03
Base Features + Average Polarity and Purity of First and Last Sentences(F13-F16)	59.97	59.94
Base Features + TF-IDF Value of First and Last Sentences(F17-F20)	60.28	59.72
Base Features + Number of Sentences(F21)	60.05	59.93
Base Features + Average Polarity of Subjective Sentences(F22)	61.27	60.27
Base Features + Average Polarity of Pure Sentences(F23)	60.19	60.02
Base Features + Average Polarity of Objective Sentences(F24)	62.47	62.64

Table 3 lists the accuracy of the emotional classification. In the upper part of the table, this article provides the results that occur when new features are added incrementally to the order listed in Table 1. In this way, the first added feature has a greater chance of increasing the accuracy of the baseline.

Thus, in the lower half of the table, this paper provides the results of the precision when the properties are added individually to the base feature.

When analyzing these results, this paper notes that the increase in accuracy is mostly through punctuation (59.97%-60.47% using SMO), increasing subjective sentences (60.70%-63.76% using SMO) and adding unrealistic characteristics (63.21%-63.76% Using SMO) to achieve.

It is also found that the statistical properties of the atomic vocabulary do not significantly increase the computational cost of the algorithm compared to the TF-IDF value and the calculation of the purity.

5. Conclusion

By analyzing the sentences in depth, this paper has developed some new features for emotional analysis at the sentence level. This article uses the new features and evaluates them in the public dataset of TripAdvisor's review. This paper found that the features of the sentence level do have some influence to the accuracy of emotional analysis. And concluded that the sentences in the emotional analysis is very important. In more diverse data sets (e.g., blogs), the role of sentence level analysis will be greater.

However, I only consider the emotional divided into two categories, the positive and negative. In the future work, I will divide emotions into more detailed categories. It is also possible to further explore the analysis based on the sentence level to identify the critical sentences in the comment, or highlight the important sentences needed to summarize the comments.

References

[1]. The TripAdvisor website. http://www.tripadvisor.com [TripAdvisor LLC]. Accessed in 2016
[2]. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found.Trends Inf. Retrieval 2(1–2), 1–135 (2008).
[3]. Turney, P.D.: Thumbs up or thumbs down: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th AnnualMeeting on Association for Computational Linguistics, pp. 417–424. Association for Computational Linguistics (2002)
[4]. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up: sentiment classification using machine learning techniques. In: Proceedings of theACL-02 conference on Empirical methods in natural language processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002).
[5]. Esuli, A., Sebastiani, F.: SentiWordNet: a publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference and Evaluation (LREC06), pp. 417–422 (2006).
[6]. Taboada, M., Brooke, J., Tofiloski, M., Voll, K.D., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. 37(2), 267–307 (2011).
[7]. Zhao, J., Liu, K., Wang, G.: Adding redundant features for crfs-based sentence sentiment classification.In: Proceedings of the 2008 Conference on Empirical Methods, pp. 117–126 (2008).
[8]. Poria, S., Gelbukh, A.F., Cambria, E.,: Enriching SenticNet polarity scores through semi-supervised fuzzy clustering. In: Vreeken, J., Ling, C., Zaki, M.J., Siebes, A., Yu, J.X., Goethals, B., Webb, G.I., Wu, X. (eds.) ICDM Workshops, pp. 709–716. IEEE Computer Society (2012).
[9]. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: separating facts identifying the polarity of opinion sentences. In: Proceedings of the 2003 conference on Empirical methods in Natural Language Processing, pp. 129–136. Association for Computational Linguistics (2003).