Sentiment Analysis of Product Reviews

Cane W. K. Leung*
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon
Hong Kong SAR
Voice: +852 2766-7311
Fax: +852 2170-0115
Email: cane.leung@gmail.com

Stephen C. F. Chan
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon
Hong Kong SAR
Voice: +852 2766-7259
Fax: +852 2170-0108
Email: csschan@comp.polyu.edu.hk

(* Corresponding author)

Sentiment Analysis of Product Reviews

Cane W. K. Leung, The Hong Kong Polytechnic University, Hong Kong SAR

Stephen C. F. Chan, The Hong Kong Polytechnic University, Hong Kong SAR

INTRODUCTION

Sentiment analysis is a kind of text classification that classifies texts based on the *sentimental orientation* (SO) of opinions they contain. Sentiment analysis of product reviews has recently become very popular in text mining and computational linguistics research. The following example provides an overall idea of the challenge. The sentences below are extracted from a movie review on the Internet Movie Database:

"It is quite boring...... the acting is brilliant, especially Massimo Troisi."

In the example, the author stated that "it" (the movie) is quite boring but the acting is brilliant. Understanding such sentiments involves several tasks. Firstly, evaluative terms expressing opinions must be extracted from the review. Secondly, the SO, or the polarity, of the opinions must be determined. For instance, "boring" and "brilliant" respectively carry a negative and a positive opinion. Thirdly, the opinion strength, or the intensity, of an opinion should also be determined. For instance, both "brilliant" and "good" indicate positive opinions, but "brilliant" obviously implies a stronger preference. Finally, the review is classified with respect to sentiment classes, such as *Positive* and *Negative*, based on the SO of the opinions it contains.

BACKGROUND

Sentiment analysis is also known as opinion mining, opinion extraction and affects analysis in the literature. Further, the terms *sentiment analysis* and *sentiment classification* have sometimes been used interchangeably. It is useful, however, to distinguish between two subtly different concepts. In this article, hence, sentiment analysis is defined as a complete process of extracting and understanding the sentiments being expressed in text documents, whereas sentiment classification is the task of assigning class labels to the documents, or segments of the documents, to indicate their SO.

Sentiment analysis can be conducted at various levels. Word level analysis determines the SO of an opinion word or a phrase (Kamps et al., 2004; Kim and Hovy, 2004; Takamura and Inui, 2007). Sentence level and document level analyses determine the dominant or overall SO of a sentence and a document respectively (Hu and Liu, 2004a; Leung et al., forthcoming). The main essence of such analyses is that a sentence or a document may contain a mixture of positive and negative opinions. Some existing work involves analysis at different levels. Specifically, the SO of opinion words or phrases can be aggregated to determine the overall SO of a sentence (Hu and Liu, 2004a) or that of a review (Turney, 2002; Dave et al., 2003; Leung et al., forthcoming).

Most existing sentiment analysis algorithms were designed for binary classification, meaning that they assign opinions or reviews to bipolar classes such as *Positive* or *Negative* (Turney, 2002; Pang et al., 2002; Dave et al., 2003). Some recently proposed algorithms extend binary sentiment classification to classify reviews with respect to multi-point rating scales, a problem known as *rating inference* (Pang and Lee, 2005; Goldberg and Zhu, 2006; Leung et al., forthcoming). Rating inference can be viewed as a multi-category classification problem, in which the class labels are scalar ratings such as 1 to 5 “stars”.

Some sentiment analysis algorithms aim at summarizing the opinions expressed in reviews towards a given product or its features (Hu and Liu, 2004a; Gamon et al., 2005). Note that such *sentiment summarization* also involves the classification of opinions according to their

SO as a subtask, and that it is different from classical document summarization, which is about identifying the key sentences in a document to summarize its major ideas.

