IOSR Journal of Computer Science (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727
PP 10-18
www.iosrjournals.org

A Survey on Feature Mining In Customer Reviews Using Soft Computing

Minal M.Thawakar[1], Prof.S.S.Patil[2], Prof.Dr.G.R.Bamnote[3]
[1,2]*Department of Computer Technology , Priyadarshini College of Engineering ,Nagpur , India*
[3]*Department of Computer Scence &Engineering , PRM Institute of Technology & Research, Badnera, Amravati , India*

ABSTRACT : *Internet is the global system which growing very fast. It is most reliable and efficient one so that the use of internet is increasing in people's day to day life. Because of increasing social networking more and more people interact with each other and share their views, emotions, experiences, feedback and opinion about anything. Feedback is the important part for selling or purchasing any product. But it is very difficult for customer to read thousands of reviews at a time which create confusion. So data mining plays an important role to mine opinion and to summarize all reviews of customer. Most of the existing methods of opinion mining show the customer reviews in the form of positive and negative comments. But it is not efficient for customers because customer will not decide whether to buy a product or not. The proposed approach mines the opinions of customers according to product features. The proposed approach not only gives the method for rating products but also gives rating according to features. This approach also compares the product according to the rating which helps customer to take decision regarding product purchasing.*

Keywords -Feature mining, Opinion mining, Sentiment, Sentiment Classification, Summarization

I. INTRODUCTION

The Web contains a wealth of opinions about products, newsgroup posts, review sites, politics and elsewhere. It is the only medium which contains a lot of information and this information is available to people online. Generally people express their views, ideas, emotions, experiences, opinion to each other, for this social media is the only source through which people can interact with each other. Feedback or opinion is very important for customer as well as producer point of view, as most of the people purchase or sale product online. There are various web sites are available which gives information about reviews of products. But it is very difficult for the customer to read hundreds or thousands of comments at a time which not only create confusion in mind but also take lot of time to take decision. So Data Mining is the only field which is used to mine opinion and summarize all reviews of customer.

Opinions are central to almost all human activities because they are key influencers of our behaviors, views and emotions. Whenever we want to take any decision, we want to know others' opinions. In the real world, businesses and organizations always need consumer or public opinions about their products and services. Similar way individual consumers also want to know the opinions of existing users of a product before purchasing it, and others' opinions about political candidates before making a voting decision in a political election. In the past, whenever an individual needed opinions, he/she asked friends and family. On the other hand an organization or a business collected public or consumer opinions by using surveys, opinion polls, and focus groups.

With the tremendous growth of social media on the Web, individuals and organizations are increasingly using the content in these media for decision making. For example reviews, blogs, micro-blogs, comments, and postings in social network sites i.e. twitter are highly available on web. Nowadays, if anyone wants to buy a consumer product, he/she does not depend on friends or family members for opinions because there are many user reviews and discussions in public forums on the Web about the product. For an organization, it may no longer be necessary to conduct surveys, opinion polls, and focus groups in order to gather public opinions because such information publicly available. Each site typically contains a huge volume of opinion text that is not always easily deciphered in long blogs and forum postings. The average human reader will have difficulty identifying relevant sites and extracting and summarizing the opinions from them. So, there is a need to extract and identify sentiments i.e. automated sentiment analysis.

Opinion mining is the field of study that analyzes people's opinions, emotions, attitudes and sentiments towards entities such as products, services, organizations, individuals. Opinion is the word which is used to denote opinion, sentiment, evaluation, attitude, and emotion. There are also many names for opinion mining but according to name the task performed by them is slightly different, e.g., sentiment analysis, sentiment mining, opinion mining, opinion extraction, subjectivity analysis, emotion analysis, review mining, etc. Opinion mining is a type of natural language processing for tracking the sentiment or thinking of the public about a particular product. It is the research area in NLP as well as Data Mining, Web mining, Machine Learning and information retrieval.

Opinion mining is useful in several ways. It acts as market intelligence for business and organization. As a huge amount of money is spend to find consumer (customer) sentiments and opinions. It is the source which gives information about whether to increase or stop the production of any product. It is also useful for customer in order to purchase or sale any product. Opinion search provide a search for the opinions, gives opinion on any product and compare two different products.

An important task of opinion mining is to extract people's opinion on the basis of features of an entity. For example, the sentence, "I like the PIM (Personal Information Management) function of Nokia Lumina" expresses a positive opinion on the "PIM function" of the Nokia phone. "PIM function" is the feature. The sentence, "The picture quality of this camera is awesome", expresses a positive opinion about picture of the camera. "Picture" is the feature. Now how to extract the features is an important problem. The process of extraction and mining of features is called as feature mining.

