A Probabilistic Approach to Study Features in Opinion Mining

Ms. Mrunal S. Pendharkar
Dept. of Information Technology
RMD Sinhgad School of engineering
Pune, India
mrun.pendhkr@gmail.com

Ms. Sweta Kale
Dept. of Information Technology
RMD Sinhgad School of engineering
Pune, India
swetakale30@sinhgad.edu

Abstract— **Opinion mining (also called as sentiment analysis) aims to understand people's thinking towards entities such as products like mobiles, etc. In opinion mining, an opinion word, or feature in short, extracts a thing or a feature of an entity on which users state their views. The proposed model is a new approach to the recognition of such features from unstructured textual reviews. The opinion features are assigned weights or frequency which is used to state the nature of reviews. It is, thus, important to extract the specific opinionated features from text reviews and associate them to opinions. The Proposed work generates syntactic relevant rules which are used to create a list of candidate rules from the given domain review database, for example, mobile or hotel databases. Secondly, for each standard feature candidate, its domain application score with respect to the domain-specific and domain independent database is calculated, which is termed as Intrinsic-Domain Relevance (IDR) score, and the Extrinsic Domain Relevance (EDR) score, respectively. Finally, candidate reviews with low IDR scores and high EDR scores are deleted. It is called as the Intrinsic and Extrinsic Domain Relevance (IEDR) criterion.**

Keywords-information search and retrieval; opinion mining; natural language processing

I. INTRODUCTION

Opinion mining also called as sentiment analysis helps to analyse people's views, thoughts, and attitudes towards products, services, and their attributes [1]. Sentiments or views expressed in textual reviews are typically analysed at various levels. For example, sentence-level opinion mining identifies the overall subjectivity or sentiment expressed on an entity (e.g., mobile or hotel) in a review document, but it does not correlate opinions with specific aspects (e.g., display, battery) of the entity. In opinion mining, an opinion feature, or feature indicates a thing or an attribute of an entity on which user's state in their opinions.

The proposed model is a novel approach to the identification of such features from unstructured textual reviews. Supervised learning model may be tuned to work well in a given domain, but the model must be analyzed if it is applied to different domains [5]. Unsupervised natural language processing (NLP) approaches [4] identify opinion features by defining domain-independent syntactic templates or rules that capture the dependence roles and local context of the feature terms. One key finding is that the distributional structure of an opinion feature in a given domain-dependent review corpus, for example, mobile reviews, is different from that in a domain-independent corpus.

The proposed method is summarized as follows: Firstly, various syntactic dependence rules are used to create a list of candidate features from the given domain review corpus, for example, mobile or hotel reviews. Then, for each predictable feature candidate, its domain relevance score with respect to the domain-specific and domain independent databases is calculated, which is termed as the intrinsic-domain relevance (IDR) score, and the extrinsic domain relevance (EDR) score, respectively. Finally, candidate features with low IDR scores and high EDR scores are eliminated. Thus, it is called as interval thresholding the intrinsic and extrinsic domain relevance (IEDR) criterion. Evaluations observed on two real-world review domains give the effectiveness of the proposed IEDR approach in identifying opinion features.

For example, consider a review of a mobile by a customer, "The battery and features of the iPhone are admirable with advanced technology but the price is too costly". Here the overall review is positive but the aspects 'price' expresses a negative opinion. Hence, a fine grained approach is required to extract the appropriate nature of review.

II. NEED OF THE WORK

One key need is that the distributional structure of an opinion feature in a given domain-dependent review corpus, is different from that in a domain-independent corpus. This leads to propose a method to identify opinion features by exploiting their distribution disparities across different corpora. The purpose of opinion mining is to categorize the overall bias or sentiment expressed in an individual review document. On any internet website, reviews by users or manufactures are randomly shown; hence, for any passionate user it is difficult to understand exact review of the product. Using opinion mining, it is easy to distinguish reviews into positive and negative review.

