RESEARCH ARTICLE OPEN ACCESS

CO-Extracting Opinion Targets and Opinion Words from Online Reviews Based On the Word Alignment Model

Aditya Rane, Sankalp Rane, Saily Sawant, Shubham Sali
Prof. Sunil Jadhav
Department of Computer Science and Engineering
Yadavrao Tasgaonkar
Institute of Engineering and Technology
Chandhai, Raigad, Mumbai
Maharashtra - India

ABSTRACT

Mining opinion targets and opinion words from online reviews are important tasks for fine grained opinion mining, the key component of which involves detecting opinion relations among words. To this end, this paper proposes a novel approach based on the partially-supervised alignment model, which regards identifying opinion relations as an alignment process. Then, a graph-based co-ranking algorithm is exploited to estimate the confidence of each candidate. Finally, candidates with higher confidence are extracted as opinion targets or opinion words. Our model captures opinion relations more precisely, especially for long-span relations. Our experimental results on three corpora with different sizes and languages show that our approach effectively outperforms state-of-the-art methods.

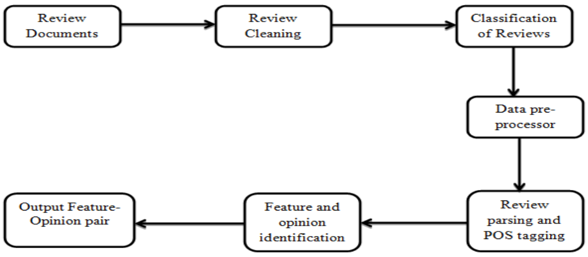
Keywords:- Data Mining, Text Mining

I. INTRODUCTION

Recently, a number of online shopping customers have dramatically increased due to the rapid growth of e-commerce, and the increase of online merchants. To enhance the customer satisfaction, merchants and product manufacturers allow customers to review or express their opinions on the products or services. The customers can now post a review of products at merchant sites, e.g., amazon.com, cnet.com, and epinions.com. These online customer reviews, thereafter, become a cognitive source of information which is very useful for both potential customers and product manufacturers. Customers have utilized this piece of this information to support their decision on whether to purchase the product. For product manufacturer perspective, understanding the preferences of customers is highly valuable for product development, marketing and consumer relationship management.

Since customer feedbacks influence other customer's decision, the review documents have become an important source of information for business organizations to take it development plans.

How does Opinion Mining System Works?


Review Documents
Review Cleaning
Classification of Reviews
Data pre-processor
Review parsing and POS tagging
Feature and opinion identification
Output Feature-Opinion pair

Among the 2 main types of textual information - facts and opinions, a major portion of current information processes methods such as web search and text mining work with the former. Opinion Mining refers to the broad area of natural language processing, computational linguistics and text mining involving the computational study of opinions, sentiments and emotions expressed in text. A thought, view, or attitude based on emotion instead of reason is often referred to as a sentiment. Hence, an alternate term for Opinion Mining, namely Sentiment Analysis. This field ends critical use in areas where organizations or individuals wish to know the general sentiment associated to a particular entity - be it a product, person, public policy, movie or even an institution. Opinion mining has many application domains including science and technology, entertainment, education, politics, marketing, accounting, law, research and development. In earlier days, with limited access to user generated opinions, research in this field was minimal. But with the tremendous growth of the World Wide Web, huge volumes of opinionated texts in the form of blogs, reviews, discussion groups and forums are available for analysis making the World Wide Web the fastest, most comprehensive and easily accessible medium for sentiment analysis. However, finding opinion sources and monitoring them over the Web can be a formidable task because a large number of diverse sources exist on the Web and each source also contains a huge volume of information. From a human's perspective, it is both difficult and tiresome to find relevant sources, extract pertinent sentences, read them, summarize them and organize them into usable form. An automated and faster opinion mining and summarizing system is thus needed.

Overview

Our work is partly based on and closely related to opinion mining and sentence sentiment classification. Extensive research has been done on sentiment analysis of review text and subjectivity analysis (determining whether a sentence is subjective or objective). Another related area is feature/topic-based sentiment analysis, in which opinions on particular attributes of a product are determined. Most of this work concentrates on finding the sentiment associated with a sentence (and in some cases, the entire review). There has also been some research on automatically extracting product features from review text. Though there has been some work in review summarization, and assigning summary scores to products based on customer reviews, there has been relatively little work on ranking products using customer reviews.