For convenience the remaining paper is organized as follows: Section 2 presents the data sources used for opinion mining. Section 3 introduces challenges in opinion mining. Section 4 presents related work of different researchers. Then applications of opinion mining are given in section 5.The sixth section is about the performance evaluation done. Last section concludes our study and discusses some future directions for research.

II. DATA SOURCE FOR COLLECTING OPINIONS

There are various sources of data for collecting views and feedback of people about a particular product. Blogs, review sites, data and micro blogs provide a good understanding of the products and services.

2.1 Ecommerce Sites

Electronic commerce is commonly known as "e-commerce" or "eCommerce". Sometimes it is also known as "e-business". E-commerce is a wide range of online business activities for products and services. It is a type of business model, which enables a firm or individual to conduct business over an electronic network, typically the internet. It pertains to any form of business transaction in which the parties interact electronically. But they don't interact by physical exchanges or direct physical contact. Electronic commerce operates in four market segments: business to business (B-to-B), business to consumer (B-to-C), consumer to consumer (C-to-C) and consumer to business (C-to-B). It can be thought of as a more advanced form of mail-order purchasing through a catalog. Almost any product as well as service can be offered by using ecommerce, from books and music to financial services and plane tickets. There are various sites available for e-commerce. The top ten sites for e-commerce according to the different categories are as follows:

- eBay.in
- Flipkart.com
- Jabong.com
- HomeShop18.com
- Tradus.in
- Shopping.IndiaTimes.com
- SnapDeal.com
- Infibeam.com
- Zovi.com
- Myntra.com

2.2 Blogs

A blog is a discussion or informational site published on the World Wide Web and consisting of "posts" typically displayed in reverse chronological order i.e. the most recent post appears first. Many blogs provide commentary on a particular subject; others function as more personal online diaries; others function more as online brand advertising of a particular individual or company. A typical blog is a combination of text, images, and links to other blogs. There are different types of blogs like media blog which discuss about any media issue or any particular media only. Political blog which discuss about any political issue or about any politician. Travel blog contains the reviews of people about travel companies or agencies and their travel experiences with them. Health blog contains the opinions of people about any disease or about any hospital. In the same way there are educational blog, device blog, corporate blog etc. The top blog sites are as follows:

- WordPress.com and WordPress.org
- Blogger.com
- Tumblr.com
- Medium.com
- ***Svbtle.com***
- ***Quora.com***

2.3 Micro-blogging

Microblogging is a broadcast medium in the form of blogging. A microblog is different from a traditional blog in that its content is typically smaller in both actual and aggregate file size. Microblogs not only allow users to exchange small elements of content but also to post short messages. These small messages are sometimes called "microposts." Microposts can be made public on a Web site and/or distributed to a private group of subscribers. Subscribers can read microblog posts online or request that updates be delivered in real time to their desktopas an instant message or sent to a mobile device as an SMS text message. Mostly used microblogging sites are as follows:

- ***Twitter***
- ***FriendFeed***
- ***Plurk***
- ***GoogleBuzz***
- ***SpotJots***

2.4 Review Sites

A review site is a website on which reviews can be posted about people, businesses, products, or services. Web 2.0 techniques can be used to gather reviews from site users or may employ professional writers to author reviews on the topic of concern for the site. Early review sites included are as follows:

- www.Epinions.com (Product reviews)
- www.Amazon.com(Product reviews)
- www.Yelp.com (restaurant reviews)
- www.CNETdownload.com(product reviews)
- www.reviewcentre.com (product reviews)
- www.dpreview.com (professional review)

2.5 Data Sets

This dataset is a subset of the opinion mining datasets released by Dr. Bing Liu's group from University of Illinois at Chicago. Their dataset is available from http://www.cs.uic.edu/~liub/FBS/sentiment analysis.html. This subset consists of plenty of review comments each for various different products.

Most of the work in the field uses movie reviews data for classification. Movie review data's are available as dataset (http://www.cs.cornell.edu/People/pabo/moviereview-data). Other dataset which is available online is multi-domain sentiment (MDS) dataset. (http:// www.cs.jhu.edu/mdredze/datasets/sentiment). The MDS dataset contains four different types of product reviews extracted from Amazon.com including Books, DVDs, Electronics and Kitchen appliances, with 1000 positive and 1000 negative reviews for each domain. Another review dataset available is http://www.cs.uic.edu/liub/FBS/CustomerReviewData.zip. This dataset consists of reviews of five electronics products downloaded from Amazon and Cnet. This subset is used for the experiments conducted in various research papers. Dr. Bing Liu's group is very much active is the field of opinion mining and has too many research papers based on the opinion mining theory. These datasets are prepared by this group for the purpose of research in the field of opinion mining.

