International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.1-12
http://dx.doi.org/10.14257/ijdta.2015.8.6.01

Opinion Objects Identification and Sentiment Analysis

Ouyang Chunping, Liu Yongbin+, Zhang Shuqing and Yang Xiaohua

School of Computer Science and Technology, University of South China, Hunan Hengyang, 421001, China
+qingbinliu@163.com

Abstract

Sentiment analysis of reviews has been the focus of recent research, which also has been attempted in different domains such as product reviews, movie reviews, and customer feedback reviews. Most sentiment analysis of reviews focused on extracting overall evaluation for a single product which makes difficult for a customer to know all the features of product and make a decision. Thus, mining this data, identifying the user opinions about different features and classify them is an important task. This paper is devoted to identify opinion object from short comments, and analyze sentiment of product based on features-level. CRFs model based on word embedding feature is adopted by identifying opinion object, which obtains a satisfied results. In addition, calculate rules based on syntax parsing are proposed to accomplish features-level sentiment analysis which extracts user's opinion on many aspects. Experimental results using short comments of movies show the effectiveness of our approach.

Keywords: *opinion objects; identification; short comments; sentiment analysis*

1. Introduction

Along with the development of Web 2.0, social media (*e.g.*, reviews, forum discussions, microblogs, comments, and postings in social network sites) is becoming a very important platform for sharing public opinions. We can obtain much of information from opinions analysis. For example, in marketing it helps judging the success of an ad campaign or new product launch, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features [1]. Opinion analysis, also called sentiment analysis, is a type of natural language processing for tracking the mood of the public about a particular product or topic. This has been a very active area of research in multiple domains, such as electronic commerce, public opinion monitoring, social recommendation and information prediction [2].

In text, sentiments can be captured at various levels of granularity: word/phrase level, sentence level and document level. Document-level sentiment classification is to determine the overall sentiment orientation of the document depends on classes which can be positive, negative or neutral. Sentence-level sentiment classification considers each sentence as independent unit and assumes that sentence should contain only one opinion [3]. Word/phrase-level sentiment classification often is used to make a featured-based opinion summary of reviews or comments.

The primary focus of early research in the field was to classify movie reviews as containing overall positive or negative sentiment [4]. This research indicated that adopting Unigram features and SVMs classifier tended to do the best performance. This method is a classical supervised algorithm for sentiment analysis at the level of the document. Turney *et al*. presented a simple unsupervised learning algorithm using PMI-IR to calculate semantic orientation and classified the review based on the average semantic orientation of the phrases [5]. However, because it is difficult to calculate the similarity between emotion words and determine the seed words

ISSN: 2005-4270 IJDTA

accurately, this approach is rarely used in the subsequent study. Most of the follow-up research focused on feature selection and emotion word identification based on supervised algorithm [6].

With the rapid development of social media in China, Chinese scientists already begin to show solicitude for sentiment analysis for Chinese microblog and short comments. They have some of the same characteristics, such as oral, non-normalized words and words number limit, which brought some difficulties for sentiment analysis. For example, people are to combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context [7].

To deal with the issues, native researchers tried to improve the representation of text features. Xie [8] combined 4 features together and used SVM method to analyze sentiments in Sina microblogs. Results showed that the highest accuracy is 66.467% and 67.283%, using subject-related and non-subject-related features separately. Xu and Lin [9] took words and structures in sentences into account, selected 9 semantic features which affect the sentiment of the whole sentence, and combined manual and automatic ways together to construct the sentiment ontology. Their method was an initial attempt in semantic based sentiment analysis. Li and Cao [10] studied from the linguistics perspective and leveraged the "sentiment tendency definition" calculation method based on weight-first to get the feature word which best represents the sentiment tendency among all the words in the phrase. Then they analyzed the sentiment tendency and its strength of the phrase according to the way the characteristic words were combined. The method is meaningful in fine-grained text sentiment analysis.

