International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

Feature Based Sentiment Analysis for Online Reviews in Car Domain

Amruta Sankhe[Á*] and Prachi Gharpure[Á]

[Á]Computer Department, Mumbai University, SPIT,India

Accepted 17 March 2014, Available online 01 April 2014, **Vol.4, No.2 (Feb 2014)**

Abstract

The rapid growth of the Web in the last decade makes it the largest publicly accessible data source in the world. There are several websites which allows user to post reviews on particular product or service. Customers also want to know the opinions of existing users before they use a service or purchase a product; even businesses want to find public or consumer opinions about their products and services. Hence there is need to do automatic feature based opinion mining for buyers to take smart decision. However, since the polarity of features is varied according to domain the task of feature based summary is challenging itself. This paper presents an approach for mining online user reviews to generate feature-based sentiment analysis that can guide a user in making an online purchase. In this work, feature based sentiment analysis on car domain is presented. In which features are extracted using domain ontology, then by using the feature description table the features are classified into positive and negative polarity. Since the polarity of each feature is varied according to the domain. The outcome of the system is a set of reviews organized by their degree of positivity and negativity based on each feature. This system helps to reduce the manual effort of evaluating reviews according to features in which user is interested. The polarity obtained for each feature from our approach is with good average accuracy.

Keywords: *opinion mining, sentiment analysis, feature-based opinion mining, summarization, semantic web*.

1. Introduction

Due to the rapid expansion of the internet, business through e-commerce has become popular. Many products are being sold on internet and the merchants selling the products ask their customers to write reviews about the products that they have purchased. This is the reason behind the abnormal increment of the number of reviews on websites. The problem occurs when for a particular popular product, the number of customer reviews reaches hundreds and sometimes in thousands. This increment makes it very difficult for potential customer as well as for the manufacturer of the product to sort the useful reviews from the one that do not contain useful information and to make a decision whether to purchase a product or not. In addition to that few reviews are too long and contain few sentences that express opinions about the product.

For example a consumer is interested in buying a particular digital camera and wants to get the knowledge of different features, strengths and weaknesses of that camera. The consumer would also like to compare the features of this camera with other brands of cameras. This process of comparing requires manual searching of related websites which would need a lot of time. Information gained by visiting few websites would provide incomplete and less information (Deshpande &Sarkar, 2010).

Opinions are central to almost all human activities and are representation of our behaviors. For this reason, when we want to buy any product or service we often check out the opinions of others. This is true for both for individuals and also for organizations. Opinions and its related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining.

One can not only communicate with people anywhere in the world but also express one's views and opinions on anything using Internet forums, blogs, review sites and social network sites. People make fewer and fewer trips to libraries, but more and more searches on the Web. Retrieving information simply means finding a set of documents that is relevant to the user query. But that is not sufficient for today's web, we require smart search for web by analyzing and mining people's opinions which indicate positive or negative sentiments. Potential customers also want to know the opinions of existing users before they use a service or purchase a product.

Sentiment analysis is a type of natural language processing for tracking the mood or opinion of the public about a particular product or topic or service. There is various ways to do the sentiment classification by machine learning or by semantic approach. But today's need is to do feature based sentiment analysis. Since polarity of each feature get changed according to the domain. To take smart buying decision sentence and document level classification is not sufficient. Feature based analysis helps

*Corresponding author: **Amruta Sankhe**

buyers to do perfect choice and business to grow their business in competitive world. The main difficulty in analyzing these reviews is that they are in the form of natural language. The processing of natural language is difficult; analyzing online unstructured textual reviews is even more difficult. However today need is feature based analysis which increases the difficulty level even more. The field of Opinion Mining (OM) is recent and as a result there are still a lot of challenges to be met. The mining of forums and online discussions is a challenge on its own. It is bit difficult for average human reader to identify relevant web sites and extract features and to do polarity summarization. Therefore automated sentiment analysis systems are thus needed. (Bing Liu, 2012)

The rest of the paper is arranged as follows. Section 2 describes work in the area of semantic web and feature based sentiment analysis, Section 3 describes the proposed feature based sentiment analysis, and Section 4 evaluates the accuracy of the approach based on obtained results. At end, we conclude and discuss the scope for future work in this field.