Sentiment analysis is closely related to *subjectivity analysis* (Wiebe et al., 2001; Esuli and Sebastiani, 2005). Subjectivity analysis determines whether a given text is subjective or objective in nature. It has been addressed using two methods in sentiment analysis algorithms. The first method considers subjectivity analysis a binary classification problem, for example, using *Subjective* and *Objective* as class labels. Pang and Lee (2005) adopted this method to identify subjective sentences in movie reviews. The second method makes use of part-of-speech (POS) information about words to identify opinions (Turney, 2002; Hu and Liu, 2004a; Leung et al., forthcoming) because previous work on subjectivity analysis suggests that adjectives usually have significant correlation with subjectivity (Bruce and Wiebe, 1999; Wiebe et al., 2001).

MAIN FOCUS

Figure 1 depicts a typical sentiment analysis model. The model takes a collection of reviews as input, and processes them using three core steps, *Data Preparation*, *Review Analysis* and *Sentiment Classification*. The results produced by such a model are the classifications of the reviews, the evaluative sentences, or opinions expressed in the reviews.

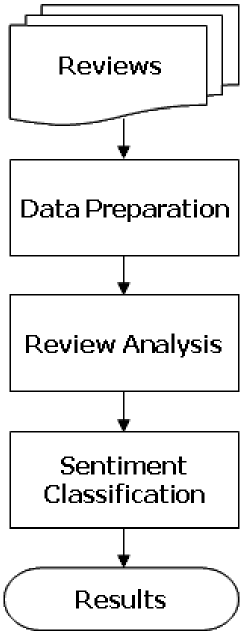

Figure 1: A typical sentiment analysis model.

Data Preparation

The data preparation step performs necessary data preprocessing and cleaning on the dataset for the subsequent analysis. Some commonly used preprocessing steps include removing non-textual contents and markup tags (for HTML pages), and removing information about the reviews that are not required for sentiment analysis, such as review dates and reviewers' names.

Data preparation may also involve the sampling of reviews for building a classifier. Positive reviews often predominate in review datasets as reported in a number of studies (e.g. Turney, 2002; Dave et al., 2003; Gamon et al., 2005). Some researchers therefore use review datasets with balanced class distributions when training classifiers to help demonstrate the performance of their algorithms (Pang et al., 2002; Leung et al., forthcoming).

Review Analysis

The review analysis step analyzes the linguistic features of reviews so that interesting information, including opinions and/or product features, can be identified. This step often applies various computational linguistics tasks to reviews first, and then extracts opinions and product features from the processed reviews. Two commonly adopted tasks for review analysis are POS tagging and negation tagging. POS tagging helps identifying interesting words or phrases having particular POS tags or patterns from reviews (Turney, 2002; Hu and Liu, 2004a; Leung et al., forthcoming), while negation tagging is used to address the contextual effect of negation words, such as “not”, in a sentence (Pang et al., 2002; Dave et al., 2003; Leung et al., forthcoming). For example, “good” and “not good” obviously indicate opposite SO. Given the term “not good”, negation tagging recognizes the existence of the word “not” and adds a special negation tag to the word “good” based on some heuristics.

The review analysis step then proceeds to extract opinions and/or product features from the processed reviews. The opinions or features extracted may be *n*-grams, which are *n* adjacent or nearby words in a sentence (e.g. Turney, 2002). Pang et al. (2002) make use of corpus statistics and human introspection to decide terms that may appear in reviews. Various algorithms adopt a more common method that extracts words or phrases having particular POS tags or patterns as opinions and product features as noted (Turney, 2002; Dave et al., 2003; Takamura and Inui, 2007, Leung et al., forthcoming).

While Hu and Liu (2004b) also make use of POS tags, they adapted the idea of frequent itemsets discovery in association rule mining to product feature extraction. In the context of their work, an itemset is a set of words that occurs together, and a “transaction” contains nouns or noun phrases extracted from a sentence of a review. They used the CBA association rule miner (Liu et al., 1998) to mine frequent itemsets, and considered each resulting itemset to be a product feature. They then processed a review sentence by sentence. If a sentence contains a frequent feature, they extracted its nearby adjective as an opinion. They also proposed methods for pruning redundant features and for identifying infrequent features.

Sentiment Classification

There are two major approaches to classifying reviews, known as the *SO approach* and the *machine learning approach*. The following subsections describe the overall idea and representative techniques of each of the approaches.