Sentence-level and document-level sentiment classification is adopted by sentiment analysis of microblogs usually, which distinguishes objective sentences that express factual information from subjective sentences and assumes that each document expresses opinions on a single entity [11]. But for reviews and short comments, we should discover what exactly people liked and did not like. In other word, we should extract users' attitude for a certain feature of product. The following excerpts from reviews of the film "Sex and the City" provide an illustration: "Beginning is outstanding and the ending is that we all yield the insipid happiness. This is a great script. But, I seemed to be feeling fatigued by when the film finishes with the equivalent promise." (Translate from Chinese review) The first sentence and the second sentence are clearly positive, with positive subjective words such as "outstanding" and "great" supporting this conclusion. The last sentence is clearly negative. Phrase such as "feeling fatigued" tip us off that the author did not satisfied with the ending of film. Sentiment does not just occur at the whole document level, nor is it limited to a single object [12]. Therefore, this paper is devoted to features-level sentiment analysis of short comments. Firstly, the opinion objects of sentiment sentence are identified. Then syntax parsing based on sentiment dictionary is used to analyze author's attitude about the objects. The aim of sentiment analysis is to help people to know sentiment acceptance for the different features of a product, and get a comprehensive knowledge of a product, such as book, movie, song, and so on.

The remainder of this paper is organized as follows. In Section 2, we introduce an overall assumption; Section 3 explains how to identify opinion objects from short comments using CRFs model, and in this section the feature selection for CRFs model is discussed. In Section 4, we adopt Stanford syntax parser and set three rules for features-level sentiment analysis. At last, conclusions and further work are made in Section 5.

2. Approach Overview

Recently, how to extract useful information from the vast amount of user generated texts (*e.g.* views, comments, attitudes) has received significant attention. So features-level sentiment analysis is applied widely in the process of classifying the emotion of user generated text into positive, negative, or neutral opinion. It is based on the idea that an opinion without its object being identified is of limited use. Realizing the importance of opinion objects also helps us understand the sentiment analysis problem better [11]. Mining information from user opinions contains several tasks. The first task is to identify and extract opinion objects that have been commented on by an opinion holder. Opinion objects are entities that can be rated by users, such as product components, functionalities, *etc.* The second task is finding the corresponding sentiment to each opinion objects. Sentiments are usually adjectives describing the quality of an object. Finally, in the last task the sentiment is analyzed to estimate the user's quality rating for each aspect where quality rating is numerical rating indicating the quality of that object [13].

In this paper, we study the problem of extracting opinion objects from movies comments, and analyzing the authors' sentiment on the objects. CRFs model is adopted to identify opinion objects, and feature selection for CRFs is the key task. Another task is parsing the dependence between opinion objects and emotion words to obtain author's sentiment of each aspect of a movie. In Figure 1 overall architecture of our proposed method is shown.