2. Related Work

In order to do the feature based sentiment analysis feature extraction is the main step the techniques to extract features are machine learning and semantic approach.
(Songbo Tan, 2008) presents an empirical study of sentiment categorization on Chinese documents. He investigated five learning methods (centroid classifier, K-nearest neighbor, winnow classifier, Naive Bayes and SVM) on a Chinese sentiment corpus. From the results he concludes that, SVM exhibits the best performance for sentiment classification. Whereas in this paper we have presented semantic approach to extract domain features. Semantic technologies have been around for a while, offering a wide range of benefits in the knowledge management field. Ontologies provide a formal, structured knowledge representation, and have the advantage of being reusable and shareable They have revolutionized the way that systems integrate and share data, enabling computational agents to reason about information and infer new knowledge . The accuracy results of opinion mining and sentiment polarity analysis can be improved with the addition of semantic techniques, as shown in (Namrata Godbole, 2007). In that work, some semantic lexicons are created in order to identify sentiment words in blog and news corpora. Then, a polarity value is attached to each word in the lexicon and such polarity is revised when a modifier appears in the text (Juana Maria Ruiz Martinez, 2012). Ontologies are thus the key for the success of the Semantic Web vision. The use of ontologies can overcome the limitations of traditional natural language processing methods and they are also relevant in the scope of the mechanisms related, for instance, with Information Retrieval, Semantic Search, and Question Answering. (JM Ruiz-Martınez, 2011; García-Sánchez, 2009; Valencia-García, 2011)

Once the feature set is finalized sentiment analysis can be performed on users review. After feature extraction next step is to build sentiment lexicon. There is various ways to build sentiment lexicon for opinion mining. Sentiment lexicons are generated using Dictionary based and corpus based methods.

Using a dictionary to compile sentiment words is an obvious approach because most dictionaries (e.g., WordNet list synonyms and antonyms for each word. Thus, a simple technique in this approach is to use a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. Specifically, this method works as follows: A small set of sentiment words (seeds) with known positive or negative orientations is first collected manually, which is very easy. The algorithm then grows this set by searching in the WordNet or another online dictionary for their synonyms and antonyms. The newly found words are added to the seed list. The next iteration begins.the iterative process ends when no more new words can be found. This approach was used in (Hu and Liu, 2004). After the process completes, a manual inspection step was used to clean up the list.

A similar method was also used by (Valitutti et al., 2004; Kim and Hovy, 2004) tried to clean up the resulting words (to remove errors) and to assign a sentiment strength to each word using a probabilistic method. (Mohammad et al., 2009) additionally exploited many antonym generating affix patterns like *X* and dis*X* (e.g., honest-dishonest) to increase the coverage. A more sophisticated approach was proposed in (Kamps et al., 2004) which used a WordNet distance-based method to determine the sentiment orientation of a given adjective.

In (Blair-Goldensohn et al., 2008) a different bootstrapping method was proposed, which used a positive seed set, a negative seed set, and also a neutral seed set. The approach works based on a directed, weighted semantic graph where neighboring nodes are synonyms or antonyms of words in WordNet and are not part of the seed neutral set. The neutral set is used to stop the propagation of sentiments through neutral words.
In (Rao and Ravichandran,2009) three graph-based semi-supervised learning methods were tried to separate positive and negative words given a positive seed set, a negative seed set, and a synonym graph extracted from the WordNet.

More recently, researchers have used the opinion mining tool SentiWordNet to determine orientation of opinions in sentiment. The main disadvantage of this approach is that the sentiment orientations of words collected this way are general or domain and context independent. In other words, it is hard to use the dictionary-based approach to find domain or context dependent orientations of sentiment words. Since the polarity of each feature is varied according to the domain.

