A Survey Paper on Identifying Candidate Features in Opinion Mining using Intrinsic and Extrinsic Domain Relevance

Janhavi Suchet Vakil

Abstract*: Opinion feature extraction is the process of obtaining candidate features from the existing set of features identified from reviews and opinions. We study few techniques and propose a novel method to identify candidate features using different pattern mining approaches and extract relevant information using a set of syntactic rules. Using Dependency Parsing (DP) we can extract Parts of Speech (POS). The POS can be used to extract candidate features using syntactic rules and thus obtain candidate features. According to previous studies candidate features that are less generic and more domain-specific are then confirmed as opinion features. Previous experimental results on two real evaluation domains show that this approach may surpass several other well-established methods for identifying opinion characteristics.

***Index Terms*: opinion mining. Sentiment analysis, Intrinsic , Extrinsic, Domain Relevance, Stanford NLP..**

I. INTRODUCTION

"What people think?" has always been one of the most important question to every individual during the decision making phase and in planning phase. Recommendations and reviews are often considered in every aspect. It is therefore very crucial for the companies and for the consumers to understand more vividly. Companies would analyze the feedbacks and reviews about their products, thus to make the future decisions about it. Therefore, opinion mining or sentimental analysis plays an important role in information gathering and analyzing the sentiments, opinions and subjectivity with respect to the product. The overall subjectivity analysis is performed at the document level retrieval. The synonyms of Opining mining that helps to understand at a beginner's level are Review mining, Sentiment analysis [1], and Appraisal mining.

The analysis of feeling is the field of study that analyzes people's opinions, feelings, assessments, assessments, attitudes and emotions with respect to entities such as products, services, organizations, Individuals, problems, events, topics. The feelings or opinions expressed in the textual comments are generally analyzed in different resolutions. For example, document-level information retrieval identifies the subjectivity or general feeling expressed on an entity (eg a cell phone or hotel) in a revision document, but does not associate opinions with Specific aspects (eg display, battery) of the entity. This problem also occurs, to a lesser extent, in the extraction of opinion at the sentence level. Many approaches have been proposed to extract the characteristics of opinion in the mines of opinion. The supervised learning model can be set to work properly in a given domain, but the model must be recycled if applied to different domains. Unsupervised natural language (NLP) [1][2] process approaches, determine domain-driven opinion options, independent grammar models, or rules that capture dependency roles and the native context of the terms of functionality. However, the rules do not work well on conversational conversations of real life that have no formal structure. Theme modeling approaches will encompass general and generic themes or aspects that are literally key elements of linguistics or aspects of particular options explicitly commented in journals. Current approaches to corpus statistics attempt to extract the options of opinion options by extracting applied mathematical models from terms characterized only in the given corpus of assessment, while not considering their characteristics of spatial arrangement in another Corpus completely different.

Extraction of information at the document level gives the overall feeling or subjectivity expressed on an entity of the examination document, but does not associate the notice with the revision document. The Document level opinion mining [3] or sentiment analysis analyzes the review sample and gives the result of that the review document is either positive or negative sentiment. Most of the time, the consumers are not satisfied with the product's ratings. They are interested to know the reasons why the products gets its ratings, positive and negative characteristics that has effects on the final ratings of the product. It is therefore essential to mine the precise opinioned features from the reviews and they are to be associated with the opinions. Opinion feature is simply the indication of an entity or an attribute of the entity on which the opinions are expressed by the users.

II. RELATED WORK

A. Literature Survey

Opinions expressed in the textual form of the reviews are analyzed at phrase, sentence and document levels. Hatzivassiloglou and Wiebe [3] have studied the effects of the Semantically oriented adjectives, Dynamic adjectives and Gradable adjectives. This is done to predict subjectivity. Therefore they have proposed a supervised classification method to predict the subjectivity of the sentence. Pang [4] et al proposed the three machine learning method and they are Naïve bayes, Maximum entropy and Support vector machines (SVM) [9].

Revised Version Manuscript Received on June 09, 2017.
Janhavi Vakil, Narsee Monjee Institute of Management Studies, Mumbai (Maharashtra), India. E-mail: janhavi.vakil16@gmail.com

The standard machine learning techniques gave good results when compared to the human generated baselines. To prevent the errors of the sentiment classifier to consider wrongful or rather misleading texts, Pang and Lee proposed a study to first employ and perform the subjectivity detector at the sentence level that further identifies whether the sentence is either an objective sentence or a subjective sentence and thus pruning the objective sentences. They are then forwarded to the sentiment classifier for the subjectivity extraction.