SO Approach

The SO approach involves two subtasks. The first subtask is to determine the SO of the opinions extracted from reviews in the Review Analysis step, while the second subtask is to determine the overall SO of a sentence or a review based on the SO of the opinions it contains.

Turney (2002) proposed an unsupervised SO determination method that computes the SO of an opinion phrase as the Pointwise Mutual Information (PMI) between the phrase and two *seed adjectives*, “excellent” and “poor”. Such information is collected using a generic search engine. Specifically, the phrase is likely to represent a positive (resp. negative) sentiment if it co-occurs frequently with the word “excellent” (resp. “poor”). The average of the SO of all opinion phrases in a review is computed, and the review is classified as *Positive* if the average is positive; and *Negative* otherwise.

Hu and Liu (2004a) presented a word-similarity-based algorithm that utilizes the semantical relationship between words to predict SO. Their bootstrapping algorithm depends on a small set of seed adjectives having known SO, such as “great” for *Positive* and “bad” for *Negative*, and automatically expands the set using the synonym and antonym sets in WordNet (Miller et al., 1990), assuming that *semantical similarity implies sentimental similarity*. Specifically, the SO of synonyms is assumed to be the same, whereas that of antonyms is opposite to each other. After predicting the SO of opinions, their algorithm classifies a sentence as *Positive* or *Negative* based on the dominant SO of the opinions in the sentence. Kamps et al.

(2004) and Kim and Hovy (2004) also described word-similarity-based methods for determining SO, but their studies deal with word-level classification.

Leung et al. (2006b) suggest that semantical similarity may not imply sentimental similarity in sentiment analysis, based on statistical observations from a movie review corpus. They therefore proposed a relative-frequency-based method for determining the SO of an opinion. Their method estimates the SO and opinion strength of a word with respect to a sentiment class as its relative frequency of appearance in that class. For example, if the word “best” appeared 8 times in *Positive* reviews and 2 times in *Negative* reviews, its strength with respect to *Positive* SO is then 8/(8+2) = 0.8. Leung et al. (forthcoming) deals with the rating inference problem, which classifies reviews with respect to rating scales. They hypothesized that some product features may be more important for determining the rating of a review, and therefore assign weights to opinions according to the estimated importance of their associated product features. They compute the weighted average SO of opinions in a review, and then rate the review by mapping the weighted average onto an *n*-point rating scale.

Machine Learning Approach

The machine learning approach is similar to topic classification, with the topics being sentiment classes such as *Positive* and *Negative* (Pang et al., 2002). It works by breaking down a review into words or phrases, representing the review as a document vector (bag-of-words model), and then classifying the reviews based on the document vectors.

Pang et al. (2002) investigated whether binary sentiment classification can be addressed using standard topic classification techniques. They applied three classifiers, including Naïve Bayes, Support Vector Machines (SVM) and Maximum Entropy, to a movie review corpus. They also attempted to incorporate various features of the reviews into the standard bag-of-words model, such as the positions of words in the reviews, but the performance of the three classifiers was found inferior to those reported for topic classification. Pang and Lee concluded that sentiment classification is more difficult than topic classification, and that discourse analysis of reviews is necessary for more accurate sentiment analysis.

Pang and Lee (2005) formulated rating inference as a metric-labeling problem. They first applied two *n*-ary classifiers, including one-vs-all (OVA) SVM and SVM regression, to classify reviews with respect to multi-point rating scales. They then use a metric-labeling algorithm to explicitly alter the results of the *n*-ary classifiers to ensure that similar items receive similar labels, determined using a similarity function. While term overlapping is a commonly-used similarity function in topic classification, it does not seem effective in identifying reviews having similar ratings (Pang and Lee, 2005). They therefore proposed the Positive-Sentence Percentage (PSP) similarity function, computed as the number of positive sentences divided by the number of subjective sentences in a review. Experimental results in general show that using metric-labeling with PSP improves the performance of the *n*-ary classifiers. Goldberg and Zhu (2006) later extended Pang and Lee's work using transductive semi-supervised learning. They demonstrated that unlabeled reviews (those without user-specified ratings) can help improve classification accuracy.