Earlier attempts at determining the semantic orientation of adjectives relied upon the use of supervised learning which involved frequency analysis and clustering on a large manually tagged corpus (V. Hatzivassiloglou and K. Mckeown, 1998). In (P. D. Turney, 2002) authors used the PMI statistic (point wise mutual information) that predicted the orientation of an adjective based on its co-occurrence with words “excellent” or “poor.”

But the adjective descriptors used for different products differ widely, it is not possible to achieve uniform accu-

-racy using this techniques.

Lexical resources such as SentiWordNet contain opinion bias scores based on individual terms, and when building a model based on this type of information there are certain challenges stemming from the nature of natural languages. Word sense disambiguation becomes relevant, since terms with potentially multiple meanings may carry different opinion bias depending on context and their use within a sentence.

Domain-specific terms are also an issue, since they may indicate a different bias than that of their more commonly seen uses. Even every domain has its own set of features, and polarity of each feature is varies according to the domain. The above issues naturally impose limitations to the effectiveness of sentiment classification using SentiWordNet.

In various research work SentiWordNet, is used to do sentiment classification of reviews, but there is need to do the feature based sentiment classification. If these user reviews classified appropriately and summarized based on features can play an instrumental role in influencing a buyers' decision.

3. Proposed feature based Summarizer and Classifier:

In this section we will explain the system design of the feature based summarizer and polarity classification implemented by us. We have taken car reviews from various car websites which post car reviews. As shown in Figure 1, our approach has following phases which explained next. These phases are
(1) Review collection and preprocessing phase, (2) feature extraction phase using domain ontology, and (3) Feature based polarity classification phase.

System architecture:

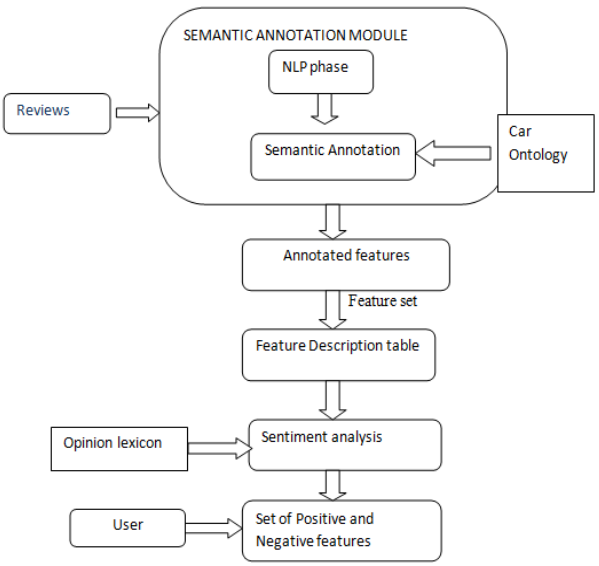

Fig 1 System Architecture

Our feature based opinion mining system needs three basic components: a lexical resource L of opinion expressions, a lexical ontology O where each concept and each property is associated to a set of labels that correspond to their linguistic realizations and a review R.

3.1Reviews collections: It collects the various car related reviews which contains positive and negative opinions and stores it in database.

3.2 NLP Module: **Natural language processing** (**NLP**) is a field of artificial intelligence, which deals with the interactions between computers (system) and human (natural) languages. It deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages. The main objective of this module is to obtain the morphologic and syntactic structure of each sentence. For this, a set of NLP tools including a sentence detection component, tokenizer, POS taggers, end of sentences and syntactic parsers has been done using the GATE framework.

In order to do the preprocessing phase following process has done for reviews.

1) Tokenization process: splits the text into very simple tokens such as numbers, punctuation and words of different types.
2) Sentence Splitting process: it deals with segmenting the review into sentences. This module helps to distinguish sentence-marking full stops from other kinds.
3) Speech tagging process: produces a part-of-speech tag as an annotation on each word or symbol.