III. CHALLENGES IN OPINION MINING

Opinion Mining/Sentiment Analysis is a somewhat recent subtask of Natural Language processing. There are several challenges in Sentiment analysis. The first is an opinion word that is considered to be positive in one situation may be considered negative in another situation. For example: take the word "long". If a customer said that the battery life of laptop was long, it indicates a positive opinion. If the customer saidthat the start-up time of laptop was long, it indicates a negative opinion.

A second challenge is that people don't always express opinions the same way. Most traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much. For example: "the movie was great" is very different from "the movie was not great".

Finally, people can be contradictory in their statements. Most reviews will have both positive as well as negative comments. For example: "the movie bombed even though the lead actor rocked it" is easy for a human to understand, but more difficult for a computer for parsing.

There are two types of opinions: direct opinion and comparative opinion. Direct opinion is also called as regular opinion. Direct opinion is nothing but sentiment expressions on some objects, e.g., products, topics, persons, events. For example: "the picture quality of this digital camera is great". Comparative opinions are nothing but the relations expressing similarities or differences of more than one object. Usually expressing an ordering. For example, "car 'p' is cheaper than car 'q'."

The basic components of an opinion are as follows: Opinion holder is the person or organization that holds a specific opinion on a particular object. Object is on which an opinion is expressed and Opinion is a view, attitude, or emotions on an object from an opinion holder. Opinion may be Positive, negative and neutral are called opinion orientations also called sentiment orientations, semantic orientations, or polarities.

IV. RELATED WORK

The following sections explain the survey of various papers. Different methods are used for extracting sentiments or opinions from the given sentences by many researchers. Following section also explain different methods that are used to rank or classify these opinions extracted from the sentences.

In [5], Minqing H., Bing L explain the problems related to feature based summarization of customer reviews of product which are sold online. Here a given set of customer reviews of a particular product is divided into three subtasks. First identifying features of the product that customers have expressed their opinions on, which is called as product features. Second for each feature, identifying review sentences that give positive or negative opinions and finally producing a summary using the discovered information.

The inputs to the system are a product name and an entry page for all the reviews of the product. The output is the summary of the reviews. In this method NLProcessor linguistic parser (NLProcessor 2000) is used, which parse search sentence and yields the part-of-speech tag of each word and identifies simple noun and noun phrases. Once the part-of-speech tags are extracted they are used to find the frequency of features referred, which in turn gives us an idea about what is the opinion direction. The limitation of this paper is, it does not show the strength of the opinion.

In [7], Satoshi M., Kenji Y., Kenji T., Toshikazu F studies the opinion mining data and presented it in the form of reputation of the product in the market. The user has to put a product name. The search engine will extract the web pages contacting any information about the product, then the opinions are extracted. The opinions are evaluated by using evaluation-expression dictionary, then opinion likeliness is calculated in which is in turn used for calculation of reputation of product.

In [8],Kushal Dave, Steve Lawrence, David M. Pennock classified the opinions into four classes: unigrams, bigrams, trigrams and distance 3 patterns. This classification is done on the basis of number of words used forgiving the opinions. Final representation of result is based on how much percentage of people has given opinion in uniform pattern, how many numbers of people has given opinion in bigrams pattern and how many numbers of people has given opinion in trigram and distance 3 patterns. This classification is done for identification of both positive and negative review sentences.

In [9], Jiaming Z., Han T., Ying L. uses automatic text summarization where frequent sequences are found by using single word, two words or multi word patterns. These frequent sequences are used as topic and again the document is searched for these topics and candidate sentences are extracted. These sentences are used to extract opinions and final customer concerns are identified.

In [10] [11], Nitin J., Bing L. studies the problem of identifying comparative sentences in text documents. The problem is related but quite different from sentiment/opinion sentence identification or classification. Identifying comparative sentences is also useful in practice, as direct comparisons are one of the most convincing ways of evaluation, which may even be more important than opinions on each individual object. It first classifies comparative sentences into different types, and then presents a novel integrated pattern discovery. A supervised learning approach is used for identifying comparative sentences from text documents.