Zhu and Goldberg (2007) proposed a kernel regression algorithm utilizing *order preferences* of unlabeled data, and successfully applied the algorithm to sentiment classification. The order preference of a pair of unlabeled data, x_i and x_j, indicates that x_i is preferred to x_j to some degree, even though the exact preferences for x_i and x_j are unknown. In the context of sentiment analysis, for example, given two reviews, one may be able to determine which review is more positive than the other without knowing the exact ratings associated with the reviews.

Zhu and Goldberg applied their algorithm to the rating inference problem, and showed that order preferences improved rating inference performance over standard regression.

FUTURE TRENDS

Most existing algorithms apply generic opinion and product feature extraction methods to reviews, such as methods based on POS tags, without considering the properties of the domain items concerned. While such generic methods allow easy adaptation of a sentiment analysis algorithm to various domains, the performance achieved by the same algorithm was often found to vary significantly when being applied to datasets from different domains (e.g. Turney, 2002; Aue and Gamon, 2005). This, currently being addressed as a domain adaptation issue (Blitzer et al., 2007), reveals a need for more intelligent sentiment analysis models that utilize domain knowledge when extracting opinions and product features from reviews.

An emerging trend regarding sentiment classification is the paradigm shift from binary classification (e.g. *Positive* vs. *Negative*) to multi-point rating inference (e.g. 1-5 "stars"). Recent studies suggest that rating inference shall not be tackled as a classical multi-category classification problem, as the ordering of class labels in rating inference is essential (Okanohara and Tsujii, 2005; Pang and Lee, 2005; Zhu and Goldberg, 2007). This opens up an interesting direction in ordered multi-category classification for future research.

Another interesting research direction is related to the utilization of the sentiments learnt from product reviews. Sentiment analysis has been used to support business and customer decision making by assisting users to explore customer opinions on products that they are interested in (Yamanishi and Li, 2002; Hu and Liu, 2004a). Leung et al. (forthcoming) recently discussed the potential use of sentiment analysis to augment ratings for collaborative filtering (CF), which is also a popular research topic in and application of data mining.

CF provides personalized recommendations to a user based on his/her preferences and the preferences of other users having similar tastes (Leung et al., 2006a). A CF-based system operates on a database of user preferences collected either explicitly by asking users to give scalar ratings on items that they have examined, or implicitly by capturing users' interactions with the system (e.g. purchase histories). Using sentiment analysis to augment ratings for CF on the one hand allows CF to use product reviews as an additional source of user preferences. On the other hand, it enables existing review hubs to utilize the user preferences learnt from reviews for personalization purpose. In view of these advantages, integrating sentiment analysis and CF is expected to be of high interests to data and text mining practitioners.

CONCLUSION

Sentiment analysis deals with the classification of texts based on the sentiments they contain. This article focuses on a typical sentiment analysis model consisting of three core steps, namely data preparation, review analysis and sentiment classification, and describes representative techniques involved in those steps.

Sentiment analysis is an emerging research area in text mining and computational linguistics, and has attracted considerable research attention in the past few years. Future research shall explore sophisticated methods for opinion and product feature extraction, as well as new classification models that can address the ordered labels property in rating inference. Applications that utilize results from both sentiment analysis and CF are also expected to emerge in the near future.

REFERENCES

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. *Proceedings of Recent Advances in Natural Language Processing*.

Blitzer, J., Dredze, M., & Pereira, F. (2007) Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Proceedings of the 45th Annual Meeting of the ACL*. Retrieved June 23, 2007, from http://acl.ldc.upenn.edu/P/P07/P07-1056.pdf

Bruce, R., & Wiebe, J. (1999). Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5(2), 187-205.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of 12th International World Wide Web Conference*.

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. *Proceedings of the ACM Conference on Information and Knowledge Management*, pp. 617-624.

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. *Lecture Notes in Computer Science*, vol. 3646, pp. 121-132.

Goldberg, A. B., & Zhu, X. (2006). Seeing stars when there aren’t many stars: Graph-based semi-supervised learning for sentiment categorization. *Proceedings of TextGraphs Workshop*, pp. 45-52.

