International Journal of Advanced Science and Technology
Vol.60, (2013), pp.1-8
http://dx.doi.org/10.14257/ijast.2013.60.01

Research Directions, Challenges and Issues in Opinion Mining

Periakaruppan Sudhakaran[1], Shanmugasundaram Hariharan[2] and Joan Lu[3]

[1]Assistant Professor, Oxford Engineering College,India
[2] Associate Professor, TRP Engineering College,India
[3]Professor, University of Huddersfield,UK
karan_sudha@rediffmail.com, mailtos.hariharan@gmail.com, j.lu@hud.ac.uk

Abstract

Rapid growth of Internet and availability of user reviews on the web for any product has provided a need for an effective system to analyze the web reviews. Such reviews are useful to some extent, promising both the customers and product manufacturers. For any popular product, the number of reviews can be in hundreds or even thousands. This creates difficulty for a customer to analyze them and make important decisions on whether to purchase the product or to not. Mining such product reviews or opinions is termed as opinion mining which is broadly classified into two main categories namely facts and opinions. Though there are several approaches for opinion mining, there remains a challenge to decide on the recommendation provided by the system. In this paper, we analyze the basics of opinion mining, challenges, pros & cons of past opinion mining systems and provide some directions for the future research work, focusing on the challenges and issues.

Keywords: *Opinion mining, user reviews, sentiment analysis, facts, NLP*

1. Introduction

The World Wide Web is increasing at an alarming rate not only in size but also in the types of services and contents provided. The users are participating and generating vast amount of new information and data. With the rapid growth of e-commerce, more and more products are sold on the web and more number of customers is also buying products online. The web has become a tremendous source for collect consumer reviews. The customer reviews collected from e-commerce websites, forums, discussion groups and blogs *etc.*, and it's useful to both potential customers and product manufacturers. With help of this reviews the customer to decide which product to buy, and the manufacturers to rectify the problems in the product. Due to the large number of reviews, it is hard for potential customer to get efficient review details, it is a challenging task.

Opinion Mining is an area of text mining that has recently received a lot of attention due to huge amount of opinion or reviews available in web documents. Thus Opinion mining is also called as sentiment analysis. Business people are spending a lot of money to find customer sentiments or opinions about products. Before the internet era opinions was to share verbally or through some written format, ask friends to suggest which product is the best. After internet has grown it has become likely to share experiences and views through forums, discussion board & blogs. In general, opinions are classified into two categories as:

1) *Direct opinions:* Direct sentiment expression on some products or services. For example: sound clarity was good for this speaker.

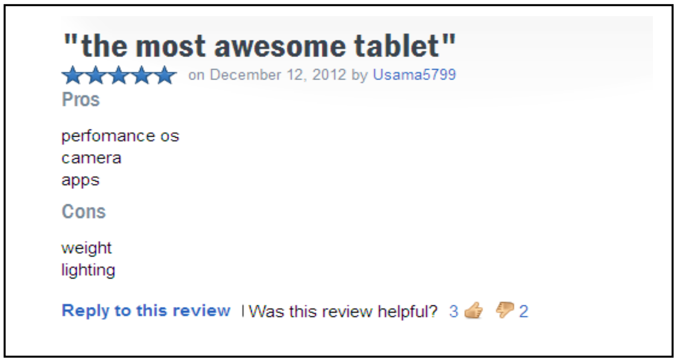
ISSN: 2005-4238 IJAST

2) *Comparative opinions*: comparisons expressing between two products or services. For example: Speaker X is better than Speaker Y.

The opinions are retrieved from customers are based on three general formats. These are

1) *Pros and Cons (Format 1):* The reviewer is invited to depict pros and cons separately.
2) *Pros and Cons and detailed review (Format 2):* The reviewer is invited to depict pros and cons separately and also write a detailed review.
3) *Free Format (Format 3):* The reviewer can write his opinions freely *i.e.*, no separation for pros and cons.

The free format review is a challenging for research area. It concerns the identification of opinions in a text and their classification as positive, negative or neutral. The identification of the opinion orientation can be manual, corpus based and dictionary based. The manual orientation requires a lot of human effort and costly. The corpus based orientation considers syntactic and statistical properties. The dictionary based orientation uses hierarchies and ontology. Automatic discovery of opinions in text is becoming increasingly important. Online review websites like Amazon[a], IMDB[b], Epinions[c], Cnet[d], eBay[e] *etc.*, allow users to express their opinions for the information they are interested in. So there is a huge amount of information available in online documents that are useful to both customers and manufacturers. A sample review for each of the formats is presented below in Figures 1, 2 and 3 respectively.