Mcdonald [5] et al investigated to predict the sentiments (opinions) at many different levels of granularity in a textual form of the review with the help of global structured model. This therefore gives an advantage to classify the decisions from one level which can influence the decisions of the other level. Since sentiments are often expressed differently in different kind of domains.

Bollegala [6] et al proposed an unsupervised learning method, a cross-domain sentiment classifier, that reviews the documents as positive (thumbs up) or negative (thumbs down) and thus the sentiments of each review document is analyzed by the average sentiment orientations in the documented review. It's dependency on the search engine was one of the limitations.

Maas [7] et al approach uses a mixture of supervised and unsupervised techniques to classify the sentiments at the sentence level and document level to learn the word vectors by capturing the meaning or the semantics of the term document and the sentiment content as well.

Whenever sentiments analysis is performed at the words level, it mainly checks for its positive or negative opinions (polarities). The opinion of the word (phrase) matters as it is generally context dependent or domain-specific and hence the polarity gives us an opinion over the phrase. Wilson et al presented an approach of predicating of these context-sentiments at the phrase (word) level [10] with the help of machine learning techniques on the various candidate feature factors.

A compositional matrix-space model at the phrase level sentiment analysis was presented by Yessenalina and Carie [8]. The advantage of this approach is that the model can handle the unseen bigrams provided the component unigram is learnt.

Back in 1940s Machine translation was the first natural language related computer based application. Weaver and Booth proposed a project which was based on computer translations subjecting to breaking the enemy codes during the world war two. Thus Weaver originated the ideas for cryptography in the language processing.

Natural language processing (NLP) is the field that focuses on the computer understanding and the manipulation of the human language. It covers the interaction between them. NLP is basically a way from which the computers understand, analyze and derive the semantics and meaning of the human language in the smartest and in a very useful way. Many tasks such as translation, Named entity recognition, sentiment analysis, and Speech recognition are performed using NLP. NLP thus allows the machine to understand how humans speak. NLP plays a key component on Artificial Intelligence (AI) and is depended on the machine learning. Chomsky [14] published the Syntactic Structures leading to better insights in the linguistics and introducing the idea of generative grammar. It further emerged to the speech recognition. Chomsky further introduced the transformational model which had the semantic concerns as the transformational generative grammars were syntactically oriented.

The semantic networks of Quillian [11], conceptual dependencies theory of Schank [12] and case grammar of Fillmore explained the syntactic anomalies and gave semantic representations. Wood's [13] augmented transition networks grew the power of phrase structure by including programming language mechanisms such as LISP.

B. System Architecture

The Text Document or the review document includes both the positive as well as the negative aspect of particular object and an entity in respect with the user's views. Generally, object's overall sentiments may hold both the polarity aspects (positive aspects and negative aspects). Therefore to find the complete aspects about the entity, strong feature-level analysis is require. It therefore requires three major steps:

1. Identifying the object features
2. Determining the opinion orientations
3. Grouping the synonyms

The Rules which are in an improper structure are not able to work well in colloquial or ordinary (familiar) real-life reviews. By applying the approaches of the topic modeling, we can extract the generic topics and the coarse-grained topics which are actually the semantic feature clusters of the precise features which are commented on explicitly in the review documents.

Therefore the proposed method is stated as the following:

1. A number of syntactic dependence set of laws are used to produce a list of candidate features from the domain review corpus.
2. For each candidate with documented functionality for domain- and domain-specific body bodies, we compute the relevance score. Intrinsic-domain relevance (IDR) is known as the domain relevance score on the domain-specific corpora and extrinsic-domain relevance score (EDR) is known as the domain relevance score of the domain independent corpora.
3. Finally, all the candidate features with high EDR and low IDR are removed or discarded. And thus, we call this as interval thresholding the intrinsic and the extrinsic domain relevance criterion (IEDR).

As such, the frequency of the domain-specific opinion features will be more than the domain-independent corpus for a given domain corpus of reviews.