Hu, M., & Liu, B. (2004a). Mining and summarizing customer reviews. *Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168-177.

Hu, M., & Liu, B. (2004b). Mining opinion features in customer reviews. *Proceedings of 19th National Conference on Artificial Intelligence*, pp. 755-760.

Kamps, J., Marx, M., Mokken, R. J., & de Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. *Proceedings of 4th International Conference on Language Resources and Evaluation*, vol. VI, 1115-1118.

Kim, S.-M., & Hovy, E. H. (2004). Determining the sentiment of opinions. *Proceedings of 20th International Conference on Computational Linguistics*, pp. 1367-1373.

Leung, C. W. K., Chan, S. C. F., & Chung, F. L. (2006a). A Collaborative Filtering Framework Based on Fuzzy Association Rules and Multiple-Level Similarity. *Knowledge and Information Systems (KAIS)*, 10(3), 357-381.

Leung, C. W. K., Chan, S. C. F., & Chung, F. L. (2006b). Integrating collaborative filtering and sentiment analysis: A rating inference approach. *Proceedings of ECAI 2006 Workshop on Recommender Systems*, pp. 62-66.

Leung, C. W. K., Chan, S. C. F., & Chung, F. L. (forthcoming). Evaluation of a rating inference approach to utilizing textual reviews for collaborative recommendation. *Cooperative Internet Computing*, World Scientific. Retrieved May 23, 2007, from http://www4.comp.polyu.edu.hk/~cswkleung/pub/cic_evalRatingInference.pdf

Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. *Proceedings of Knowledge Discovery and Data Mining*, pp. 80-86.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (Special Issue)*, 3(4), 235-244.

Okanohara, D., & Tsujii, J. (2005). Assigning polarity scores to reviews using machine learning techniques. In R. Dale, K.-F. Wong, J. Su and O. Y. Kwong (Eds.), *Natural Language Processing - IJCNLP 2005*, Springer-Verlag, pp. 314-325.

Pang, B., & Lee, L. (2005) Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of 43rd Annual Meeting of the ACL*, pp. 115-124.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pp. 79-86.

Takamura, H., Inui, T., & Okumura, M. (2007) Extracting semantic orientations of phrases from dictionary. *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the ACL*, pp. 292-299.

Turney, P (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of 40th Annual Meeting of the ACL*, pp. 417-424.

Wiebe, J., Bruce, R., Bell, M., Martin, M., & Wilson, T. (2001). A corpus study of evaluative and speculative language. *Proceedings of 2nd ACL SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark.

Yamanishi, K., & Li, H. (2002). Mining open answers in questionnaire data. *IEEE Intelligent Systems*, 17(5), 58-63.

Zhu, X., & Goldberg, A. B. (2007). Kernel Regression with Order Preferences. *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, Retrieved May 23, 2007, from http://www.cs.wisc.edu/~jerryzhu/pub/orderssl_aaai07.pdf

KEY TERMS AND THEIR DEFINITIONS

Collaborative Filtering: A recommendation technique that provides personalized recommendations to a user based on his/her expressed interests and the interests of other users having similar preferences.

Opinion Strength: Indicates how strong the opinion is given its sentimental orientation. It is also known as the intensity of an opinion.

Rating Inference: Refers to sentiment classification with respect to multi-point rating scales. It can be viewed as an *n*-ary classification problem in which the class labels are scalar ratings.

Sentiment Analysis: The process of analyzing the sentiments expressed in texts, and then classifying and/or summarizing the sentiments based on their polarity.

Sentiment Classification: A core step in sentiment analysis that classifies a text with respect to sentiment classes, such as *Positive* and *Negative*. An *n*-ary sentiment classification problem is also known as rating inference.

Sentiment Summarization: Summarizes the opinions expressed in a document or in a set of documents towards a product.

Sentimental Orientation: Indicates the polarity, such as *Positive* or *Negative*, of a text known to be subjective.

Subjectivity Analysis: Determines whether a text is subjective or objective (factual) in nature.