Opinion Feature Extraction via Domain Relevance

Vaishnavi S. Baste
Computer Department,
MET BKC Adgaon, Nasik, Savitribai Phule Pune University, Maharashtra, India

ABSTRACT
Rich web resources such as discussion forum, review sites, blogs and news corpus available in digital form, tends the current research to focus on the area of sentiment analysis. Researchers are intended to develop a system that can identify and classify opinion or sentiment as represented in an electronic text. Accurate prediction methods can enable us, to extract opinions from the internet and make predictable decisions which will help economic or marketing research. The majority of existing mining approaches for opinion feature extraction depend on a single review corpus, ignoring word distributional characteristic across different domain. In this paper, a novel method is proposed to recognize opinion features from online assessment by determining the difference in opinion feature across two corpora, one domain-related corpus and one domain-independent corpus, which is a variant in method proposed in [1].

Keywords
Sentiments, opinion features, opinion mining.

1. INTRODUCTION
Opinions are required for all humans when they need to make some decision and opinions manipulate a lot on human actions and choices made. Our choices towards any object mainly depend on how others feel about the same object. Because of this reason, when we need to make a verdict we seek out the judgment and opinion of others. Organizations follow the same process and hence organizations conduct different types of information gathering through survey, interviews to know feedback from their customers. Sentiment analysis with help of natural language processing works to sketch the mood of the customers about a particular product. Sentiment analysis also called opinion mining collects data from online review, inspect the data and after some processing draw conclusions on review data specifying whether negative or positive opinions are expressed in review [3]. A sample review taken from website is given as below

"I bought an iPhone and my friend bought a Samsung Grand a month ago. Its picture quality is amazing but the photos from my phone are not that great and battery life is short too. My friend is happy with his phone and I am going to buy a new Samsung Grand tomorrow."

The above sample review expresses contradictory opinions related with different attributes or aspects of cellphone. A positive opinion is made on the cell phone for its picture quality but battery life is too short which poses a negative opinion on the review. Smart consumers nowadays are no longer satisfied with just the overall opinion rating of a product. They want to understand why it receives the rating, that is, which the positive aspects of the product are and which are the negative attributes of product that contribute to the final rating of the product. It is important to mine the exact opinion features from reviews and classify them to fie grained opinions [1].

In opinion mining, an opinion feature indicates an entity towards which user express their specific opinions. This paper proposes an approach to the identification of such features from unstructured textual reviews.

1.1 Levels of Sentiment Analysis
In general, sentiment analysis has been classified mainly at three levels:

1. ***Document level:*** This level classify a whole opinion document such that it expresses a positive or negative sentiment. For example, given a movie review, the system simply identifies whether the review expresses positive opinion towards a movie or negative opinion. It does not describe what exactly is liked or disliked.
2. ***Sentence Level:*** This level increases the granularity of opinion mining and sentences wise extraction of opinions take place. Thus, it determines whether opinion expressed by sentence is positive, negative, or neutral.
3. ***Entity and Aspect level:*** Both analysis levels fail to associate corresponding likes and dislikes of people towards a product. Aspect level performs finer grained analysis. Instead of looking at language constructs, aspect level straightforwardly looks at the opinion itself.

The rest of the paper is structured as follows: Section 2 presents the literature review and related work, and Section 3 describes proposed method. Section 4 comprises of system architecture. Section 5 describes algorithms used and finally we draw conclusions in end.

2. LITERATURE REVIEW
Natural language processing (NLP) have extensive history; but little investigations had been done about people's opinions and sentiments. After the year 2000 many researches are done in field of NLP. Reasons behind this comprise widespread use of internet and technology. Sentiment has a wide variety of applications in different domains. Customers try to analyze the review, opinion of other people before buying any product. This provides a strong motivation for research [3]. It also offers many new research problems, which must be studied to improve performance in Sentiment analysis. Opinion text was available in digital form before year 2000. Since the year 2000, the field has grown and majority researches are focused in NLP [3].