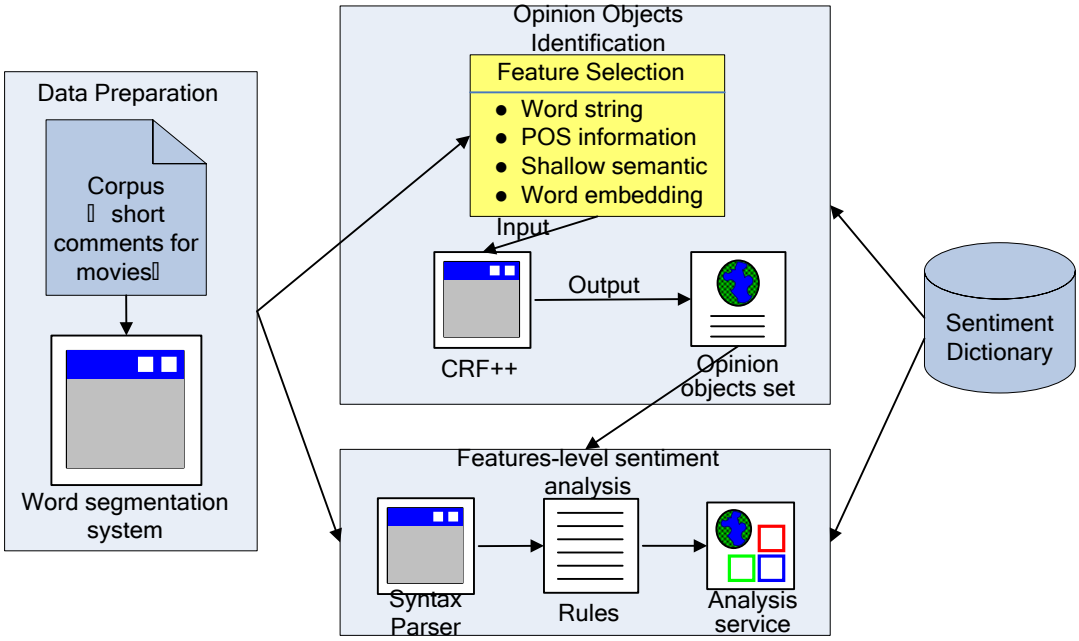


Figure 1. Overall Architecture of Our Proposed Method

In our approach, three tasks will be executed. The first task is data preprocessing including corpus annotated, sentiment dictionary expansion, and Chinese word segmentation. Then we aim to extract opinion objects that are defined as the common features that users (authors) are interested in. In this task, four features, word string, POS, shallow semantic and word embedding, are selected for training CRFs model. After extracting opinion objects, the semantic relationship between opinion objects and emotion words should be identified. An opinion object and emotion words may associate in feature-opinion pair (dependence pair) if the feature usually occurs accompanies with emotion words in short comments. The final task

of sentiment analysis of product features is completed by detecting dependence pairs using three rules self-defined.

3. Opinion Objects Identification

3.1. Conditional Random Fields Model

Conditional Random Fields is one noTable variant of a Markov random field. Conditional Random Fields are a type of discriminative undirected probabilistic graphical model. The nodes in the graph represent random variables; the edges between the nodes represent dependencies between random variables.

X is given the observed values, and Y is a set of random variables. Conditional Random Fields are used to calculate the conditional probability of random variables Y on observations X. CRFs define as follows:

Definition [14]: Let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of *G*. Then (X, Y) is a conditional random field in case, when conditioned on *X*, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v | X, Y_w, w \neq v) = p(Y_v | X, Y_w, w \sim v)$, where $w \sim v$ means that *w* and *v* are neighbors in *G*.

From the view of the fundamental theorem of Markov random field, Conditional Random Fields is composed of an undirected graph and a set of potential functions cliques. The conditional probability distribution represented by a Conditional Random Field is given by

$$P(Y|X) = \frac{1}{Z_x} \prod_{c \in C} \Phi_c(x_c, y_c) \tag{1}$$

Where Z_x, known as the partition function(Normalization factor), is given by

$$Z_X = \sum_y \bullet \prod_{c \in C} \Phi_c(x_c, y_c) \tag{2}$$

$\Phi_c(x_c, y_c)$ is the potential function for the state of the variables that appear in *c* clique, *c* represents *a* clique in all cliques *C*, x_c and y_c represent the variables participated in clique *c*. Conditional Random Fields are often conveniently represented as log-linear combination. Each clique potential is replaced by an exponentiate weighted sum of features of the state:

$$\Phi_c(x_c, y_c) = \exp(\sum_k \lambda_{kc} f_{kc}(x_c, y_c)) \tag{3}$$

Conditional random fields commonly used in the labeling or analysis of sequence data, such as Natural Language Processing, *e.g.* [15]. In principle, the graph model layout of conditional random fields can be arbitrarily given, but popular layout is a linear chain. It is because that there are efficient algorithms on the training, inference, or decoding of the linear chain CRFs.

Opinion objects identification is a sub problem of named entity recognition, which is also a sequence tagging problem. So a linearchain CRFs is employed for our problem. The important issue is the feature set selection for CRFs, since it is the key factor for the accuracy of CRFs. Moreover, there is the ineviTable problem of the feature sparsity problem, which is serious in our case due to the limited size of our tagged data set. We present our approach for these tasks in the following section.

3.2. Features Selection for CRFs

An important characteristic of Conditional Random Fields model is that various features can be flexibly defined and additional independence assumptions or internal constraints do not need to be considered. On the other hand, it also shows that feature selection has a great influence on the final results. According to the research result, the

combination of multiple features in comparison with a single feature generally shows a reasonable enhancement of any classification system [16]. Consequently, through manually reviewing the short comments corpus and their language specific characteristics, the feature sets for opinion objects are defined, as shown in the Table 1.

