ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Feature-Based Opinion Mining: A Review

Juveria Fatima[1], Dr. Deepak Arora[2]

P.G. Student, Department of Computer Science & Engineering,Amity University, Lucknow, Uttar Pradesh, India[1]

Professor, Department of Computer & Engineering, Amity University, Lucknow, Uttar Pradesh, India[2]

ABSTRACT: Internet is accessible to almost everyone these days and therefore people spend most of their time in surfing the internet whether it is for education, entertainment, online shopping, etc. Not only do they use the available information but they can also give feedbacks or suggestions. As a result there are various reviews available online which can benefit both, the customers and the manufacturers. Customers can use it for making smart purchase decisions and the manufacturers can take its help for locating the areas of their products which require improvements. These reviews are basically the opinions of various individuals. Opinion mining is a discipline that requires natural language processing, information retrieval and data mining. It aims at classifying the opinions available on the web as positive, negative or neutral. It is done at three different levels- sentence, document and feature. Feature based opinion mining aims at summarizing the opinions for each feature of a product and then classifying it as positive, negative or neutral. It helps in determining which features of the products do customers like or dislike. This paper discusses various methods proposed for the feature based opinion mining.

KEYWORDS:Feature based opinion mining, opinion mining, polarity detection, sentiment classification.

I. INTRODUCTION

The rapid way in which Web 2.0 has emerged has led the explosion of social media contents and e-commerce sites. It has become a platform for the people to communicate, shop, give feedbacks or suggestions, etc. People do not only read the information available on the internet but they can contribute to it too by providing suggestions/feedbacks. As a result, there is a lot of data available on the web. A person who is willing to buy anything and wants to take help from the reviews available online might get confused because of the variety of opinions by customers on a varied range of platforms like, blogs, social sites, e-commerce sites, etc. Therefore, to deal with this problem there is a discipline called opinion mining which aims at analyzing opinions. Opinions are basically attitude, emotions or sentiments of an individual on some entity. These are available in the form of reviews on the web. Mining the relevant opinions can help the customers as well as the organizations. Customers can benefit themselves through opinion mining by making smart purchase decisions and they will not have to be dissonant about their decision later. On the other hand, organizations can derive help from opinion mining by making better marketing plans, better production, in locating the defects of their products and thereby improving them too, etc. It is a very challenging task as there is no particular structure followed in reviews. Automated extraction of reviews from those available online require natural language processing, machine learning, information retrieval, data mining etc.

Opinion mining is done at different levels- sentence level, document level and feature level. In the sentence level opinion mining, the polarity of each sentence is identified and classified as positive, negative or neutral. In the document level opinion mining, the polarity of whole document is identified and it is assumed that the document holds the opinion of a single opinion holder. In the feature based opinion mining, the opinions corresponding to every feature is summarized after the features have been identified. It determines which feature of the product is liked or disliked by the customers. There are various products available on the e-commerce sites and each product has various features on which various customers express their reviews. It becomes difficult for the customers to analyze the reviews available online. It is a very tedious task. Feature based opinion mining aims at solving this problem by producing a summary of opinions. However, the customers do not generally use any fixed format or structure for writing reviews. Identifying features from the reviews is a challenging task as some reviews may explicitly contain features and the others may have

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

it implicitly. For example, *"The size of this phone is very big"*. This review has explicitly stated the feature '*size'*. Consider another review - "*It is light and can be carried anywhere easily"*. This review has stated the feature '*size'* implicitly which is difficult to be identified. Not much work has been done to identify the implicit features. For the process of mining reviews from various platforms like social blogs, e-commerce sites, etc, the reviews are firstly downloaded and then the pre-processing of the downloaded reviews is done. It includes removing HTML tags and stop words. The downloaded reviews are pre-processed until the unnecessary information has not been removed. After this step, features and opinion words are identified, for which many approaches have been proposed. This paper provides an insight of various methods proposed for feature based opinion mining. The rest of the paper is organized as follows: Section (II) discusses about the data sources, Section (III) discusses the basic tasks involved in mining the features and opinion words, Section (IV) discusses the various approaches used for the feature-based opinion mining, Section (V) discusses about various performance measures, Section (VI) discusses about the challenges involved in this task and Section (VII) concludes the paper.

II. DATA SOURCES

There are various sources of data available online from which the reviews can be collected. Since, the use of web has evolved tremendously, every user finds it tempting to look for the reviews online before making any purchase decisions. These reviews are available in the blogs, e-commerce sites as well as other social networking sites. These sources are described in detail below:

A. Blogs

Now-a-days, the number of blog pages is increasing rapidly. Blog pages contain the opinions of a single individual. The writer might wish to write about any current trending issue, any product that he had used and liked/disliked it, etc. Few blogs gain popularity and get constant readers. Readers can therefore benefit themselves by reading the corresponding reviews.