III. RELATED WORK

Opinions and sentiments expressed in text reviews can be generally analyzed at the document, sentence, or even phrase (word) levels. The purpose sentence-level opinion mining is to classify the overall subjectivity or sentiment expressed in an individual review document (sentence). To prevent a sentiment classifier from considering irrelevant or even potentially misleading text, Pang and Lee [8] proposed to first employ a sentence-level subjectivity

detector to identify the sentences in a document as either subjective or objective, and later discarding the objective ones. Then the resulting subjectivity extract was applied via sentiment classifier, with improved results. Review rating assessment is a much more complicated problem compared to binary sentiment organization. Generally, opinions are expressed in a different way in different domains. The sentiment classification methods mentioned above can be tuned to work very well on a given domain; however, they may fail to classify opinions in a different domain.

Bollegala [2] proposed a fully automatic method to create a thesaurus that is sensitive to the sentiment of words expressed in different domains. It utilizes both labelled and unlabeled data available for the source domains and unlabeled data from the target domain. A fundamental problem when applying a sentiment classifier trained on a particular domain to classify reviews on a different domain is that words (hence features) that appear in the reviews in the target domain do not always appear in the trained model. To prevail over this feature difference problem, the author constructed a sentiment sensitive thesaurus that captures the relatedness of words as used in different domains.

Bing Liu [7] studied the problem of generating feature-based summaries of customer reviews of products sold online. Here, features broadly mean product features (or attributes) and functions. Given a set of customer reviews of a particular product, the task involves three subtasks: identifying features of the product that customers have expressed their opinions on (called product features); for each feature, classifying review sentences that give positive or negative opinions; and producing a summary using the discovered information. B. Liu also proposed an association rule mining (ARM) approach to mine frequent item sets as probable opinion features, which are nouns and noun phrases with high sentence-level frequency (or support). However, ARM, which relies on the frequency of item sets, has the following limitations for the task of feature identification, frequent but invalid features are extracted incorrectly, and rare but valid features may be overlooked.

IV. METHODOLOGY

An opinion feature such as “battery” in mobile reviews is typically domain-specific. That is, the feature appears often in the given review domain, and infrequently outside the domain such as in a domain-independent corpus. As such, domain-specific opinion features will be mentioned more often in the domain databases of reviews, compared to a domain-independent corpus.

Figure.1 shows the workflow of our proposed method. Given a domain-dependent review database and a domain independent database, first a list of candidate features from the review corpus are extracted by manually defined syntactic rules defined in TABLE I. For each extracted candidate feature, its IDR is calculated, which represents the statistical relation of the candidate to the given domain corpus, and its extrinsic relevance, which gives the statistical relation of the candidate to the domain-independent database.