Table 1. Feature Sets for Opinion Objects Identification

Feature	Feature symbol	Description
Word string	Token	The original tokens, as well as stemmed tokens, as features using the word segment tool.
POS information	POS	Chinese word is tagged "noun, verb, adj, adv, nw, idiom, prep", and statistics for each part of speech in the sentence frequency.
Shallow semantic	SRL	Shallow semantic parsing is used to label the semantic role in a sentence.
Word embedding	W2V	Each short comment is represented as a sequence of sentiment words and underlying states.

Word string features. Token feature refers to the word string feature after word segmentation, word strings is a very important feature of opinion object extraction. Observing the corpus of short comments, the opinion object often can be a compound word multiple noun, pronoun or phrases. But because there is no obvious boundary between Chinese words, and word segmentation tool has error with compound word, the opinion object can't be identified accurately, which often with the context of the other noun or phrase to form a compound word. Then the errors in combining words are likely to lead to the compound word can't be identified as the opinion object. Therefore, when successive none occurs, we choose merging continuous noun to a noun phrase. For example, after word segmentation, a phrase "Ancient costume/nz movie /n superstar /n", according to the rule, it is combined into "Ancient costume movie superstar /n". For word strings feature selection in this paper, we use NLPIR, a Chinese word segmentation system, developed by Dr. Zhang Huaping [17].

Part-of-speech features. POS features is part of speech tagging of each word after word segmentation, such as noun(/n), verb(/v), adjectives(/a) and so on. In general, polysemy often exists in Chinese text, and in different context, the same word often have different parts of speech. In short comments corpus, most of opinion objects are a noun or noun phrases, part of speech features is the important feature of opinion objects, which can provide more information to help identifying the opinion objects, so it is chosen as input of CRFs model.

Shallow semantic parsing features. Shallow semantic parsing is a simplified method of semantic analysis, which does not consider the temporal and the anaphora relationship. Semantic Role Labeling is implementation of shallow semantic parsing. It does not analyze the semantic of whole sentence in detail, but it annotates the semantic roles for a given predicate (verb, noun and so on) in the sentence. This feature can better reveal the relationship between the opinion object and a predicate. For example, "I like this nostalgic film.", according to semantic role analysis, "nostalgic film" is patient role for emotion word "like".

Word embedding features. The above three features are commonly used in sentiment analysis and entity recognition, but these features are difficult to reflect the correlation between each word and emotion words. Mikolov [18] found that the relationship between the two term vectors can be directly reflected from the difference of two vectors. So in order to express the relationship between each word and emotion words, we use the most popular method "word embedding" to express every word in corpus. However, this approach will inevitably lead to the feature sparsity problem, in order to solve this

problem, clustering based the distance between word embeddings is proposed. These features can be suitable for expressing the correlation between words and emotion words. For example, phrase "A wonderful story" and phrase "A wonderful beginning". Hypothesis word "wonderful" is embedded to c(wonderful), word "story" is embedded to c(story), word "beginning" is embedded to c(beginning), then the distance between them is calculated, we can found c(story)- c(wonderful) ≈ c(beginning)- c(wonderful), therefore it can be inferred that word "story" and word "beginning" belongs to the similarity class. They are both likely to be opinion object.

Three steps is executed to obtain word embedding features, a description of each of these steps is follows.

Step 1: Trying to utilize meaningful and useful text features, we turned to word2vec due to its speed and ease of use. Word2vec computes vector representations of words using a few different techniques, two of which are continuous bag-of-words (CBOW) and an architecture called a Skipgram [19]. Due to our dataset is not very big, so Skipgram model is adopted, and dimension for word2vec is set to 300.

Step 2: Calculated the distance between each word vectors and each emotion word vector using the cosine distance formula. Avoiding the influence of the accuracy of CRFs, the sparse results will be clustered in next step.

Step 3: In an attempt to quickly find clusters of similar features, we used Kmeans++ [20] to cluster the word vectors. This data is used to try to find semantic-related terms with the goal of using the results for a feature of CRFs model.