B. E-Commerce sites

Almost every e-commerce site has the option for the users to comment about any product i.e. post reviews about the products. Before buying any product, buyer can read the reviews about the products and make better decisions after he had compared the reviews are various products. Example of such e-commerce site is Amazon.com, Flipkart.com, etc.

C. Social Networking sites

A lot of users express their opinions, views or perceptions about anything on the social networking sites. Twitter and Facebook are the most popular social networking sites available today. People use it as a platform to express their opinions. These are considered to be a potential source for opinion mining.

III. BASIC TASKS INVOLVED IN FEATURE-BASED OPINION MINING

In order to extract features and their corresponding opinions from the web, following tasks are required to be performed:

A. Crawl reviews

Firstly, the reviews from the required data source are downloaded and stored in the database. For this purpose web crawlers can be used. Web crawlers visit sites and are capable of downloading reviews from it.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

B. Pre-Processing

After the reviews have been downloaded, pre-processing of the stored reviews is done. It involves removing the unnecessary information from the downloaded content. This step helps in the removal of HTML tags, URLs and the other tasks like stemming, removal of stop words, tokenization and normalization.

C. POS tagging

Once the reviews have been pre-processed, POS tagger is used to tag the words of each sentence with their corresponding part of speech type. For example, the review "*This phone has an awesome display"* will be tagged as:

This/DT phone/NN has/VBZ an/DT awesome/JJ display/NN

This tagging can be done with the POS taggers that are readily available on the web such as Stanford POS tagger, etc. After the sentence has been tagged, further processing is done.

D. Feature extraction

Features of any product are generally nouns. So after the sentence has been tagged, the nouns can be extracted easily. These nouns are usually the features of a product. Like in the above example, words that are tagged as nouns (phone and display) are the required features.

E. Opinion word extraction

Opinion words are generally the adjectives that describe the nouns. Once the sentence has been tagged, the adjective words can be extracted which are usually the potential opinion words. Like in the above example, word that is tagged as adjective (awesome) is the required opinion word that describes the display of a phone.

F. Determining the sentiment orientation of opinion words

After the opinion words have been identified, the polarity of the opinion words is determined. For determining the polarity of the opinion words, various approaches have been proposed. One can use the standard dictionary available online like SentiWordNet which is capable of defining the polarity of the opinion words.

G. Determining the polarity of the sentence

The sentiment orientation of the sentence is determined by averaging the polarity values of the opinion words in a sentence. Usually, the polarity of the sentence depends upon the majority of the particular polarity in a sentence. If a sentence has maximum positive opinion words then the polarity of the whole sentence will be positive, otherwise negative.

H. Summary generation

After every opinion word and the features have been identified, feature wise summary can be generated in which every review of the corresponding feature is classified as positive, negative or neutral.

IV. VARIOUS APPROACHES FOR FEATURE-BASED OPINION MINING

The following approaches have been proposed for mining the opinions from the web:

A. Dictionary based approach

This approach is based on a pre-built dictionary. This dictionary defines semantic orientation of words. An example of such dictionary is 'SentiWordNet' which is the standard dictionary available today. In order to determine the semantic orientation of words i.e., if a word indicates positivity or negativity, this dictionary can be used. The

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

approaches that use this approach first identifies the semantic orientation of each opinion words in a sentence through this dictionary and then the semantic orientation of the whole sentence is identified by averaging the orientation values of each opinion words present in a sentence. The methods that use dictionary approach use the rule-base algorithm for detecting the overall polarity of reviews. Zhu et al. 2011 [1] used Chinese sentiment lexicon Hownet for classifying the reviews of a restaurant in Chinese language. For identification of features, it applied bootstrapping method. The polarity of the review was calculated by summing the sentiment orientation values of the opinion words occurring in a sentence. The overall semantic orientation of a review is greater than 0 for a positive review, less than 0 for a negative review and equal to 0 for a neutral review.

B. Machine Learning

Machine learning is the discipline that aims at obtaining the work from machines that is comparable with that of humans. In other words, it aims at building machines that can think and act like humans. Its first step is preparing a training dataset which may or may not be labeled with the corresponding sentiments and then presenting the reviews as the vector of features. In the next step, a classifier is trained so that it can differentiate among the reviews based on their sentiment labels. After it has been trained, it is expected to classify the reviews correctly that are presented to it after training. This process may be supervised, in which a labeled training dataset is available or unsupervised, in which it is not available. Supervised technique like Naïve-Bayes has been used by David D. Lewis. [11] andPedro Domingos and Michael J. Pazzani [12] and it showed great results for identifying the features and classifying the reviews as positive, negative or neutral. Bing, Liu et al. [13] used the unsupervised approach. It used double propagation method. It is based on the assumption that the features and the opinions are co-related. Firstly, the features are extracted and then through it opinions are extracted. The extracted opinions are then used again to find out the other features. This process continues until there are no opinion words or features left to be extracted. It worked well for the corpora with medium size but did not produce satisfactory results for small and large sized corpora.