2.1 Related Work
W. Jin and H. Hay [4] proposed a Lexicalized HMM-based approach to extract opinions from online reviews. They proposed a framework that is capable of classifying product linked entities from product reviews. The system initially recognizes probable product related other objects and opinion of such entities from the reviews. It then extracts opinion sentences which show each identified product entity. This

then finally determines opinion course for each identified product entity. N. Jakob and I. Gurevych [5] employ the supervised algorithm which represents the state-of-the-art on the employed data. They had used Conditional Random Fields (CRF) for opinion targets extraction which tackles the problem of domain portability. Their proposed work evaluates the performance of the system in cross domain. S. M. Kim and E. Hovy [6] proposed a method extraction of opinion in the form of triplets as opinion, holder, and topic for identifying an opinion from online news media texts. They identified word expressing opinion and assigning labels describing semantic roles to the word in the sentence. It also finds the holder and the opinion word among the labeled semantic roles.

B. Pang, L. Lee and S. Vaithyanathan [7] employed document level opinion mining. Rather classifying the document on topic basis, they classified the overall sentiment determining whether a review is positive or negative. They had used this technique to classify the movie review and rate it using thumbs up and thumbs down approach. Ratings were automatically extracted and converted into one of three categories: positive, negative, or neutral. All the above methods classify opinions within given domain and does not calculate its relevance score. Proposed method enables us to calculate domain relevance score against different domains by calculating its intrinsic and extrinsic domain relevance, which helps to give more accurate opinions results in regard with the particular domain.

3. PROPOSED METHOD

The proposed novel method identifies opinion features from online reviews by utilizing the dissimilarity in opinion feature across two domain sets, one domain-specific corpus and one domain-independent corpus. The proposed method captures this difference via a measure called domain relevance (DR). Domain Relevance resembles the relevance of a term in a particular domain.

First a list of opinion features is created which are initially candidates of opinion features. Opinion Features are extracted from online review using a set of syntactic dependency rules. For each extracted opinion feature we have estimated two scores Intrinsic Domain Relevance(IDR)and Extrinsic Domain Relevance (EDR). The domain relevance is calculated with help of opinion and its reliance upon the domain. Opinion feature when calculated on a domain related review gives *intrinsic-domain relevance.* In the same fashion, opinion feature when computed on a different, independent domain gives *extrinsic-domain relevance*.IDR represents how much the candidate feature is related to the given domain corpus and EDR represents the relevance of the candidate to the domain independent corpus. High IDR and low EDR of the candidate are expected. Candidate features that are less generic and more domain-specific are then confirmed as opinion features. This thresholding approach is called the intrinsic extrinsic domain relevance (IEDR) criterion [1].

4. SYSTEM ARCHITECTURE

4.1 System Flow

1. Several syntactic dependence rules are designed which are used to extract a catalog of candidate features from online review, for example, cell phone or hotel reviews.
2. For each documented candidate feature, its domain relevance score is calculated with respect to domain-specific review and domain independent review. For domain-specific review corpus we calculate intrinsic-domain relevance and for domain independent corpus we calculate extrinsic-domain relevance.
3. After computation of domain relevance, candidate features having low IDR scores and good EDR scores are pruned.

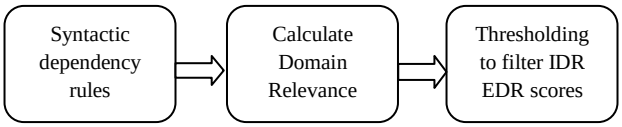

Fig 1: Proposed method workflow

4.2 Candidate Feature Extraction

Candidate features are extracted in the following manner: for each word, first determine if it is a noun if so, apply the Verb Object (VOB), Subject Verb (SBV), and Preposition Object (POB) rules sequentially. A noun matching any of the rules is extracted as a candidate feature. In addition, when a clause contains only a noun phrase without any verbs, the headword of the noun phrase is also a candidate. Due to the colloquial nature of online reviews, it is complicated and nearly impossible to collect all possible syntactic roles of features [8]. Thus, only three aforementioned primary patterns are used to extract an initial set of candidates. Dependence Grammar explores asymmetric governor dependent relationship between words, which are then combined into the dependency structure of sentences. The three dependency relations SBV, VOB, and POB correspond to the three aforementioned patterns. For each relation, a rule is defined with additional restrictions for candidate feature extraction, as shown in Table 1.