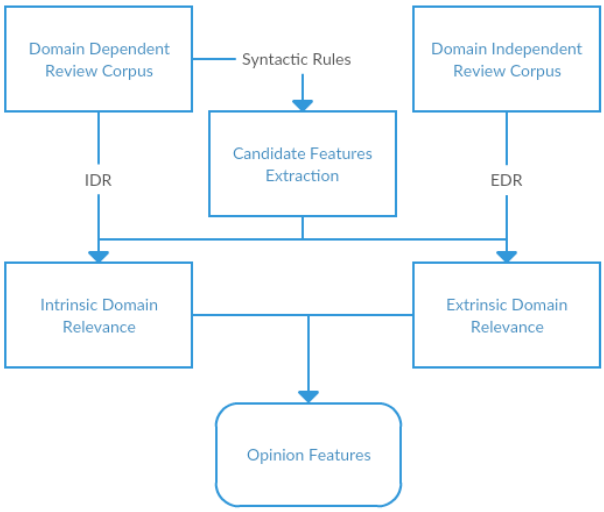

Figure 1. System Architecture

Fig 1 explains the flow of the proposed method. With the help of the manually stated syntactic rules, we first mine the list of the candidate features from the given review corpus. Later, the IDR and the EDR is computed. The IDR gives the statistical association of the candidate feature to the given domain corpus whereas, the EDR replicates the statistical relevance of the candidate feature to the domain-independent corpus. Valid opinion features are extracted as the candidate features with IDR gains (high score of IDR) greater than the predefined intrinsic relevance threshold and EDR scores (low EDR score) less than the extrinsic relevance threshold.

A. Candidate feature extraction

The opinion features are made up of noun phrases or nouns which generally emerge as the subject or the object of the review statement.

The subject opinion feature has a syntactic relationship of type subject verb (SBV) with the statement predicate and the object opinion feature has a dependence relationship of verb-object (VOB) on the predicate. This is in the cases of dependence grammar. In further addition to it, it also has a dependence relationship of the preposition object (POB) on the prepositional word in the statement. Therefore we present the three syntactic rules based on the above dependence relations i.e. SBV, VOB, and POB. They are as follows:

Table 1 Syntactic Rules

Rules	Interpretation
NN+SBV CF	Identify NN as CF, If NN has a SBV dependency relation.
NN+VOB CF	Identify NN as CF, If NN has a VOB dependency relation.
NN+POB CF	Identify NN as CF, If NN has a POB dependency relation.

.

The procedural mechanism of the candidate feature extraction is as the following:

Firstly, in a given review corpus, we need to recognize or identify the syntactic organization of each statement. Dependence parsing (DP) is employed.

The proposed candidate feature extraction method is language reliant. The three syntactic rules (from the table given above) are later applied to the recognized dependence structures. When a rule is fired or implemented, the corresponding nouns or noun phrase are mined as a set of candidate features.

Domain relevance: The way in which a term is related to a particular domain is based on two types of the statistics which are dispersion and deviation. This is described by the domain relevance. Dispersion gives us the frequency or counts the number of times a particular term is referred across the documents by computing the distributional importance of that term across many different documents in the entire domain. This is generally known as the horizontal significance.

B. Opinion feature Identification

Intrinsic and extrinsic domain relevance: The domain relevance of particular opinion feature calculated on the given domain specific review corpus is known an Intrinsic Domain Relevance (IDR). The domain relevance of particular opinion feature calculated on the given domain-independent review corpus is known an Extrinsic Domain Relevance (EDR). The IDR gives the statistical association of the candidate feature to the given domain corpus and the EDR illustrates the statistical association of the candidate feature to the domain-independent corpus, as the candidate features are related to either one corpus or the other but never related to both the corpora at the same time. Therefore in such a case, the EDR also gives the irrelevance of the candidate feature to the given review corpus. Also, there are some common terms that are used everywhere and also used in the review corpus as candidate features.

C. Proposed Architecture

Number citations consecutively in square brackets [1]. The sentence punctuation follows the brackets [2]. Multiple references [2], [3] are each numbered with separate brackets [1]–[3]. When citing a section in a book, please give the relevant page numbers [2]. In sentences, refer simply to the reference number, as in [3]. Do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] shows ... .” Number footnotes separately in superscripts (Insert | Footnote).[1] Place the actual footnote at the bottom of the column in which it is cited; do not put footnotes in the reference list (endnotes). Use letters for table footnotes (see Table I).