II. EXISTING SYSTEM

Existing Systems on feature-based opinion mining have applied various methods for feature extraction and refinement, including NLP and statistical methods. However, these analyses revealed two main problems. First, most systems select the feature from a sentence by considering only information about the term itself, for example, term frequency, not bothering to consider the relationship between the term and the related opinion phrases in the sentence. As a result, there is a high probability that the wrong terms will be chosen as features. Second, words like 'photo,' 'picture,' and 'image' that have the same or similar meanings are treated as different features since most methods only employ surface or grammatical analysis for feature differentiation. This results in the extraction of too many features from the review data, often causing incorrect opinion analysis and providing an inappropriate summary of the review analysis.

Level of Opinion Mining

The opinion mining tasks at hand can be broadly classified based on the level at which it is done with the various levels being namely,

- The document level,
- The sentence level and
- The feature level.

At the document level, sentiment classification of documents into positive, negative, and neutral polarities is done with the assumption made that each document focuses on a single object O(although this is not necessarily the case in many realistic situations such as discussion forum posts) and contains opinion from a single opinion holder. At the sentence level, identification of subjective or opinionated sentences amongst the corpus is done by classifying data into objective (Lack of opinion) and subjective or opinionated text. Subsequently, sentiment classification of the aforementioned sentences is done moving each sentence into positive, negative and neutral classes. At this level as well, I make the assumption that a sentence contains only one opinion which as in our previous levels is not true in many cases. An optional task is to consider clauses.

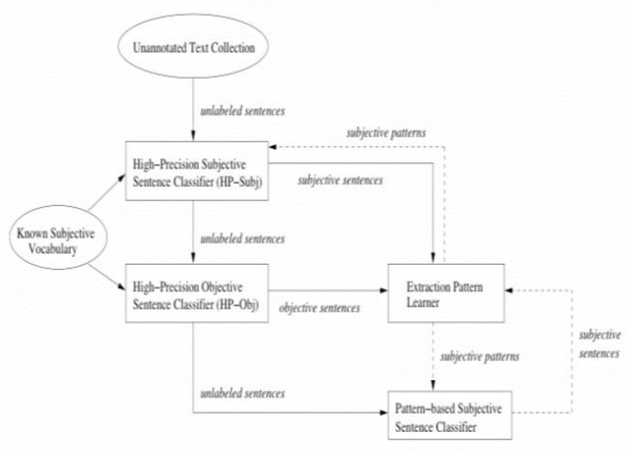
At the feature level, the various tasks that are looked at are:

- Task1: Identifying and extracting object features that have been commented on in each review/text.
- Task 2: Determining whether the opinions on the features are positive, negative or neutral.
- Task 3: Grouping feature synonyms and producing a feature-based opinion summary of multiple reviews/text.

When both F (the set of features) and W (synonym of each feature) are unknown, all three tasks need to be performed. If F is known but W is unknown, all three tasks are needed, but Task 3 is easier. It narrows down to the problem of matching discovered features with the set of given features F. When both W and F are known, only task 2 is needed.

III. SENTENCE-LEVEL SENTIMENT ANALYSIS

The sentiment classification at the document-level is the most important field of web opinion mining. However, for most applications, the document-level is too coarse. Therefore it is possible to perform finer analysis at the sentence-level. The research studies in this field mostly focus on a classification of the sentences whether they hold an objective or a subjective speech, the aim is to recognize subjective sentences in news articles and not to extract them. The sentiment classification as it has been described in the document-level part still exists at the sentence-level; the same approaches as the Turney's algorithm are used, based on likelihood ratios. Because this approach has already been described in this paper, this part focuses on the objective/subjective sentences classification and presents two methods to tackle this issue. The first method is based on a bootstrapping approach using learned patterns. It means that this method is self-improving and is based on phrases patterns which are learned automatically.


Unannotated Text Collection
unlabeled sentences
subjective patterns
High–Precision Subjective Sentence Classifier (HP–Subj)
subjective sentences
Known Subjective Vocabulary
unlabeled sentences
High–Precision Objective Sentence Classifier (HP–Obj)
objective sentences
Extraction Pattern Learner
subjective sentences
subjective patterns
unlabeled sentences
Pattern–based Subjective Sentence Classifier

The input of this method is known subjective vocabulary and a collection of annotated texts.

• The high-precision classifiers find whether the sentences are objective or subjective based on the input vocabulary. High-precision means their behaviors are stable and reproducible. They are not able to classify all the sentences but they make almost no errors.

• Then the phrases patterns which are supposed to represent a subjective sentence are extracted and used on the sentences the HP classifiers have let unlabeled.

• The system is self-improving as the new subjective sentences or patterns are used in a loop on the unlabeled data.

This algorithm was able to recognize 40% of the subjective sentences in a test set of 2197 sentences (59% are subjective) with a 90% precision. In order to compare, the HP subjective classifier alone recognizes 33% of the subjective sentences with a 91%precision.Along this original method, more classical data mining algorithm are used such as the naïve bayes classifier.