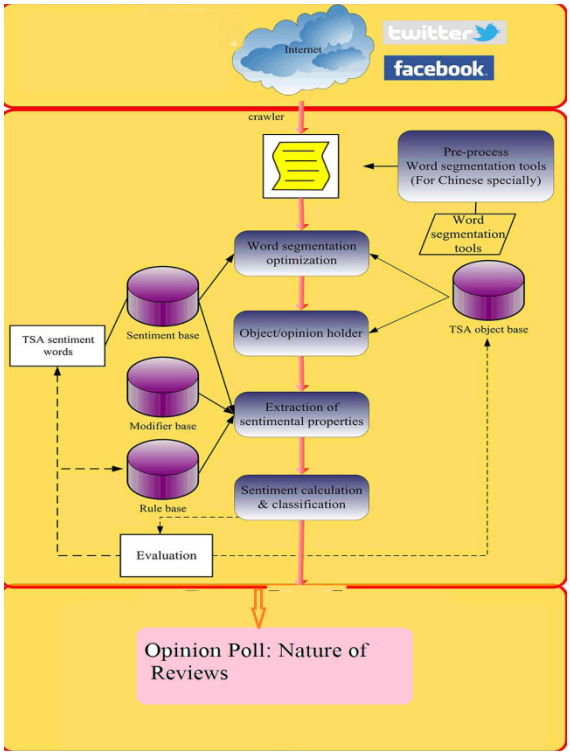

Figure 1. Proposed System Architecture

A. Extracting Candidate Features

The reviews are extracted from any website using web crawler and opinion features are to be identified. Opinion features are generally nouns or noun phrases, which normally appear as the subject or object of a review sentence. In the case of dependence grammar [10], the subject opinion feature has a syntactic relationship of type subject-verb (SBV) with the predicate. The object opinion feature has a dependence association of verb-object (VOB) on the predicate. In addition, it also has a dependence association of preposition-object (POB) on the prepositional word in the sentence.

TABLE I. SYNTACTIC RULES

Rules	**Interpretation**
NN+SBV→CF	Identify NN as a CF, if NN has a SBV dependency relation
NN+VOB→CF	Identify NN as a CF, if NN has a VOB dependency relation
NN+POB→CF	Identify NN as a CF, if NN has a POB dependency relation

Given example illustrates the corresponding dependence tree in Figure. 2. As shown in example, the opinion feature “price” (underline), which is associated with the adjective “expensive” (italic), is the subject of the sentence. It has a dependence association of SBV with the adjective predicate. From various dependence relations we can represent three syntactic rules “NN” and “CF” denotes nouns phrases and candidate features.

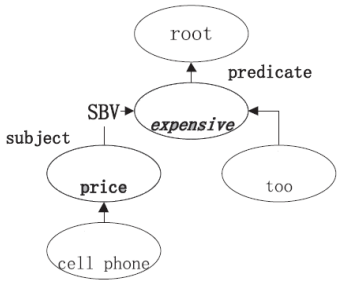

Figure 2. SBV Dependency Relation

The candidate feature extraction process works in the following steps: 1) Dependence parsing (DP) is first employed to identify the syntactic structure of each sentence in the given review corpus; 2) the three rules are applied to the identified dependence structures, and the equivalent nouns or noun phrases are extracted as candidate features whenever a rule is fired. There could be many invalid features in the extracted candidate feature list, the next step is to prune the list using proposed algorithms.

B. Identify Opinion Features

Domain relevance characterizes how much a term is related to a particular corpus (i.e., a domain) based on two kinds of statistics, namely, dispersion and deviation. Dispersion quantifies how significantly a term is mentioned across all documents by measuring the distributional significance of the term across different documents in the entire corpus (horizontal significance).

Deviation reflects how frequently a term is mentioned in a particular document by measuring its distributional significance in the document (vertical significance). Both dispersion and deviation are calculated using the well-known term frequency-inverse document frequency (TF-IDF) term weights. Each term T_i has a term frequency TF_{ij} in a document D_j, and a global document frequency DF_i. The weight w_{ij} of term Ti in document D_j is then calculated as follows:

$$w_{ij} = \{(1+\log TF_{ij}) * \log \frac{N}{DFi.} \text{if } TF_{ij} >0\}, \quad \text{otherwise } 0 \tag{1}$$

where i are total number of terms and j are total number of documents in the corpus.

Dispersion thus measures the normalized average weight of term Ti. It is high for terms that appear frequently across a large number of documents in the entire corpus. The dispersion $disp_i$ of each term Ti in the corpus is defined as follows:

$$disp_i = \overline{w_t} / s_i \tag{2}$$

where s_i is standard deviation.

Deviation $devi_{ij}$ indicates the degree in which the weight w_{ij} of the term T_i deviates from the average w_{ij} in the document D_j. The deviation thus characterizes how significantly a term is mentioned in each particular document in the corpus.The deviation $devi_{ij}$ of term Ti in document Dj is given by

$$devi_{ij} = w_{ij} - \overline{w_j} \tag{3}$$

where the average weight w_j in the document D_j is calculated over all M terms as follows:

$$w_j = \frac{1}{M} \sum_{i=1}^{M} w_{ij}$$

The domain relevance dr_i for term T_i in the corpus is finally defined as follows:

$$dr_i = disp_i * \sum devi_{ij} \tag{4}$$

Clearly, the domain relevance dr_i incorporates both horizontal (dispersion $disp_i$) and vertical (deviation $devi_{ij}$) distributional significance of term Ti in the corpus which reflects the ranking and distributional characteristics of a term in the entire corpus.

