Extracting Features in Opinion Mining using Inter Dependent Domain Relevance

1R. Abirami and 2S. Ramesh

1Student of Computer Science and Engineering, Anna University Regional Centre Madurai, Tamilnadu, India
2Faculty of Computer Science and Engineering, Anna University Regional Centre Madurai, Tamilnadu, India

ARTICLE INFO

Article history:
Received 12 October 2014
Received in revised form 26 December 2014
Accepted 1 January 2015
Available online 25 February 2015

Keywords:
Feature Extraction, Dispersion, Deviation, Domain Relevance, camera, iPod.

ABSTRACT

Opinion mining is the task of extraction and analyzes the user comments and classifies the reviews, documents into Positive or Negative. It analyzes the user view on several domains such as product, movie, news media, social blogs etc. The automatic extraction of opinion from the old user reviews is more prominent to new users and developer. Normally the opinion mining is carried out in single domain. In multi domain, the extraction of relevant features from the reviews is more difficult. The Inter Dependent Domain Relevance (IDDR) mechanism is extracting the relevant and contrast domain features using domain relevance score. The features are extracted with the help of the Part-Of-Speech (POS) tagging tool. The tool generates the POS of each word in the reviews and the Noon (NN) is considered as Feature. The domain relevance score is used to predict the relevance of those features in the domain. Domain relevance score is calculates with the help of the Dispersion and Deviation. The domain relevance score is greater than the threshold value then the features are selected. The common feature is extracted and important features in that domain are also extracted. The performance of Inter Dependent Domain Relevance is evaluated using two real word domains like Canon S100 and iPod over intrinsic domain relevance and extrinsic domain relevance algorithm.

INTRODUCTION

The textual information consists of two parts they are fact and opinion. Nowadays, the opinion mining task is mainly focus on web due to the large volume of opinionized text. Opinions are more important because whenever the people want to buy any product or deal any situation or characters of famous person they want to hear others opinion about it. In olden, opinion mining deals the two important terms such as opinion from individual (family and friends) and business. The opinion in the form of surveys, focus groups and consultants.

Opinion mining is also termed as sentiment analysis. The prediction of opinion, sentiments and emotions expressed in the text is main goal of opinion mining. The people express their opinion on anything such as movie reviews, forum, discussion groups and blogs like Facebook; twitter etc. product reviews like amazon, flip cart etc. Day-by-day it will increases due to easy accessibility of reviews, document on the web. Machine learning in natural language processing and information retrieval were increased due to development of practical method at making these widely available corpora. Recently many researchers focus on opinion mining and sentiment analysis. They are trying to fetch opinion information and analyze it automatically with computers. In product based opinion mining improve the productivity, quality etc. in developer side and the purchasing is efficient in customer side. In news blog or news media based opinion mining, the different communities, organization; people express their opinion in different form. Several blog reviews, social sites review based pinion mining is worked by researchers.

The Part-Of-Speech tagging is the method to the features and opinions are extracted with help of the Part-Of-Speech tagging tools. Several free tools available in online or offline to extract the features and opinion from the given reviews. Opinion words are adjective and features are nouns. Consider the following example. “This is good book” – this/DT is/VBZ good/JJ book/NN. In above sentence, book (product feature) is noun(NN) and good (opinion word) is adjective(JJ). In Part-Of-Speech tagging, each word in review is tagged with its Part-Of-Speech such as noun(NN), adverb (RB), verb(VB), conjunction (CC), pronoun (PRP) etc. After POS tagging,

Corresponding Author: R. Abirami, Student of Computer Science and Engineering, Anna University Regional Centre Madurai, Tamilnadu, India
E-mail: abiucev@gmail.com
now it is ready to retrieve nouns as features and adjectives as opinion words. The freely available POS taggers like Stanford POS tagger, Tree tagger, CRF tagger etc.

Based on domain relevance score, the features present in various reviews can be selected for classification. Initially the features are extracted and pruned. Finally the relevant features only selected for further classification. The performance of IDDR algorithm is appropriately high compared to intrinsic domain relevance and extrinsic domain relevance algorithm. The proposed method considered two domains such as Canon S100 and iPod. The method called Inter Dependent Domain Relevance (IDDR) is predicted the features with more accuracy compared to the Intrinsic Domain Relevance (IDR), Extrinsic Domain Relevance (EDR) and Intrinsic Extrinsic Domain Relevance (IEDR).