3.3 Feature extraction using domain ontology

Ontology has been defined as the specialization of the conceptualization by (Gruber, 1993).The main aim of ontology is to provide knowledge about specific domains that are understandable by both the computers and developers. It also helps to interpret a text review at a finger granularity with shared meanings and provides a sound semantic ground of machine understandable description of digital content. Ontology improves the process of information retrieval and reasoning thus results in making data interoperable between different applications (Zhou and Chaovalit, 2007)

After the pre-processing step, we use domain ontology for feature identification and extraction. For each expression, the system tries to find out whether the expression under question is an individual of any of the classes of the car ontology. Then, system retrieves all the annotation respective to expression. Each class in the ontology is defined by means of a set of relations and data type properties. Then, when an annotated term is mapped onto an ontological individual, its data type and relationships constitute the potential information which is possible to obtain for that individual. The following are tasks done by ontology:

Structure features: ontology is main part of semantic web; it helps to define meaning, concepts, relationships and entities that describe a domain with unlimited number of terms. These sets of terms are very helpful for extracting explicit and implicit features. For example, in

Table 1 Feature Description Table

Feature	Positive polarity	Negative polarity
comfort	Complacent, enjoyable, satisfying, useful, cozy,	Dissatisfied, unpleasant, unsuited, unfriendly, discontented, hopeless, pitiable, disagreeable, troubled,
price	Cheap, reasonable, Valuable, inexpensive, worthy, affordable, good, low	Expensive, pricey, extravagant, invaluable, worthless, exorbitant, costly, high-priced, high.
speed	Fast, quick, swift, speedy, nimble, brisk, high-speed, high, good, rapid	Slow, bad,
quality	Awesome, cool, good, satisfactory, superfine, bang-up, fabulous, fantastic,	Terrific,bad, inferior, low-grade, substandard, unsatisfactory, atrocious, awful, execrable, pathetic, poor,
performance	High, good	Poor,terrible,pathetic,bad,worst
engines	Good, awesome, cool, satisfactory, superfine, bang-up, fabulous, quiet	Bad,worst,terrific,loe-grade,unsatisfactory,pathetic,poor,terrible

Table 2 Overall accuracy

Features	Using FD table		Using SENTIWORDNET		Average accuracy of polarity(FD)	Average accuracy of polarity(SENTIWORDNET)
	Total	Total negative	Total positive	Total negative		
cost	17	08	09	16	0.6	0.3
speed	20	9	15	7	0.8	0.6
safety	16	09	14	12	0.6	0.5
comfort	18	07	12	18	0.7	0.4
mileage	15	12	11	13	0.6	0.4
Lighting system	21	6	16	10	0.8	0.6
Internet service	12	11	11	10	0.6	0.4
seat	19	8	18	9	0.7	0.7
Total					0.6	0.4

the following restaurant review: cold and not tasty the negative opinion not tasty is ambiguous since it is not associated to any lexicalized feature.

However, if the term cold is stored in the ontology as a lexical realization of the concept quality of the cuisine, the opinion not tasty can be easily associated to the feature cuisine of the restaurant.

Extract features: ontologies provide structure for these features through their concept hierarchy but also their ability to define many relations linking these concepts. This is also a valuable resource for structuring the knowledge obtained during feature extraction task. This way it helps to extract domain features.

3.4 Feature Description Table: now once we got the feature set from previous method next is to build feature description table. We have created the feature description table for each feature related to car domain with positive and negative polarity associated with each feature. Since the orientation of the adjectives describing the elements of the extracted feature set is useful in performing the sentiment analysis. System will determine the polarity of opinion words involves using an initial list of adjectives with known orientations which is subsequently expanded by looking up its synonyms and antonyms using the lexical resource WordNet. The feature description table as shown below:

3.5 Sentiment analysis

The main objective of this module is to classify the set of features obtained in the previous module according to their polarity: positive or negative. For any retrieved features which have been annotated, the sentiment orientation or sentiment polarity value is computed.

However, for positive value the system adds 1positive, and substrates 1 for negative opinion.