C. Intrinsic and Extrinsic Domain Relevance

The domain relevance of opinion words, which is done on a domain-specific review database, is called intrinsic-domain relevance. Similarly, the domain relevance of those opinion words calculated on a domain-independent database is called extrinsic-domain relevance. IDR gives the frequency score of the feature to the domain review corpus (e.g., mobile reviews), while EDR gives the frequency score of the feature to the domain-independent database. Naturally, a candidate term is relevant to either one or the other, but not both. As such, EDR also characterize the insignificance of a feature to the given domain review corpus.

Granted, there do exist some relatively common terms that are used almost everywhere and also in a review corpus as features. For example, the term “price” usually appears as a feature in many review domains, such as mobile and hotel reviews. Therefore, the success of the proposed work approach gets down to the careful selection of a domain independent corpus that is as different from the domain specific review corpus as possible.

D. The IEDR Algorithm

The procedure for computing the domain relevance is the same regardless of the corpus, as concise in Algorithm 1. When the procedure is applied to the domain-specific review database, the scores are called IDR, otherwise they are called EDR. Candidate features with overly high EDR scores or gloomily low IDR scores are eliminated using the inter-corpus criterion of IEDR. Algorithm 1 gives the proposed IEDR method, where the minimum IDR threshold i^{th} and maximum EDR threshold e^{th} can be determined experimentally.

Algorithm 1:

Input:

- Domain review corpus R and domain independent corpus D

Output: A validated list of opinion features with the nature of reviews

Method:

- Begin
- Extract candidates from the review corpus R;
- For each candidate feature CF_i do

- Compute IDR score idr_i using (1), (2), (3), (4) on the review corpus R;
- Compute EDR score edr_i using (1), (2), (3), (4) on domain independent corpus D;

- If ($idr_i \geq i^{th}$) AND ($edr_i \leq e^{th}$) then
  - Confirm candidate CF_i as a feature

  End for
- Return a validated set of opinion features

V. EXPERIMENTAL EVALUATION AND RESULTS

Various experiments are conducted to comprehensively evaluate the IEDR performance on two real-world review domains, mobile and hotel reviews. First reviews are extracted from a website using crawler. Then using sentiment analyzer frequently occurring features are extracted. Secondly, the weights of opinion words is calculated using term frequency tool which assigns a weight comparing it to a threshold of initial testing on a database. Then these features denote the nature of reviews using sentiment analysis via IEDR method. Various methods are compared to the proposed method as follows:

- Intrinsic-domain relevance (IDR), which uses only the given review corpus to extract opinion features,
- Extrinsic-domain relevance (EDR), which uses only the domain-independent corpus to extract opinion features,
- Latent Dirichlet allocation (LDA) [9], which is a generative probabilistic graphical topic model,
- Association rule mining (ARM) [7], which mainly discovers frequent nouns or noun phrases as opinion features,
- Mutual reinforcement clustering (MRC) [4],
- Dependency parsing (DP) [10], which uses synthetic rules to extract features.