The paper is organized as the following sections. Section II describes the related work of Feature based Opinion mining. Section III portrays the Inter Dependent Domain Relevance. The Experimental results are shown in the Section IV.

2. Related Work:

Intrinsic Extrinsic Domain Relevance[Zhen Hai et al. (2014)] method extracts the features from two different domains. The product reviews[A. Popescu and O. Etzioni (2005)] are crawled and find the opinions related to the product. The rating is available in the website to describe the opinion about the product. Hotel reviews are considered and predict the opinion about the particular hotel. Opinion about product and political candidates are used in SentiWordNet [A. Esuli et al. (2006)]. The automatic extraction of opinion of PN-polarity of subjective term word net is used to estimate the opinion into three numerical scores. They are obj(s), pos(s), neg(s). Identify how the term contained in SYNSET, 3 scores are derived by combining the results produced by a committee of eight ternary classifier. Within opinion mining, several subtasks are available; determining text SO-polarity, Determining Text PN polarity, determining the strength of text PN-polarity. A simple unsupervised learning algorithm is used to classify the reviews in the form of recommended (thumbs up) or not recommended (thumbs down) [P.D. Turney (2002)]. The classification is carried on with the help of semantic orientation of the given phrases in the reviews that contain adjective and adverbs. The association is predicted based on semantic orientation. The good association is revealed by positive semantic orientation, the bad association is revealed by negative semantic orientation. The average semantic orientation is calculated, based on this decision is made. Finally the classification is done as recommendation or not recommended.

The PMI-IR (Point wise Mutual Information- Information Retrieval) algorithm [P.D. Turney (2002)] is to estimate the semantic orientation of a phrase. This method has measure the similarity of pair of words or phrases to a positive reference word (“Excellent”) with its similarity to a negative reference word (“poor”). Joint structure tagging [F. Li et al. (2010)] method for review mining using conditional random field mechanism. Rich features to jointly extract the positive, negative opinion and object feature is made. Linguistics representation is integrated into modular representation, instead of linear chain. The design structure and syntactic tree structure is created (object feature in which the opinion expressed on). The generation of summarization is more useful to new users and producers. Basically the opinions are ranked by their frequency. The method for identification an opinion with its holder and topic of online news media text [S.M. Kim and E. Hovy (2006)] exploiting semantic structure of a sentence attached to an opinion bearing as adjective or verb semantic role labeling method is used as intermediate step to label an opinion holder and topic using data from FrameNet.

Phrase level sentiment [T. Wilson et al. (2005)] analysis checks whether an expression is neutral or polar. Automatic identification of contextual polarity for lare subset is done. Sometimes the entries are tagged with priori prior polarity. They create corpus and add contextual polarity judgment to the existing annotations in the multi-perspective question answering (mpqa) opinion corpus annotations of subjective expressions. Subjective expressions is any word or phrase used to express an opinion, emotion, evaluation, speculation etc. annotators were instructed to tag the polarity of subjective expression as positive, negative, both or neutral. Multi-knowledge based approach is proposed which include wordnet [L. Zhuang et al. (2006)], statistical analysis and domain knowledge. Mining and summarization method is different from text mining. Initially identification of the feature is carried out, then identifies the opinions wordnet is used to generate a keyword list using movie cast and labeled training data. The grammatical rules are used to identify the features and opinion pair. Finally organize the sentence. Most of the review mining and summarization is concentrate on product reviews. But here focus on different domain called movie review. It has unique characteristics. The user wrote a comment for a particular movie not only a movie element (e.g. screenplay, vision effects, music) and also movie-related people (director, actor and screenwriter).

Inter Dependent Domain Relevance:

The extraction of relevant features from the reviews is more difficult in multi domain. Whereas the Inter Dependent Domain Relevance (IDDR) mechanism is extracting the relevant and contrast domain features using domain relevance score. The features are extracted with the help of the Part-Of-Speech (POS) tagging tool. This tool generates the POS of each word in the reviews and the Noun is considered as Feature. Features (NN) are
extracted based on the outcome of the tool. Based on the feature count, the Domain relevance score is generated. It is used to predict the relevance of those features in the domain. Domain relevance score is calculated with the help of the Dispersion and Deviation. The Dispersion is calculated on basis of average weight and Standard Variance of each term present in the features. Deviation is examined by average weight of each document and weight for term in the non-repeatable features. The domain relevance score is greater than the threshold value the features are selected. The common feature is extracted and important features in that domain are also extracted.