The general concept is to split each sentence in features -- such as presence of words, presence of n-grams, and heuristics from other studies in the field -- and to use the statistics of the training data set about those features to classify new sentences. Their results show that the more features, the better. They achieved at best a 80-90%recall and precision classification for subjective/opinions sentences and a 50% recall and precision classification for objective/facts sentences. The sentence-level sentiment classification methods are improving, this results from research studies in 2003 show that they were already quite efficient then and that the task is possible.

Feature and Opinion Learner

This module is responsible to analyze dependency relations generated by document parser and generate all possible information components from them. The dependency relations between a pair of words w1 and w2 is represented as relation type (w1; w2), in which w1 is called head or governor and w2 is called dependent or modifier. The relationship relation type between w1 and w2 can be of two types- i) direct and ii) indirect. In a direct relationship, one word depends on the other or both of them depend on a third word directly, whereas in an indirect relationship one word depends on the other through other words or both of them depend on a third word indirectly. An information component is defined as a triplet < f; m; o >, where f represents a feature generally expressed as a noun phrase, o refers to opinion which is generally expressed as adjective, and m is an adverb that acts as a modifier to represent the degree of expressiveness of the opinion. As pointed out in, opinion words and features are generally associated with each other and consequently, there exist inherent as well as semantic relations between them. Therefore, the feature and opinion learner module is implemented as a rule-based system, which analyzes the dependency relations to identify information components from review documents. For example, consider the following opinion sentences related to Nokia N95:

(i) The screen is very attractive and bright.

(ii) The sound sometimes comes out very clear.

(iii) Nokia N95 has a pretty screen.

(iv) Yes, the push email is the \Best" in the business.

In example (i), the screen is a noun phrase which represents a feature of Nokia N95, and the adjective word attractive can be extracted using nominal subject *nsubj* relation (a dependency relationship type used by Stanford parser) as an opinion. Further, using *advmod* relation the *adverb* very can be identified as a modifier to represent the degree of expressiveness of the opinion word attractive. In example (ii), the noun sound is a nominal subject of the verb comes, and the adjective word clear is adjectival complement of it. Therefore, clear can be extracted as opinion word for the feature sound. In example (iii), the adjective pretty is parsed as directly depending on the noun screen through *amod* relationship. If pretty is identified as an opinion word, then the word screen can be extracted as a feature; likewise, if screen is identified as a feature, the adjective word pretty can be extracted as an opinion. Similarly in example (iv), the noun email is a nominal subject of the verb is, and the word Best is direct object of it. Therefore, Best can be identified as opinion word for the feature word email.

Based on these and other observations, we have defined different rules to tackle different types of sentence structures to identify information components embedded within them.

Rule-1: In a dependency relation R, if there exist relationships nn(w1;w2) and nsubj(w3;w1) such that POS(w1) = POS(w2) = NN_, POS(w3) = JJ* and w1, w2 are not stop-words, or if there exists a relationship nsubj(w3;w4) such that POS(w3) = JJ*, POS(w4) =NN* and w3, w4 are not stop-words, then either (w1;w2) or w4 is considered as a feature and w3 as an opinion.

Rule-2: In a dependency relation R, if there exist relationships nn(w1;w2) and nsubj(w3;w1) such that POS(w1) = POS(w2) = NN_, POS(w3) = JJ* and w1, w2 are not stop-words, or if there exists a relationship nsubj(w3;w4) such that POS(w3) = JJ*, POS(w4) =NN_ and w3, w4 are not stop-words, then either (w1;w2) or w4 is considered as the feature and w3 as an opinion. Thereafter, the relationship advmod (w3; w5) relating w3 with some adverbial word w5 is searched. In case of presence of advmod relationship, the information component is identified as < (w1; w2) or w4; w5; w3 > otherwise < (w1; w2) or w4; -; w3 >.

Rule-3: In a dependency relation R, if there exist relationships nn(w1;w2) and nsubj(w3;w1) such that POS(w1) = POS(w2) = NN_, POS(w3) = V B_ and w1, w2 are not a stop-words, or if, there exist a relationship nsubj(w3;w4) such that POS(w3) = V B*, POS(w4) = NN* and w4 is not a stop-word, then we search for acomp(w3;w5) relation. If acomp relationship exists such that POS (w5) = JJ_ and w5 is not a stop-word then either (w1; w2) or w4 is assumed as the feature and w5 as an opinion. Thereafter, the modifier is searched and information component is generated in the same way as in Rule-2.

The need to identify and interpret possible difference in the linguistic style of texts–such as formal or informal–is increasing, as more people use the Internet as their main research resource. There are different factors that affect the style, including the words and expressions used and syntactical features. Vocabulary choice is likely the biggest style marker. In general, longer words and Latin origin verbs are formal, while phrasal verbs and idioms are informal (Park, 2007). There are also many formal/informal style equivalents that can be used in writing.