C. Statistical Approach

It is based on the idea that the opinion words that occur together generally have the same polarity. It determines the polarity of a any new word by calculating its frequency with other opinion words. This new word will then probably possess the same orientation as the word with which it occurs most frequently. For this approach, it maintains large sized corpora. The corpora are kept larger in size in order to deal with the problem of unavailability of some words.

Point wise mutual information method was proposed by Peter Turney [14], [15]. It calculated the PMI by evaluating the log of the ratio of frequency of two words that occur together to the individual frequencies of two words. Semantic orientation is then determined by evaluating the difference of PMI values computed against negative words from PMI values computed against positive words.

D. Semantic Approach

This approach is based on a simple principle that words that are semantically related should have similar sentiment orientation values [2]. The dictionary that is used to determine the sentiment polarities of words is WordNet. It is capable of determining different types of relations among words. A word can have different meanings that vary according to the context for which it is used. This leads to ambiguity and using WordNet solves this problem to some extent as it determines the senses of words. In this approach, two sets of seed lists are used. One seed list has positive opinion words and the other has negative opinion words. It serves as the base for constructing the dictionary. Based on this dictionary, the semantic orientation of the words is determined. Hu and Liu [3], [4], [5] used WordNet for obtaining a list of sentiment words by expanding the seed list iteratively using antonyms and synonyms. It has a drawback that it did not deal with the opinion words that are context dependent.

Ding, Liu and Yu [6] used a holistic lexicon based approach for dealing with the context dependent opinion words. This approach implemented special linguistic rules. Instead of looking at the current sentence only, it also refers to the other sentences. It helped in dealing with the context related opinion words. This approach was the improvement of the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

method used in [3].

M. Taylor et al [8], [9], [10] applied the feature-based opinion mining techniques on the reviews of hotels and restaurants. For extracting features, association rule mining was applied. Holistic lexicon based approach [6] was then used for the sentiment classification. This approach was capable of extracting only 35% of the explicit features.

M. Eirnaki et al. 2011 [7] proposed a feature-based opinion mining technique. For identifying the features from reviews, it used High Adjective Count and for the purpose of sentiment classification it used Max Opinion Score algorithm. For identifying the polarity of reviews, features extracted from title are considered separately.

E. Syntactic Approach

This approach is based on the idea that if the reviews can be parsed such that it corresponds to some syntactic structure, then extracting features and opinions from that structure might get easier. Chinsha T C, Shibily Joseph. [16] used this approach effectively. It used dependency parser to obtain the syntactic structure of each sentence. It then used the dictionary (SentiWordNet) and the combinations of adverb-adverb, adverb-verb and adverb-adjective to extract the features and their corresponding opinions. It showed an accuracy of 78.04%. Another work using this approach was proposed in [17]. It was based on pattern matching. In this method, firstly the patterns are constructed using the adjectives, adverbs, verbs and nouns of variable lengths. It then works by segmenting the reviews into fragments with simpler structures. Then the fragments are matched with the patterns to extract the features and their corresponding opinions. Finally, feature grouping method is used to prune the features and to extract the other infrequent features too.

V. PERFORMANCE MEASURE

responseThe performance of these approaches can be evaluated using three most important measures other than entropy and purity. These are the methods that are generally used for evaluating the information retrieval approaches. These measures are briefly described below:

A. *Recall*

It is the ratio of the retrieved features or opinion words that are relevant to the search, to the relevant opinion words or features. For the feature or opinion word extraction approach to be successful, this measure should be high.

B. *Precision*

It is the ratio of the retrieved opinion words or features that are relevant to the search, to the retrieved opinion words or features. This measure should be high too because it is required that all the relevant opinion words or features should be successfully retrieved.

C. *Accuracy*

It is the measure of the percentage of reviews that are successfully classified. For any feature extraction or sentiment detection approach to be successful, its accuracy should be high.

VI. CHALLENGES AND LIMITATIONS

To mine the opinions from the reviews available on the web is a very challenging task. Since the users of the web do not follow any fixed structure while writing the reviews, it is difficult to use any fixed approach for mining them. Following problems occur while mining reviews from the web:

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

1) Users do not follow any fixed format while writing reviews.