\[ w_{ij} = \frac{f_{ij}}{\sqrt{|D_j|}} \]

\[ d_{ij} = \frac{f_{ij}}{|D_j|} \]

**Fig. 1:** System Architecture.

Figure 1 shows the steps for the extraction of valid set Opinion feature. Consider two different product domain such as Canon S100 domain and iPod domain. The reviews are crawled from those product reviews. The part-of-speech (POS) of each word present in the reviews is identified with the help of the POS Tagger tool called Stanford POS tagger. The nouns are gathered from the tagger tool is considered as features. The non-repeatable features are extracted by eliminating the repeated features. The Domain relevance score is generated with the help of the frequency count of each non-repeatable features present in each reviews. Based on the threshold value the validated set of features is selected.

The reviews belonging to the product domain is send as input for Part-Of-Speech Tagger tool such as Stanford POS Tagger tool. The outcome from the tool is in the form Part-Of-Speech for each word belonging to the given reviews. The Part-Of-Speech tagging is the method of extracting the Part-Of-Speech of each word given in the input space. The features and opinions are extracted with help of the Part-Of-Speech tagging tools. Several free tools available in online or offline to extract the features and opinion from the given reviews. Opinion words are adjectives and features are nouns.

**Fig. 2:** Maximum Entropy Part of Speech Tagger.

Figure 2 demonstrates the simple example for the Part-Of-Speech tagging. Consider the following example. “This is good phone” – this/DT is/VBZ good/JJ book/NN. In above sentence, phone (product feature) is noun (NN) and good (opinion word) is adjective (JJ). In Part-Of-Speech tagging, each word in review is tagged with its Part-Of-Speech such as noun(NN), adverb(RB), verb(VB), conjunction(CC),pronoun(PR) etc.

After POS tagging, now it is ready to retrieve nouns as features and adjectives as opinion words. Normally opinion mining use single domain to classify their thoughts. Some researcher’s perform cross domain sentiment classification to improve the performance of the classification. The Features are normally in the form of Noun. So i have to extract the noun from each tagged reviews for both Canon S100 and iPod. The sentences are crawl from the review are used to extract the features. The features are extracted based on the part of speech.

The dispersion is identified how frequency a term is used across all documents by measuring the distributional significance of the term across different document in the whole domain. The deviation quantifies how frequently a term is used in a particular document by measuring its distributional significance in the document. The dispersion and deviation are calculated with the help of the frequency-inverse document frequency (TF-IDF) term weight. For each \( t_i \) in a document have a term frequency \( TF_i \) in a particular document \( D_j \).

\[ w_{ij} = \frac{f_{ij}}{\sqrt{|D_j|}} \]

\[ d_{ij} = \frac{f_{ij}}{|D_j|} \]

(i) The weight \( w_{ij} \) of each Term \( T_i \) in a particular document is calculated as follows.
\[ w_{ij} = \begin{cases} 1 + \log(1 + \log(TF_{ij})), & \text{if } TF_{ij} > 0 \\ 0, & \text{Otherwise} \end{cases} \quad (1) \]

Where \( TF = \text{Term Frequency} \)
Where \( i = 1 \ldots M \) for a total number of \( M \) terms in a document
Where \( j = 1 \ldots N \) for a total number of \( N \) document in the corpus

(ii) The average weight \( \overline{w_i} \) of term \( T_i \) is calculated as follows

\[
\overline{w_i} = \frac{1}{N} \sum_{j=1}^{N} w_{ij}
\]

(iii) The Standard Variance \( S_i \) for term \( T_i \) is calculated as follows

\[
S_i = \sqrt{\frac{\sum_{j=1}^{N} (w_{ij} - \overline{w_i})^2}{N}} \quad (2)
\]

(iv) The dispersion \( disp_i \) of each term \( T_i \) in the Review Corpus is defined as follows.

\[
disp_i = \frac{\overline{w_i}}{S_i} \quad (3)
\]

(v) The \( \overline{w_j} \) be an average weight for a particular document across all \( M \) terms is given as

\[
\overline{w_j} = \frac{1}{M} \sum_{i=1}^{M} w_{ij}
\]

(vi) The Deviation \( dev_{ij} \) of a term \( T_i \) in a document is expressed as follows.

\[
dev_{ij} = w_{ij} - \overline{w_j} \quad (4)
\]

The deviation is quantified how significantly a term is used in each document in the corpus.