Table1. Candidate Feature Extraction

Relation	Rules	Interpretation
VOB	(NN ,VOB) => CF	If term is noun and depends on anothercomponent with relation VOB, extract as candidate
SBV	(NN, SBV) => CF	If term is noun and depends on anothercomponent with relation SBV, extract as candidate
POB	(NN, POB) => CF	If term is noun and depends on anothercomponent with relation POB, extract as candidate

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 in Table 1 are applied to the identified dependence structures, and the corresponding nouns are extracted as candidate features whenever a rule is fired.

4.3 Methodology

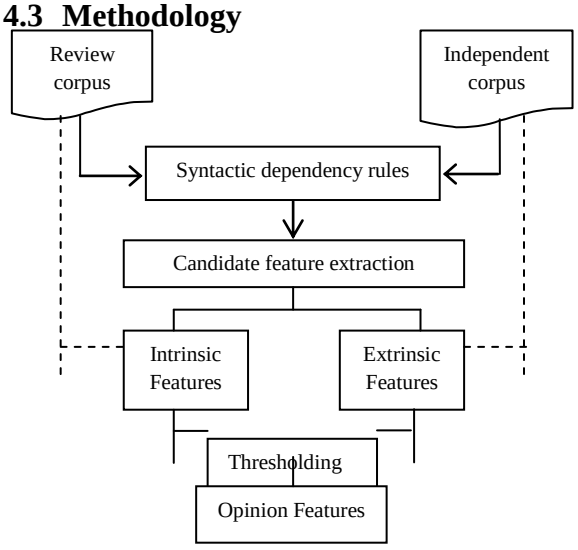

Fig 2: System Architecture

Fig2 shows the architecture of proposed method using a domain-dependent review corpus and a domain independent corpus. In first step extract a list of candidate features from the review corpus via manually defined syntactic rules. These rules must be applied to domain dependent and domain independent corpus to extract the noun features. Candidate features are extracted in the following manner: for each word, first determine if it is a noun; if so, apply the VOB, SBV, and POB rules sequentially as shown in Table 1. A noun matching any of the rules is extracted as a candidate feature [8]. For each extracted candidate feature, estimate its intrinsic domain relevance, which represents the statistical association of the candidate to the given domain corpus, and extrinsic- domain relevance, which reflects the statistical relevance of the candidate to the domain independent corpus. Only candidates with IDR scores exceeding a predefined intrinsic relevance threshold specified by user and EDR scores less than another extrinsic relevance threshold are confirmed as valid opinion features. In short, identify opinion features that are domain-specific and at the same time not overly generic via the inter-corpus statistics IEDR criterion [1].

5. ALGORITHMS USED

5.1 Pearson Correlation

The correlation coefficient is a measure of how well two domain data fit on a straight line. A correlation of 1 means both domains have perfect positive linear relationship and -1 indicates negative relationship [9].