In [12], Zhongwu Z., Bing L., Hua X., focuses on clustering. Clustering is nothing but grouping of synonym features. The whole process is divided into three steps. The first step is to connect feature expressions using sharing words, e.g., "customer service", "customer support", "service". The next step is merge components using lexical similarity. Lexical similarity based on WordNet and which is widely-used in the NLP area to measure the similarity of two words.. For example, "picture" and "image" has very high similarity in WordNet. In final step it selects the leader components as labeled data.

In [13], Lei Z., Bing L., Suk H. L., Eamonn O'Brien-Strain focuses on the problems of double propagation. Double propagation assumes that features are nouns/noun phrases and opinion words are adjectives. Opinion words are usually associated with features. Thus, opinion words can be identified by identified features, and features can be identified by known opinion words. The extracted opinion words and features are used to identify new opinion words and new features, which are used to extract more opinion words and features. This propagation process ends when no more opinion words or features can be found. The advantage of the method is that it requires no additional resources. But it requires an initial opinion lexical analyzer.

In [1],Zhu Jian explained the use of neural networks in sentiment classification and proposed an individual model based on Artificial neural networks to divide the movie review corpus into positive , negative and fuzzy tone which is based on the advanced recursive least squares back propagation training algorithm. In [2], Long-Sheng Chen proposed a neural network based approach, which combines the advantages of the machine learning techniques and the information retrieval techniques.

V. APPLICATIONS

Opinions are so important that whenever one needs to make a decision, one wants to hear others' opinions. This is true for both individuals and organizations. The technology of opinion mining thus has a tremendous scope for practical applications

Opinion mining can be useful in many ways. If you are in marketing, for example, it can help you to judge the success of an ad campaign or new product launch, to determine which versions of a product or service are popular and even to identify which demographics like or dislike particular features. For example, a review for the digital camera is highly positive, but due to high weight of the camera the review is negative. Being able to identify this kind of information in a systematic way gives the vendor a much clearer picture of public opinion than surveys or focus groups, because the customer creates a data.

Opinions matter a great deal in politics. Some work has been focused on understanding what voters are thinking. It is widely used in business, as it acts as marketing intelligence as well as product and service benchmarking and improvement. It is used to understand the voice of the customer as expressed in everyday communications.

VI. EVALUATION AND DISCUSSION

Table 1. represents the objectives and performances of various approaches along with mining techniques, data sources and feature selection.

Sr.No.	Studies	Mining technique used	Feature selection	Data source	objectives
1	KaiquanXu(2011)	Multiclass SVM	Linguistic Feature	Amazon reviews	structural risk minimization principle is used, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors
2	Long Sheng(2011)	BPN	Point wise Mutual information	Movie review	combines the advantages of the machine learning techniques and the information retrieval techniques.
3	Rui Xia (2011)	Naïve bayes, Maximum entropy, SVM	Uni gram, bi grams, deendency grammar, joint feature	Movie review, Multi domain dataset,Amazon	to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories
4	XueBai (2011)	Naïve bayes,	Information gain, two stage markov blanket classifier	Moview review,	to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories
5	Ziqiong (2011)	Naïve bayes,SVM	Information gain	Cantonese reviews	to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories

6	Gamgamso mprasti (2010)	Maximum Entropy	Dependency relation	Amazon reviews	Mining product feature by applying dependency relation and ontological knowledge
7	Gang li (2010)	K-means Clustering	TF-IDF	Moview review	clustering or grouping of synonym features,voting mechanism is used to extract a more stable clustering result
8	Yulan He (2010)	Sentiment lexicon, General expectation criteria	Self trained features	Movie review	to create a novel framework for sentiment classifier learning from unlabeled documents
9	Zhu Jian (2010)	Back propogation	Odds ratio	Movie review	Develop an individual model based on Artificial neural networks to divide the movie review corpus into positive , negative and fuzzy tone which is based on the advanced recursive least squares back propagation training algorithm
10	Melville (2009)	Bayesian classification	n-grama	Blogs	to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories
11	Rudy(2009)	ID3,SVM,Hybrid	Document frequency	Moview review, mySpace comments	SVM classifier is considered the best text classification method
12	QingliangMi ao (2009)	Lexical resource	POS,Apriori	Amazon reviews	Use Naïve bayes classifier to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories
13	Songho tan (2008)	Centroid classifier, K-Nearest neighbourhood,W innow Classifier, SVM	POS,Apriori	Chnsenticor p	Use Naïve bayes classifier to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories
14	Kennedy and Inkpen (2006)	support vector machines, term counting	term frequencies	Movie review	evaluate a negation model in document-level polarity classification
15	Hu and Liu (2005)	Opinion word extraction and aggregation	Opinion words	Amazon Cnn.Net	Opinion extraction and summarization of customer reviews

Table 1: Summary of different approaches by researchers

VII. CONCLUSION

Opinion mining is a new field of study. It is most important field in this competitive world because every customer try to compare multiple products before purchasing. Also the organization needs customer opinion about their products to be in the competition and to put improvements in their products. This is a recent trend in research also.