(vii) The Domain Relevance \( dr_i \) is calculated using dispersion and deviation as follows.

\[
dr_i = disp_i \times \sum_{j=1}^{N} dev_{ij} \quad (5)
\]

The value of \( disp_i \) (for term \( T_i \)) is high, which indicates that the term frequently occurred across the entire document in the corpus.

DRS and Threshold value play a vital role in the selection of validated set of features. Based on the threshold value the common and most relevant features are extracted. Set 1.5 as a threshold value for IDDR algorithm.

The performance of the Inter Dependent Domain Relevance Mechanism can be analyzed by comparing my proposed algorithm with Intrinsic Domain Relevance (IDR), Extrinsic Domain Relevance (EDR) and Intrinsic Extrinsic Domain Relevance (IEDR) algorithm. The F-Score of my proposed algorithm is good compared to the above mentioned algorithm.

4. Experimental Results:

This section describes the performance analysis to validate the proposed algorithm. In the Inter Dependent Relevance Score (IDDR) algorithm, the features with more accurately compared to the Intrinsic Domain Relevance (IDR), Extrinsic Domain Relevance (EDR) and Intrinsic Extrinsic Domain Relevance (IEDR) Algorithm.

The performances are measured using the standard evaluation measures of precision \((p)\), recall \((r)\) and F-score \((F)\), \[ F = \frac{2pr}{p+r} \]. They are able to improve the recall dramatically without much loss in precision. The gains in F-scores are dramatic. Naturally as the number of given opinion words increases, the improvement decreases slightly.

![RECALL](image)

**Fig. 3:** Recall.

In Figure 3, the IDDR curve lies well above the IDR, EDR and IEDR curve for all. This is perfectly acceptable since Recall values at high levels are more practical. Across all Recall levels, the largest gap of IDDR over IDR is 0.3 in r3. The Proposed IEDR thus achieved a significant improvement over IDR, EDR and IEDR. In Figure 4, the IDDR curve lies well above the IDR, EDR and IEDR curve for all. This is perfectly
acceptable since Precision values at high levels are more practical. Across all Precision levels, the largest gap of IDDR over IDR is 0.3. The Proposed IEDR thus achieved a significant improvement over IDR, EDR and IEDR. In information retrieval, recall is the fraction of the documents that are relevant to the query that are successfully retrieved. In opinion word extraction or feature extraction algorithm, recall will be the fraction of the relevant opinion words or features that are relevant to that are successfully retrieved.

In Figure 5, the IDDR curve lies well above the IDR, EDR and IEDR curve for all. This is perfectly acceptable since F-Score values at high levels are more practical. Across all F-Score levels, the largest gap of IDDR over IDR is 0.3. The Proposed IEDR thus achieved a significant improvement over IDR, EDR and IEDR. In information retrieval, precision is the fraction of retrieved documents that are relevant to search. Similarly, in opinion word or feature extraction algorithm, precision is the fraction of retrieved opinion words/features that are relevant to search.

Conclusion And Future Enhancement:

Inter Dependent Domain Relevance (IDDR) to opinion feature extraction based on the Domain Relevance Score which utilizes the disparities in distributional characteristics of features across two corpora, one domain-specific (Canon S100) and one domain-independent (iPod). IDDR identifies Valid features that are specific to the given review domain and also specifies the most important features in the both domain. Experimental results demonstrate that the proposed IDDR not only leads to noticeable improvement over either Intrinsic Domain Relevance (IDR) or Extrinsic Domain Relevance (EDR), but also outperforms on Intrinsic Extrinsic Domain Relevance feature extraction results. Both IDR and EDR are in-domain feature extraction mechanism and IEDR is cross-domain feature extraction. The extraction of features is not only the common features but also the most important features in each domain. In addition, since a good quality domain-independent corpus is quite important for the proposed approach.

The F-Score measure indicates that the proposed method has to improve the prediction of relevant feature. This improves the efficient summarization of product reviews using the extracted features. For future work, (1) Consider more than two domains and perform IDDR feature extraction mechanism with improved accuracy. (2) To jointly identify opinion features, including non-noun features, infrequent features, as well as implicit features. (3) Classify the product reviews as positive, negative and neutral polarity as well as weakly positive, mildly positive, strongly positive, weakly negative, mildly negative and strongly negative using the extracted validated set of features. (4) I plan to further test the IDDR opinion feature extraction in several other opinion mining systems. (5) Summarization of features with respect to their opinion sentence or opinion word is performed.

REFERENCES


Dave, K., S. Lawrence, D. Pennock, 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews.