Next, if the semantic polarity value of review is less than 0, the news is labeled as negative. In contrast, if the value is higher than 0, the review is labeled as positive.

4. Empirical Evaluation and Results

We collected over 25 online reviews for car domain from several popular review websites using a web crawler. We applied Preprocessing steps like sentence boundary detection, spell-error correction on the review dataset as explained earlier in the preprocessing phase. We obtained the feature set as explained in the feature extraction phase, generated the feature description table and determined the polarity according to the features of domain. Since polarities of features are varied according to domain our approach helps to get results with high accuracy.

Since our aim is to do the feature based sentiment analysis we have done the polarity classification based on features as follows:

The results we obtained show that polarity classification using feature description table has high accuracy then polarity by Sentiwordnet.

Domain: CAR

Graph represents the polarity classification using our system:

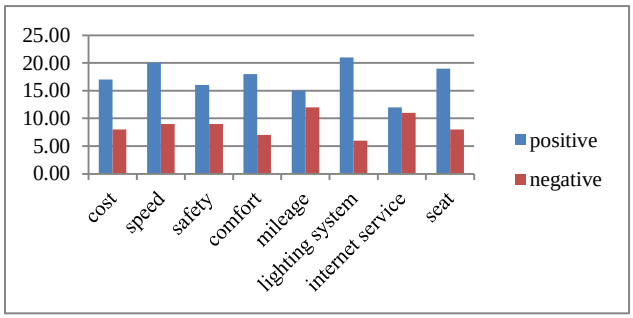

Fig 2 sentiment classification of features

5. Conclusion and future work

Opinion mining has become a fascinating research area due to the availability of a huge volume of user-generated content, e.g., reviewing websites, forums, and blogs. Aspect-based opinion mining, which aims to extract item aspects and their corresponding ratings from online reviews, is a relatively new sub-area that attracted a great deal of attention recently. The extracted aspects and estimated ratings not only ease the process of decision making for customers but also can be utilized in other opinion mining systems.

Our method is promising because the use of the ontology allows improving the feature extraction and the association between an opinion expression and object features. On the one hand, the ontology is useful thanks to its concept list which brings a lot of semantic data in the system. Using concept labels the ontology allows to recognize terms which refer to the same concepts and brings some hierarchy between these concepts.

The proposed methodology is supported by natural language processing methods to annotate car reviews in accordance with car ontology, then by using the feature description table the features are classified into positive and negative polarity. Since the polarity of each feature is varied according to the domain. The outcome of the system is a set of reviews organized by their degree of positivity and negativity based on each feature. The proposed system helps to give uniform accuracy. This system helps to reduce the manual effort of evaluating reviews according to features in which user is interested. In the future, we want to extend feature based opinion mining on various domains. We would also like to extend our work to find out the strength of various adjectives which help to increase the accuracy of sentiment analysis.

References

Bing Liu. (2012), Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers.

Juana Maria Ruiz Martinez (2012),Semantic Based Sentiment analysis in financial news, *ceur-ws.org/Vol-862/*

Juana Maria (2011), Ontology Population: An Application For The E-Tourism Domain, International Journal of Innovative Computing, Information and Control Volume 7 , Number 11

García-Sánchez, F., Valencia-García, R., Martínez-Béjar, R., Fernández-Breis, J.T(2009), An ontology, intelligent agent-based framework for the provision of semantic web services, Expert Systems with Applications 36(2) Part 2, pp.3167–3187

Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J.T.(2011), OWLPath: an OWL ontology-guided query editor, IEEE Transactions on Systems, Man,Cybernetics: Part A, vol 41(1), pp. 121 – 136

Bar-Haim, Roy, Elad Dinur, Ronen Feldman, Moshe Fresko, and Guy Goldstein. (2011), Identifying and Following Expert Investors in Stock Microblogs, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2011).

V. Hatzivassiloglou and K. Mckeown, (1998),Predicting the semantic orientation of adjectives, in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL '98), pp. 174–181.

P. D. Turney (2002), Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424.