D. POS patterns and candidate generation

In this model, we extract the combination of nouns and noun phrases and adjectives from the review sentences since according to the observations the aspects are the nouns. We

[1]It is recommended that footnotes be avoided (except for the unnumbered footnote with the receipt date on the first page). Instead, try to integrate the footnote information into the text.

introduce the heuristics combinations in the table which are the experimentally extracted POS patterns. The first row gives the heuristic combinations selects the candidate aspects from the noun phrase patterns like "NN", "NNS", "NN NN" , etc. The second row uses patterns like "JJ NN", "JJ NNS", "JJ NN NN", etc. The third row selects candidates based on the pattern ‘‘DT JJ", etc. The last row uses heuristic patterns like ‘‘DT VBG’’, ‘‘VBG NN’’ and ‘‘NN VBG NN’’.

Table 2 Heuristic Rules

Heuristic combination POS patterns for candidate generation	
Description	**Pattern**
Nouns	Unigram to four-gram of NN and NNS
Nouns and adjectives	Bigram to four-gram of JJ, NN and NNS
Determiners and adjectives	Bigram of DT and JJ
Nouns and verb gerunds	Bigram to trigram of DT, NN, NNS and VBG

Once we find the candidate features, we move to the next level of aspect identification.

We therefore start with the heuristics and the experimentally extracted rules. The two rules in the aspect detection model are:

Rule #1: The aspects for which there are no opinion words within the sentence are removed.

Rule #2: The aspects that contain stop words are removed.

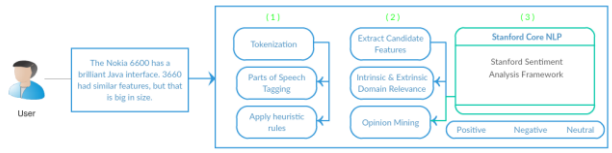

Figure 2 Proposed Architecture

The purpose of extracting the aspects is to construct a sentimental analysis system. The aspect is not very valuable if there are no opinion words that appear with it. Hence, we employ Rule #1 mentioned above. Opinion words are words that help us understand if the people have given a positive or negative opinion. In this study we check adjective phrases for opinion words in Rule #1 as most of the opinion words are adjectives in the sentence and therefore we extract adjective phrases from review sentences to construct a polarity lexicon.

We will demonstrate the working to the review sentences ‘‘Signal strength will affect the battery life.’’ and ‘‘Battery life is very good, I use it every day and I have to charge it every 4 or 5 days or so.’’ Both of these sentences talk about the aspect ‘‘battery life’’. The first sentence is not an opinionated sentence and states a fact about battery life, whereas, the second sentence expresses an opinion or sentiment about ‘‘battery life’’. On applying Rule #1 we ignore sentences without opinions like the first sentence for candidate aspect extraction.

By using the Rule #2 candidate aspects that contain stop words are removed as they are considered not to contribute any semantic weight. For instance, pattern ‘‘JJ NN’’ from Table 2 can extract some incorrect aspect candidates like ‘‘other cellphone’’. According to Rule #2 this ‘‘other cellphone’’ should be removed for the set of candidate aspects. These heuristic rules turned out to improve the performance of aspect detection model. Through these performed experiments.

E. Identification of implicit aspects

In this section we focus on identifying the implicit aspects. We therefore consider that an implicit aspect should satisfy the following criteria:

1. In the review sentence, the related aspect word does not occur explicitly.
2. In the review sentence, the aspect can be discovered by the opinion words or its surrounding words.

Table 2 Penn Diagram

POS Tag	Expression
CD	CARDINAL
FW	FOREIGNWORD
UH	INTERJECTION
LS	LISTMARKER
NN, NNS, NNP, NNPS	NOUN
PRP	INTERJECTION
MD	MODAL
PB, RBR, RBS	ADVERBS

Table 3 Review Aspects

Examples of implicit aspects in review sentences for Nokia 6610.	
Review sentence	**Implicit aspects**
It is small	Size
I like my phone to be small so I can fit it in my pockets	Size
This is a very light phone	Weight

In the above given table there are three examples of implicit aspects in review sentences for Nokia 6610 from www.amazon.com. For identifying implicit aspects in the reviews, we propose a graph-based approach. We draw a graph for aspects and opinion words by utilizing a list of predefined aspects, and a polarity lexicon. From the polarity lexicon, the graph uses an opinion word as a node. It maps this node to the set of the aspects nodes. In the graph, if a pair of nodes co-occurs together in a review sentence, we set an edge to it and we assign initial weight w to the edge as the number of their co-occurrence.