Pearson correlation is defined by the following equation. x and y represents two cross domain dataset -

$$corr(x,y) = \frac{c_i(x,y)}{s_i(x) \times s_i(y)}$$

where

$c_i$is covariance between x and y

s_i is standard deviation and is calculated as -

$$s_i = \sqrt{\frac{\sum_{j=1}^{N}(w_{ij} - \overline{w}_i)^2}{N}}$$

TF_{ij} = term frequency for term Ti,

D_j = document

DF_i = global document frequency

Now, weight w_{ij} of T_i in D_j is calculated as follows*:*

$$w_{ij} = \begin{cases} (1 + \log TF_{ij}) \times log(\frac{N}{DF_i}) & if TF_{ij} > 0, \\ 0 & otherwise \end{cases} \quad (1)$$

where-

i = 1, 2.....M for total number of M terms

j = 1, 2.....N for total number of N documents

The average weight $\overline{w}_i$ of term T_i across all documents is calculated by:

$$\overline{w}_i = \frac{1}{N}\sum_{j=1}^{N} w_{ij}$$

5.2 Algorithm for Calculating Cross Domain Relevance

The procedure for computing the domain relevance is the same regardless of the corpus. When the procedure is applied to the domain-specific review corpus, the scores are called IDR, otherwise they are called EDR.

Input: A domain specific or domain independent corpus

Output: Domain relevance scores

for*each candidate feature* **do**

for*each document in the corpus C***do**

Calculate *weight*

Calculate *dispersion*

Calculate *standard deviation*

Calculate *correlation*

for*each document in the corpus C***do**

Calculate *deviation*

Calculate *correlation*

Compute *domain relevance*

return A list of domain relevance (IDR/EDR) scores for all candidate features;

Candidate features with overly high EDR scores or miserably low IDR scores are pruned using the intercorpus criterion of IEDR, where the minimum IDR threshold i^{th} and maximum EDR threshold eth can be determined experimentally.

6. CONCLUSION

This paper proposes a variation in intercorpus information method to extract opinion feature based on the IEDR filtering criterion. This utilizes the different word characteristics features across two corpora, one domain-specific and one domain-independent. To improve the confidence we have used distance measure to calculate domain relevance. IEDR recognizes candidate features that are more specific to the given review and yet not very broad. In addition, since two domains with maximum different features and characteristics are important for the proposed approach. Also size of both domains may be same or different will yield good opinion feature extraction results[1]. In future work, non-noun feature extraction would be considered and also fine grained topic

modeling would be employed for accuracy of opinion features. Fine grained topic modeling will improve the accuracy and efficiency of the result. This technique would also be used for extraction of opinions in different languages. Neutral opinion will be considered as one of the significant response from the user.

7. REFERENCES

[1] Zhen Hai, Kuiyu Chang, Jung-Jae Kim, and Christopher C. Yang, "Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance," IEEE Transactions on Knowledge and Data Engineering, Vol.26 No.3, March 2014, pp 623-634.

[2] G. Vinodhini, RM. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," Proceedings of International Journal of Advanced Research in Computer and Software Engineering, Vol. 2, June 2012.

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

[4] W. Jin and H.H. Ho, "A Novel Lexicalized HMM- Based Learning Framework for Web Opinion Mining," Proc. 26th Ann. Int'l Conf. Machine Learning, pp.465-472, 2009.

[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] S. M. Kim and E. Hovy, " Extracting Opinions, Opinion Holders, and Topics Ex- pressed in Online News Media Text," Proc. ACL/COLING Workshop Sentiment and Subjectivity in Text, 2006.

[7] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment Classification Using Machine Learning Techniques," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 79-86, 200.

[8] Z. Hai, K. Chang, Q. Song, and J.-J. Kim, "A Statistical NLP Approach for Feature and Sentiment Identification from Chinese Reviews," Proc. CIPS-SIGHAN Joint Conf. Chinese Language Processing, pp. 105-112, 2010.

[9] Priyanka U Chavan, P M Yawalkar and D V Patil. Article: A Hybrid Approach for Recommendation System in Web Graph Mining. International Journal of Computer Applications 95(24):23-27, June 2014

[10] Bollegala, Danushka, David Weir, and John Carroll. "Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[11] Manjunathan, N. "Cross-Domain Opinion Mining Using a Thesaurus in Social Media Content."

[12] Kumar, Guntupalli Manoj, and B. Gobinathan. "Sentiment classification of Customer reviews for online products using Cross Domain Sentiment Classifier."