The formal style is used in most writing and business situations, and when speaking to people with whom we do not have close relationships. Some characteristics of this style are long words and using the passive voice. Informal style is mainly for casual conversation, like at home between family members, and is used in writing only when there is a personal or closed relationship, such as that of friends and family. Some characteristics of this style are word contractions such as “won’t”, abbreviations like “phone”, and short words. We discuss the main characteristics of both styles.

Characteristics of Informal Style Text

The informal style has the following characteristics:

1. It uses a personal style: the first and second person (“I” and “you”) and the active voice (e.g., “I have noticed that...”).
2. It uses short simple words and sentences (e.g., “latest”).
3. It uses contractions (e.g., “won’t”).
4. It uses many abbreviations (e.g., “TV”).
5. It uses many phrasal verbs in the text (e.g., “find out”).

6. Words that express rapport and familiarity are often used in speech, such as “brother”, “buddy” and “man”.
7. It uses a subjective style, expressing opinions and feelings (e.g.“pretty”, “I feel”).
8. It uses vague expressions, personal vocabulary and colloquialisms (slang words are accepted in spoken text, but not in written text (e.g., “wanna” = “want to”))

IV. CHARACTERISTICS OF FORMAL STYLE TEXT

The formal style has the following characteristics:

1. It uses an impersonal style: the third person (“it”, “he” and “she”) and often the passive voice (e.g., “It has been noticed that...”).
2. It uses complex words and sentences to express complex points (e.g., “state-of-the-art”).
3. It does not use contractions.
4. It does not use many abbreviations, though there are some abbreviations used in formal texts, such as titles with proper names (e.g., “Mr.”) or short names of methods in scientific papers (e.g., “SVM”).
5. It uses appropriate and clear expressions, precise education, and business and technical vocabularies (Latin origin).
6. It uses polite words and formulae, such as “Please”, “Thank you”, “Madam” and “Sir”.
7. It uses an objective style, citing facts and references to support an argument.
8. It does not use vague expressions and slang words.

V. CONCLUSIONS

This paper proposes a novel method for co-extracting opinion targets and opinion words by using a word alignment model. Our main contribution is focused on detecting opinion relations between opinion targets and opinion words. Compared to previous methods based on nearest neighbor rules and syntactic patterns, in using a word alignment model, our method captures opinion relations more precisely and therefore is more effective for opinion target and opinion word extraction. Next, we construct an Opinion Relation Graph to model all candidates and the detected opinion relations among them, along with a graph co-ranking algorithm to estimate the confidence of each candidate. The items with higher ranks are extracted out. The experimental results for three datasets with different languages and different sizes prove the effectiveness of the proposed method. In future work, we plan to consider additional types of relations between words, such as topical relations, in Opinion Relation Graph. We believe that this may be beneficial for co-extracting opinion targets and opinion words.

ACKNOWLEDGMENT

It is matter of great satisfaction and pleasure to present seminar on **“CO-EXTRACTING OPINION TARGETS AND OPINION WORDS FROM ONLINE REVIEWS BASED ON THE WORD ALIGNMENT MODEL”.** We wish to express our sincere thanks to **H.O.D. Mrs. Prof. Nilima Nikam** who extended their valuable support during the course of seminar. We wish to express our sincere thanks and gratitude to our honourable guide **Mr. Prof. Sunil Jadhav** for his constant guidance and motivation. We also thank her for her valuable support and encouragement throughout the preparation of seminar without which the seminar would have not been completed. We also thank our colleagues who have helped in successful completion of the seminar. Last but not least we would like to thank all our friends, who helped us not directly or indirectly. Helpful hand rendered by all of them

will remain for long time in our memory. Finally we admit the cooperation, coordination & hard work are our keywords for success.

REFERENCES

[1] K. Liu, L. Xu, and J. Zhao, "Opinion target extraction using word-based translation model," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju, Korea, July 2012, pp. 1346–1356.

[2] M. Hu and B. Liu, "Mining opinion features in customer reviews," in Proceedings of the 19th the National Conference on Artificial Intelligence (AAAI), San Jose, California, USA, 2004, pp. 755–760.

[3] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, 2005, pp. 339– 346.

[4] G. Qiu, L. Bing, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," Computational Linguistics, vol. 37, no. 1, pp. 9–27, 2011.

[5] B. Wang and H. Wang, "Bootstrapping both product features and opinion words from chinese customer reviews with cross-inducing," in Proceedings of the third International Joint Conference on Natural Language Processing, Hyderabad, India, 2008, pp. 289–295.