We use extracted aspects and opinion words from the previous sections in this proposed approach. Since using only the co-occurrence of aspect and opinion word for identifying implicit aspects are not enough we therefore define a function to measure the aspect association and opinion word as:

F. Key Index Parameters for Result Classification

In information retrieval with binary classification instances are documents and the task is to return a set of relevant documents given a search term.

The positive predictive value known as precision is the fraction of retrieved instances that are relevant. Precision is the ratio of the correct positive observations. The formula for precision Is (True Positives) / (True Positives+ False Positives).

Recall also known as sensitivity is the fraction of the relevant instances that are retrieved. It is the proportion of all real positive observations that are correct. Recall is the ability of the classifier to give all the positive samples. The formula for Recall is (True Positives) / (True Positives + False Negatives). Therefore, Precision & Recall values are based upon understanding and measuring relevance.

A high recall means that an algorithm has yielded the most relevant results whereas high accuracy means that an algorithm returns significantly more relevant than irrelevant results.

The most important category measurements for binary categories are:

Precision	Recall	F Measure
P= TP / (TP+FP)	R= TP/(TP + FN)	2*((precision*recall) /precision)

Where, TP= True Positive
FP= False Positive
FN= False Negative

III. CONCLUSION

In this research we study opinion mining and study sentiment analysis for the review documents. It is often expensive and time consuming to construct labeled data for training purposes and it is desirable to develop a model or algorithm that can do without labeled data while dealing with mining reviews online. In this paper, we therefore proposed language-independent and an unsupervised domain model for detecting the implicit and the explicit aspects from the review documents.

There are three major bottlenecks: domain dependency, the need for labeled data, and implicit aspects. The proposed model is able to deal with it. We also proposed a number of novel techniques for mining aspects. We used the influence of an opinion word on detecting an explicit aspect and the inter-relation information between words in a review. Further, we described an approach which uses a co-occurrence metric to calculate the association between the explicit aspect and the opinion words to identify implicit aspects. The conclusion can be drawn that the model can be used in practical settings, particularly where high precision is required.

ACKNOWLEDGMENT

I would like to thank my mother Mrs Varsha Suchet Vakil and my father Mr. Suchet R. Vakil for supporting me and providing me with neccessary financial and emotional aids. I am also thankful to Mr. Nikhil Brahmbhatt for providing the necessary facilities for the preparation of the paper.

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
2. B. Liu, “Sentiment Analysis and Opinion Mining,” Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167, May 2012.
3. Hatzivassiloglou and j.m. Wiebe, “effects of adjective orientation and gradability on sentence subjectivity,” proc. 18th conf. Computational linguistics, pp. 299-305, 2000.
4. 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.
5. R. Mcdonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar, “Structured Models for Fine-to-Coarse Sentiment Analysis,” Proc. 45th Ann. Meeting of the Assoc. of Computational Linguistics, pp. 432- 439, 2007.
6. 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.
7. Yessenalina and C. Cardie, “Compositional Matrix-Space Models for Sentiment Analysis,” Proc. Conf. Empirical Methods in Natural Language Processing, pp. 172-182, 2011.
8. A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts, “Learning Word Vectors for Sentiment Analysis,” Proc. 49th Ann. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies, pp. 142- 150,2011.
9. S.-M. Kim and E. Hovy, “Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text,” Proc. ACL/COLING Workshop Sentiment and Subjectivity in Text, 2006.
10. T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis,” Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347-354, 2005.
11. Quillian, M. Ross (1966) Semantic Memory, PhD dissertation, Carnegie Institute of Technology (now CMU). Abridged version in Minsky (1968) pp. 227-270.
12. Schank, Roger C., & Larry G. Tesler (1969) A conceptual parser for natural language, Proc. IJCAI-69, 569-578.
13. Woods, William A. (1975) What’s in a link: foundations for semantic networks, in D. G. Bobrow & A. Collins, eds. (1975) Representation and Understanding, New York: Academic Press, pp. 35-82.
14. Chomsky, N., "A transformational approach to syntax", Proceedings of the 1958 University of Texas Symposium on Syntax

AUTHORS PROFILE

Janhavi Vakil is pursuing her degree from MPSTME, NMIS Shirpur, her hobbies includes reading technical blogs, her research interest is Natural Language Processing.