3.3. Data and Model Training

Data. Experimental data is from Douban (a Chinese social website for comments sharing about books, songs and movies), including hundreds of thousands of short comments for 1500 films. Download this 264M corpus which for free from datatang (www.datatang.com). For word embedding features selection, the whole corpus is input of word2vec. In additional, the emotion words used in calculation of two word embedding is from sentiment thesaurus constructed by Dalian University of Technology, which is enriched by some network vocabulary and emoticons. While for the other three features, we use simplified corpus, which includes 300 films are classified into 3 types according to the place of origin: American film, Chinese film and Korea film. In each category, 5000 short comments are selected randomly to be annotated by manual. Those datasets had been labeled with polarity for three weeks by 3 annotators, who are both native Chinese speakers. The 5000 annotated short comments are used as training dataset, and the rest of the 10000 short comments are used as test dataset.

Model Training. We train our CRFs model using CRF++, a highly efficient general purpose CRFs toolkit written in C++ [21]. CRF++ allows the definition of both unigram and bigram features, where unigram features are related to the prediction of a single observation in a sequence (first order Markov) and bigram features are related to the prediction of pairs of observations (second order Markov). Unigram features generate a total of L×N distinct features, where L is the number of output classes and N is the number of unique features. Bigram features generate L×L×N distinct features. In our task, unigram template is adopted by CRFs model, and the size of feature template window is 3.

When running the train command "%crf_learn template_file train_file model_file>> info.txt", the parameter "-f NUM" is set to "-c 5", which meant that only the features appears not less than 5 times will be calculated when CRF++ is training. Additionally, in the case of the CRFs trained on independent observations, we remove all but node features, so as to avoid an artificial decrease in performance. At last, command "%crf_test –m model_file test_file >>result.txt" is running, and then the accuracy of model is achieved by assessing the file "result.txt".

3.4. Results and Observations

In our experiments, NLPIR-ICTCLAS 2013 released by CAS is employed to segment words and to tag POS. Semantic role labeling is implemented by using the algorithm proposed by Wang [22].CRF++0.58 is employed to train model. Three evaluation metrics, that is, Precision (P), Recall (R), and F-measure (F) are used for performance evaluation for binary classification of sentiment. The formulas of evaluation standards are listed as follows.

$$\text{Precision} = \frac{\#system\_correct}{\#system\_proposed} \tag{4}$$

$$\text{Recall} = \frac{\#system\_correct}{\#annotated} \tag{5}$$

$$F\text{-}measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

In above formulas, #system_correct denotes the number of the submitted results which match with the manually annotated results, #system_proposed denotes all the number of the submitted results, #annotated denotes the number of the manually annotated results.

For creating the results, we carried out four groups of experiments, respectively according to American film, Chinese film, Korea film and unclassified film. In Table 2 the comparison results for identifying the opinion objects of films are provided. From the results, recall has lower value than precision, which shows that there are still a fair number of unrecognized objects. But for experiments on specific category, Chinese film has the most efficient result of precision, and the recall value of American film is the lowest. However, Korea film achieves good results on F-measure.

We observe the results and the annotated corpus, many transliterations in person name, location name and organization name are missing, which lead to the lowest recall of American film. In this regard, Chinese film do better, the lower recall of Chinese film because that there are many colloquial language in comments. Korea film has the most F value for two reasons. The first, person name, location name in Korea film are similar to Chinese words, which led to identified easily. Secondly, China audiences make relatively less complaints with oral language for Korea film. Taken as a whole, it is that a satisfactory result can be acquired on CRFs model for opinion objects identification.

Table 2. Experimental Results of Objects Identification

	Precision	Recall	F-measure
American Films	0.7422	0.5213	0.6124
Chinese Films	0.8103	0.6835	0.7415
Korea Films	0.7811	0.7124	0.7451
All	0.7864	0.6511	0.7123

4. Sentiment Analysis

Document-level sentiment analysis and sentence-level sentiment analysis are hot research in nature language processing, especially for the microblog. But the vast amount of user generated data is from Internet, such as electronic commerce website, discussion forums, comments-centered social network, and so on. Sentiment analysis for those data, we should extract users' opinion about the different aspect of comments. Aspects (also called product features) are entities that can be rated by users, such as product

components, functionalities, *etc.* In hence, this paper focuses on features-level sentiment analysis, which is based on the idea that an opinion consists of a sentiment (positive or negative) and a feature of products.

4.1. Syntax Analysis

Syntactic structure consisting essentially of the relationship between words and words, this relation is called dependence. A dependence relation connecting two words, which is the core words and modifiers. Dependencies can be divided into different types, and represents the syntactic relation between two words.