2) Users may use informal language and might use various adjectives to describe a single feature.

3) A word may have different meaning in different contexts. It gets difficult to understand that in which context a word has been used.

4) Users also use slangs or shorthand while writing reviews which cannot be detected using Standard English rules.

5) Users also use emoticons which are hard to detect.

6) Every approach is designed on a particular domain and might not work well on the other domains.

7) Reviews are generally written by different users from across the world, so there is difference in languages being used by them.

8) A lot of fake reviews and the spam reviews are available on the net too.

VII. CONCLUSION

Feature-based opinion mining has a great use for the customer as well as for the manufacturers. Customers find it beneficial to consult reviews before buying anything. Manufacturers on the other hand use this information to check how their product is being perceived by its customers and they can also keep a check on their competitors' products. This paper discusses several approaches being proposed for extracting the features and their corresponding opinions as well as the challenges that occur while mining reviews. It has a great scope and requires a lot of research to be made for addressing various issues like spammed reviews, fake reviews, handling dual negative words, etc.

REFERENCES

[1] Zhu, Jingbo, Huizhen Wang, Muhua Zhu, Benjamin K. Tsou, andMatthew Ma.,"Aspect-based opinion polling from customer reviews." IEEE Transactions on Affective Computing, vol. 2, pp. 37-49, 2011.
[2] Tsytsarau, Mikalai, and Themis Palpanas, "Survey on mining subjectivedata on the web.",Data Mining and Knowledge Discovery vol. 24, pp.478-514, 2012.
[3] Hu, Minqing, and Liu Bing, "Mining and summarizing customerreviews", In Proceedings of the tenth ACM SIGKDD internationalconference on Knowledge discovery and data mining, ACM, pp. 168-177, 2004.
[4] Hu, Minqing, and Liu Bing. "Mining opinion features in customer reviews", In AAAI, vol. 4, pp. 755-760, 2004.
[5] Liu, Bing, Minqing Hu, and JunshengCheng,"Opinion observer:analyzing and comparing opinions on the web", In Proceedings of the14th international conference on World Wide Web, ACM, pp. 342-351,2005.
[6] Ding, Xiaowen, Bing Liu, and Philip S. Yu, "A holistic lexicon-basedapproach to opinion mining", In Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, pp. 231-240, 2008.
[7] EirinakiMagdalini, ShamitaPisal, and Japinder Singh, "Feature-basedopinion mining and ranking",Journal of Computer and System Sciences,vol. 78, pp. 1175-1184, 2012.
[8] Marrese-Taylor, Edison, Juan D. Velásquez, Felipe Bravo-Marquez, andYutaka Matsuo,"Identifying customer preferences about tourism products using an aspect-based opinion mining approach." , ProcediaComputer Science, vol. 22, pp. 182-191, 2013.
[9] Marrese-Taylor, Edison, Juan D. Velásquez, and Felipe Bravo-Marquez, "Opinion Zoom: A Modular Tool to Explore Tourism Opinions on theWeb", In Web Intelligence (WI) and Intelligent Agent Technologies(IAT), IEEE/WIC/ACM International Joint Conferences, vol. 3, IEEE,pp. 261-264, 2013.
[10] M. Taylor, Edison, Juan D. Velásquez, and Felipe Bravo-Marquez, "Anovel deterministic approach for aspect-based opinion mining in tourismproducts reviews", Expert Systems with Applications, vol. 41, pp. 7764-7775, 2014.
[11] David D. Lewis,"The Independence Assumption in Information Retrieval", In *Proc. of the European Conference on Machine Learning (ECML)*, p.p. 4–15, 1998.
[12] Pedro Domingos and Michael J. Pazzani, "On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. Machine Learning", 29(2-3):103–130,1997.
[13] Qiu, gang., Bing, Liu., Jiajun Bu and Chun Chen,"Expanding Domain Sentiment Lexicon Through Double Propogation", In Proceedings of IJCAI, 2009.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

[14] Turney, Peter D. "Thumbs up or thumbs down? Semantic orientationapplied to unsupervised classification of reviews", In Proceedings of the40th annual meeting on association for computational linguistics,Association for Computational Linguistics, pp. 417-424, 2002.
[15] Turney, Peter D., and Michael L. Littman, "Measuring praise andcriticism: Inference of semantic orientation from association" ACM Transactions on Information Systems (TOIS), vol. 21, pp. 315-346,2003.
[16] Chinsha T C, Shibily Joseph, "A Syntactic Approach for Aspect Based Opinion Mining",Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015).
[17] Yao pan, yuwang ."Mining product features and opinions based on pattern matching" International Conference on science and network technology,2011.