Opinion Mining has become a latest trend in the information mining industry. There is plenty of future scope for opinion mining as it requires Natural Language Processing, Data mining and also Artificial Intelligence. A more work is needed on further improving the performance measures. The main challenging aspects exist in use of other languages, dealing with negation expressions; produce a summary of opinions based on product features/attributes, complexity of sentence/ document, handling of implicit product features, etc. More future research is dedicated to these challenges.

References

[1] ZHU Jian , XU Chen, WANG Han-shi, Sentiment classification using the theory of ANNs, The Journal of China Universities of Posts and Telecommunications, July 2010, 17(Suppl.): 58–62 .

[2] Long-Sheng Chen , Cheng-Hsiang Liu, Hui-Ju Chiu , A neural network based approach for sentiment classification in the blogosphere, Journal of Informatics 5 (2011) 313–322.

[3] Ana-Maria P., Oren E., Extracting Product Features and Opinions from Reviews. In Proceeding of HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing Pages 339-346.

[4] Nozomi K., Kentaro I., Yuji M. Extracting Aspect Evaluation and Aspect-of Relations in Opinion Mining. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLPCoNLL).

[5] Minqing H., Bing L. Mining Opinion Features in Customer Reviews. In Proceedings of AAAI'04 Proceedings of the 19th national conference on Artificial intelligence Pages 755-760.

[6] Nikolay A., Anindya G., Panagiotis G. I. Show me the Money! Deriving the Pricing Power of Product Features by Mining Consumer Reviews. In Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2007).

[7] Satoshi M., Kenji Y., Kenji T., Toshikazu F. Mining Product Reputations on the Web. In the Proceedings of KDD '02 Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. Pages 341-349.

[8] Kushal Dave, Steve Lawrence, David M. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In the Proceedings of WWW '03 Proceedings of the12th international conference on World Wide Web. Pages 519 528.

[9] Jiaming Z., Han T., Ying L. Gather customer concerns from online product reviews – A text summarization approach. In the Proceedings of Journal on Expert Systems with Applications: An International Journal Volume 36 Issue 2, March,2009. Pages 2107-2115.

[10] Nitin J., Bing L. Identifying Comparative Sentences in Text Documents. In the Proceedings of the 29th Annual International ACM SIGIR Conference2006

[11] Nitin J., Bing L. Mining Comparative Sentences and Relations. In the Proceedings of AAAI'06 The21st national conference on Artificial intelligence -Volume 2 Pages 1331-1336.

[12] Zhongwu Z., Bing L., Hua X., Peifa Jia. Clustering Product Features for Opinion Mining. In the Proceedings of WSDM '11 The fourth ACM international conference on Web search and data mining. Pages 347-354.

[13] Lei Z., Bing L., Suk H. L., Eamonn O'Brien-Strain. Extracting and Ranking Product Feature in Opinion Documents. In the Proceedings of COLING '10 The 23rd International Conference on Computational Linguistics. Pages 1462-1470.

[14] Dongjoo L., Ok-Ran J., Sang-goo L. Opinion Mining of Customer Feedback Data on the Web. In the Proceedings of ICUIMC '08 The 2nd international conference on Ubiquitous information management and communication. Pages 230-235.

[15] Soo-Min K., Eduard H. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In Proceedings of the COLING/ACL, an ACL Workshop on Sentiment and Subjectivity in Text.

[16] Ali H., Michel P., Gerard D., Mathieu R., François T., Pascal P. Web Opinion Mining: How to extract opinions from blogs?. In Proceedings of CSTST '08the 5th international conference on Soft computing as trans disciplinary science and technology. Pages 211-217.

[17] Xiaowen D., Bing L., The utility of linguistic rules in opinion mining. In Proceedings of SIGIR '07the 30th annual international ACM SIGIR conference on Research and development in information retrieval. Pages 811-812

[18] Andrea E., Fabrizio S., SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06).