According to the given grammar rules, syntax analysis can complete the following work: deriving the grammar structure of sentences automatically, analyzing the relationship between grammatical unit included in sentences, and converting the sentences into a structured syntax tree. After syntactic analysis completion, all the relations appearing in the sentence will be expressed to the dependence between words pairs with the unified two tuple form. In addition, the types of relationship labeled are also made in results. These relationships are not limited to two of adjacent words, and those relationships between distant words can be better solved.

The Stanford Parser is used by to analyze the dependence between opinion object and emotion words in this paper. Stanford parser is a program that works out the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb [23]. Consider the following example: “最后一幕机器人转身离开的场面有些悲凉(The last scene of the robot turned and left let us feel some sad)”. After parsing the sentence, the typed dependencies is obtained and shown in Table 3.

Table 3. Typed Dependencies Example

Abbreviation	Description	Dependence
amod	Adjectives modify	amod(幕-3, 最后-1)
nummod	Quantifier	nummod(幕-3, 一-2)
nn	Compound nouns	nn(转身-5, 幕-3), nn(转身-5, 机器人-4)
nsubj	Noun theme	nsubj(离开-6, 转身-5), nsubj(悲凉-10, 场面-8)
relcl	Predicated Noun dependents	relcl(场面-8, 离开-6)
mark	Special clausal	mark(离开-6, 的-7)
advmod	Adverbial modifier	advmod(悲凉-10, 有些-9)

However, sentiment analysis doesn’t process all the typed dependencies. Hypothesis word “场面(scene)” is opinion object, then some related dependencies (*e.g.*, nsubj(悲凉(sad), 场面(scene)), relcl(场面(scene), 离开(left)), advmod(悲凉(sad), 有些(some))) will be chosen to carry out a further analysis.

4.2. Features-Level Sentiment Analysis

The task of features-level sentiment analysis is to determine the polarity of opinion object on features classes: positive and negative. In order to reduce the influence of error dependencies made by syntactic parser on results, we define the distance between two words is 8 (*i.e.* if the distance of two words in a pair of dependence is greater than 8, we believe that the dependence is invalid, not be considered).

We construct a 2-tuple <*EW*, *FW*> to compute the polarity of sentiment. *EW* denotes emotion word, *FW* denotes feature word, and *S*(*w*) denotes sentiment of the feature word. Emotion words used by rules belong to the sentiment thesaurus constructed by Dalian

University of Technology, which is enriched by some network vocabulary. When $S(w)$ is larger than 0, sentiment of the feature word is positive. On the contrary, when $S(w)$ is smaller than 0, sentiment of the feature word is negative.

Followed that, based on the results of the syntactic analysis, we set a few rules to filter sentiment analysis tuples (*i.e* the dependencies associated with opinion objects), which is described as follows.

Rule 1. If emotion word only appears in the *nsubj* dependence, then $S(w)=EW_{emotion}$, $EW_{emotion}$ is given by sentiment thesaurus.

Rule 2. If at the same time emotion word appears in the *nsubj* dependence, the *advmod* dependence, and the *neg* dependence, then $S(w)=-1*EW_{emotion}$.

Rule 3. If emotion word appears in an iterative dependence, for example, *nsubj(ew,fw),advmod(ew1,ew),dep(ew2,ew1)*, at this time *EW* can be expressed as a nested tuple, then $S(w)=\prod_{i=1} EW_{emotion}(i)$, in other word, the sentiment value of feature word is equal to the product of all emotion words value.

In the above rules, the sentiment analysis results depend on the results of syntax parsing. But syntax parsing is not a research subject in this paper, and Stanford syntax parsing is used directly. Hence, the sentiment analysis results are not evaluated according to recall and precise, but from another perspective, many statistical results of sentiment analysis are offered to users. Take a movie "RoboCop" as an example, the statistical result of emotion proportion for hot features is shown in Figure 2.