TABLE II. SUMMARY OF EVALUATED RESULTS

Method	Characteristics	Remarks
IEDR	Intrinsic and Extrinsic domain relevance criterion	Outperforms the best and gives exact nature of reviews
IDR	Intrinsic domain relevance	Performs best for any review database
EDR	Extrinsic domain relevance	Performs best only if the reviews are domain independent
LDA	Topic modeling	Performs only for related topics
ARM	Frequent item set mining	Performs for mining datasets which are frequently occurring
MRC	Mutual reinforcement principle	Used for calculating scores of frequently occurring features
DP	Dependence parsing	Finding relation between words

A. Precision and Recall

Firstly candidate features from the review domains used i.e. hotel reviews and mobile reviews are extracted, using the syntactic rules defined in TABLE I. Based on the same set of features extracted, a graph of precision and recall is plotted taking the existing results of methods summarized in TABLE II. Different methods and their characteristics are mentioned in the table. After extracting features, the opinion poll is calculated which gives the rating for the extracted features out of 10. Percentage of the opinion poll is calculated which describes how frequently the feature occurred in the corpus.

The final results are still in progress. The final results will be implemented online using a server which will be useful for the users to post their reviews, can make decisions using other reviews and understand the nature of reviews which gives result into positive and negative review opinion poll in tabular form. The existing system was machine oriented; it did not post the result on the web server.

B. Discussion

The opinion feature extraction performance of IEDR (as well as all the competing methods) on the hotel reviews is not as good as that on the mobile reviews. This is because hotel reviews

- are longer and more complex, which makes feature mining much more challenging, and
- Contain large number of noisy domain-unrelated user reviews, which overlap with documents in the domain-independent corpus. In other words, the division between the intrinsic-domain relevance and the extrinsic-domain relevance of a feature is not that clear-cut for the hotel reviews.

As a result, IEDR (and all other competitors) is less successful in identifying features on the hotel domain compared to the mobile domain .Given a particular review domain, IEDR repayment greatly from a highly different domain-independent corpus for opinion feature extraction. It is, thus, very difficult to choose the right domain-independent corpus. According to the experiments carried out, neither corpus size nor topic number in corpus has a large effect on IEDR feature extraction, but the different nature of the domain-independent corpus/topic from the given review domain makes a big difference.

VI. CONCLUSION

The proposed model is an original inter-related data approach to opinion feature classification based on the IEDR feature-filtering principle, which uses the disparities in distributional description of features across two databases, one related domain and one correlated domain. IEDR identifies candidate description that are specific to the given review domain and yet not overly basic (domain independent). Experimental results reveal that the proposed IEDR not only leads to noticeable improvement over either IDR or EDR, but also overcomes the limitations observed in four conventional methods,

namely, LDA, ARM, MRC, and DP, in terms of feature withdrawal performance as well as opinion mining results. Moreover, as domain related database is important, it also requires good size data and topic selection on features on which the performance is calculated. Using a domain-independent corpus of a similar size will give way good opinion feature extraction results are found.

ACKNOWLEDGMENT

First and foremost, I would like to express my sincere gratitude to my guide Ms. Sweta Kale, for her continuous support and head of the department Ms. D.T. Kurian in completion of my project stage-I. I will always be grateful for her patience, motivation, enthusiasm, and immense knowledge. This work is an extension of the previous research efforts of Zhen Hai, Kuiyu Chang, Jung-Jae Kim and Chrishtopher C. Yang. I am very much indebted to them for their inspiring work which is a boost for future researchers. Finally, I would like to pay my respect and love to my parents, for their encouragement throughout my career.

REFERENCES

[1] Hai et al., "Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance", IEEE Transactions on Knowledge and Engineering, Vol 26, No. 3, March 2014.

[2] D. Bollegala, D. Weir, and J. Carroll, "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 8, pp. 1719-1731, Aug. 2013.

[3] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167, May 2012.

[4] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," Computational Linguistics, vol. 37, pp. 9-27, 2011.

[5] N. Jakob and I. Gurevych, "Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 1035-1045, 2010.

[6] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su,"Hidden Sentiment Association in Chinese Web Opinion Mining," Proc. 17th Int'l Conf. World Wide Web, pp. 959-968, 2008.

[7] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 168-177, 2004.

[8] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics, 2004.

[9] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, Mar. 2003.

[10] L. Tesniere, Elements de la syntaxe structurale. Librairie C. Klincksieck, 1959.