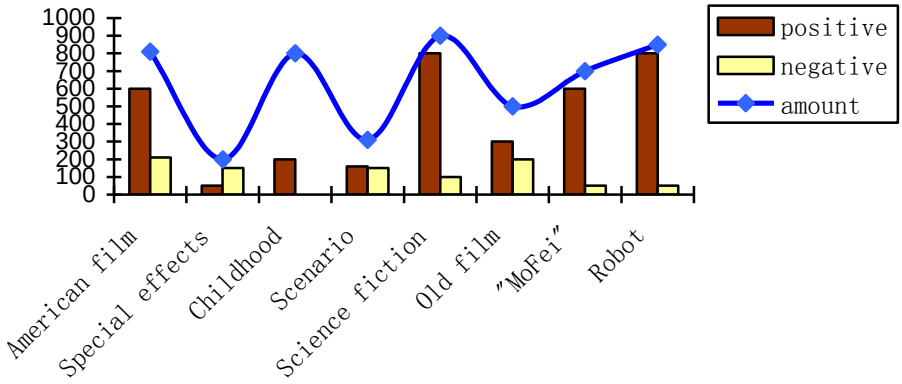


Figure 2. The Emotion Proportion for a Movie Features

In the above figure, fold line denotes the total amount of feature word, columnar section denotes positive and negative respectively. From the shape of the curve, we can see which features are attractive. In addition, which features have more positive opinion is clear and unambiguous. And from the overall trend, the audiences give the movie a high rating.

5. Conclusions and Future Work

Sentiment analysis is the process of extracting knowledge from the peoples' opinions, appraisals and emotions toward entities, events and their attributes. In this paper, we have studied the problem of analyzing author opinions expressed on the Web about movies and their features. To address the problem, we designed three tasks: data preprocessing, opinion objects identification and features-level sentiment analysis. The latter two are the key research contents. We formulated the second task as a sequence tagging problem and solved it using a CRF. For optimizing the accuracy of model, we introduced word embedding features based on deep learning technology, which is one of the main contributions of this paper. Experimental results indicate that opinion object identification

based on CRFs perform well in movie short comments. In addition to aspects of movies and audients' opinions, we also proposed a features-level sentiment analysis, which is a method based on syntax parsing and rules. By using this approach we can view the strength or weakness of the movie more detail, and the results generated by our method are summarized and helpful for user in decision making.

Currently we are working towards all reviews regarding particular movie from various sites and it will classify all the reviews based on sentiments present in reviews. The need for domain-adaptable sentiment lexicons and more sophisticated feature extraction drives us further in our research. Furthermore, different domain has different characteristics, our method is only in the film domain experiment was carried out, so we plan to apply this technique on sentiment analysis of audio review, book review and product review. By processing the data in a different domain, in order to our model is optimized.

Acknowledgement

This research work is supported by National Natural Science Foundation of China (No.61402220), Hunan Provincial Natural Science Foundation of China (No.13JJ4076), the Scientific Research Fund of Hunan Provincial Education Department for excellent talents (No.13B101), Foundation of University of South China (No.2012XQD28), the Construct Program for the Key Discipline in University of South China (No.NHxk02), the Construct Program for Innovative Research Team in University of South China.

References

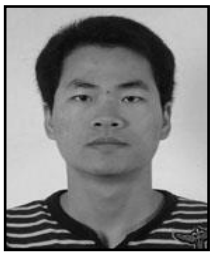
[1] Vinodhini G. and Chandrasekaran R. M., "Sentiment analysis and opinion mining: a survey", International Journal, vol. 2, no. 6, **(2012)**.
[2] Zhang C., Zeng D., Li J., Wang F. Y. and Zuo W., "Sentiment Analysis of Chinese Documents: From Sentence to Document level", Journal of the American Society for Information Science and Technology, vol. 60, no. 12, **(2009)**.
[3] Jagtap V. S. and Pawar K., "Analysis of different approaches to sentence-level sentiment classification", International Journal of Scientific Engineering and Technology, vol. 2, **(2013)**.
[4] Pang B., Lee L. and Vaithyanathan S, "Thumbs up?: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Grenoble, France, **(2002)**.
[5] Turney P. D., "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", Proceedings of the 40th annual meeting on association for computational linguistics, Philadelphia, Penn, **(2002)**.
[6] O. Chunping, Y. Xiaohua, L. Longyan, X. Qiang, Y. Ying and Liu Zhiming, "Multi-strategy Approach for Fine-grained Sentiment Analysis of Chinese Microblog", Acta Scientiarum Naturalium Universitatis Pekinensis, vol. 50, no. 1, **(2014)**.
[7] Y. Wu, S. Jingjing and T. Jiqiang, "A Study on the Classification Approach for Chinese MicroBlog Subjective and Objective Sentences", Journal of Chongqing Institute of Technology, vol. 27, no. 1, **(2013)**.
[8] X. Lixing, Z. Ming and S. Maosong, "Hierarchical Structure Based Hybrid Approach Sentiment Analysis of Chinese Microblog and Its Feature Extraction", Journal of Chinese Information Processing, vol. 26, no. 1, **(2012)**.
[9] X. Linhong and L. Hongfei, "Discourse Affective Computing Based on Semantic Features and Ontology", Journal of Computer Research and Development, vol. 44, no. 3, **(2007)**.
[10] D. Xishuang, Z. Qibo and G. Yi, "A Survey on Sentiment Analysis Models", Scientific Journal of Psychology, vol. 1, no. 2, **(2012)**.
[11] Liu B., "Sentiment analysis and opinion mining", Synthesis Lectures on Human Language Technologies, vol. 5, no.1, **(2012)**.
[12] Fink C. R., Chou D. S. and Kopecky J. J., "Coarse- and Fine-Grained Sentiment Analysis of Social Media Text", Johns Hopkins APL Technical Digest, vol. 30, no. 1, **(2011)**.
[13] Moghaddam S. and Popowich F., "Opinion polarity identification through adjectives", arXiv preprint arXiv:1011.4623, **(2010)**.
[14] J. Lafferty, A. McCallum and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", Proceeding of the 18th International Conference on Machine Learning, Williamstown, MA, USA, **(2001)**.
[15] Liu B., "Sentiment analysis and subjectivity", Handbook of natural language processing, vol. 2, **(2010)**.

[16] Das D. and Bandyopadhyay S., "Emotion Tagging–A Comparative Study on Bengali and English Blogs", ICON, vol. 9, **(2009)**.
[17] Zhou L. and Zhang D., "NLPIR: A theoretical framework for applying natural language processing to information retrieval", Journal of the American Society for Information Science and Technology, vol. 54, no. 2, **(2003)**.
[18] Mikolov T., Chen K., Corrado G., "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781, **(2013)**.
[19] Goldberg Y. and Levy O., "word2vec Explained: deriving Mikolov *et al*.'s negative-sampling word-embedding method", arXiv preprint arXiv:1402.3722, **(2014)**.
[20] Arthur D. and Vassilvitskii S., "k-means++: The advantages of careful seeding", Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. New Orleans, Louisiana, **(2007)**.
[21] Kudo T., "CRF++: Yet another CRF toolkit", Software available at https://code.google.com/p/crfpp/downloads/list, **(2014)**.
[22] W. Rongyang, J. Jiupeng, L. Shoushan and Z. Guodong, "Opinion Objects Identification and Sentiment Analysis", Journal of Chinese Information Processing, vol. 26, no. 2, **(2012)**.
[23] Chen D. and Manning C. D., "A fast and accurate dependency parser using neural networks", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, **(2014)**.

Authors

Ouyang Chunping, received her Ph.D. degree from University of Science Technology Beijing, China, in 2011. Now, she is working in University of South China, where she is an associate professor and the dean of software engineering department. Her research interests are in Semantic Web, Social Computing and Sentiment Analysis of Text.

Liu Yongbin, born in 1978, he received his Ph.D. degree from University of Science Technology Beijing, China, in 2013. He is a lecturer in School of Computer and Technology, University of South China. His current research interests focus on Semantic Web and Data Mining.



Zhang Shuqing, received his B.E. degree from Department of Computer Science of Huanghe Science and Technology College, Zhengzhou, in 2011. Now, he is working his M.E. Degree at University of South China. His research interests include Natural Language Processing and Data Mining.

Yang Xiaohua, received his Ph.D. degree from CAS, China, in 1999. From 2000 to 2001, he was a visiting scholar at Wollongong University. Yang is vice-president of University of South China, and professor of computer science, doctoral tutor. He has authored and coauthored over 50 research publications in peer-reviewed reputed journals. He has served as the program committee member of various international conferences and reviewer for various international journals. His research interests include Natural Language Processing and Information Retrieval.