"the most awesome tablet"
★★★★★ on December 12, 2012 by Usama5799
Pros

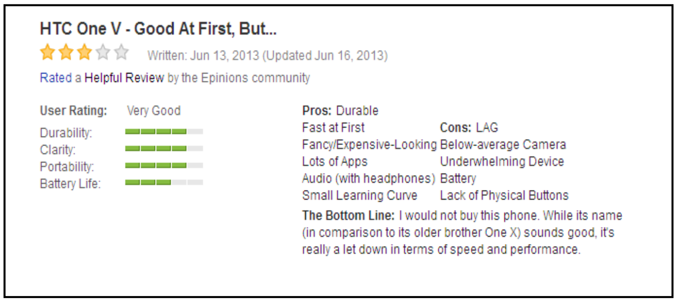
perfomance os
camera
apps

Cons

weight
lighting

Reply to this review | Was this review helpful? 3 👍 👎 2

Figure 1. Sample Product Review for Format 1



HTC One V - Good At First, But...
★★★☆☆ Written: Jun 13, 2013 (Updated Jun 16, 2013)
Rated a Helpful Review by the Epinions community

User Rating: Very Good
Durability: 🟩🟩🟩🟩⬜
Clarity: 🟩🟩🟩🟩⬜
Portability: 🟩🟩🟩🟩⬜
Battery Life: 🟩🟩🟩⬜⬜

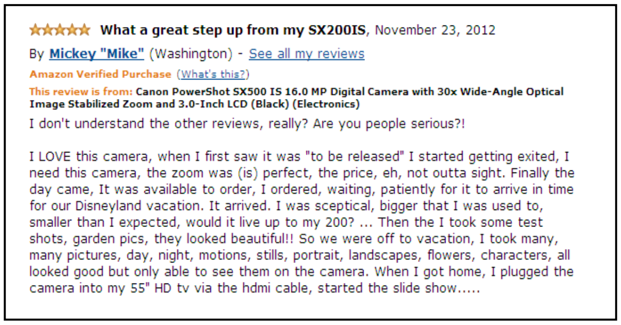
Pros: Durable
Fast at First
Fancy/Expensive-Looking
Lots of Apps
Audio (with headphones)
Small Learning Curve

Cons: LAG
Below-average Camera
Underwhelming Device
Battery
Lack of Physical Buttons

The Bottom Line: I would not buy this phone. While its name (in comparison to its older brother One X) sounds good, it's really a let down in terms of speed and performance.

Figure 2. Sample Product Review for Format 2

[a]www.amazon.com, [b]www.cnet.com, [c]www.ebay.com,[d]www.epinions.com, [e]www.imdb.com



★★★★★ **What a great step up from my SX200IS**, November 23, 2012

By **Mickey "Mike"** (Washington) - See all my reviews

Amazon Verified Purchase (What's this?)

This review is from: Canon PowerShot SX500 IS 16.0 MP Digital Camera with 30x Wide-Angle Optical Image Stabilized Zoom and 3.0-Inch LCD (Black) (Electronics)

I don't understand the other reviews, really? Are you people serious?!

I LOVE this camera, when I first saw it was "to be released" I started getting exited, I need this camera, the zoom was (is) perfect, the price, eh, not outta sight. Finally the day came, It was available to order, I ordered, waiting, patiently for it to arrive in time for our Disneyland vacation. It arrived. I was sceptical, bigger that I was used to, smaller than I expected, would it live up to my 200? ... Then the I took some test shots, garden pics, they looked beautiful!! So we were off to vacation, I took many, many pictures, day, night, motions, stills, portrait, landscapes, flowers, characters, all looked good but only able to see them on the camera. When I got home, I plugged the camera into my 55" HD tv via the hdmi cable, started the slide show.....

Figure 3. Sample Product Review for Format 3

The rest of the paper is structured as follows. While basics introduction is presented in Section 1, Section 2 presents some background studies on opinion mining. Section 3 presents the existing method and section presents the proposed method.

2. Background Study

Yue Lu *et al.*, [5] focused on the problem of generating a “rated aspect summary" of short comments. The input is a large number of short comments about a target entity. Each comment is associated with an overall comment. The solution to the problem involved three steps: (1) Aspect Discovery and Clustering (2) Aspect Rating Prediction (3) Representative Phrases Extraction. The effectiveness of this method was demonstrated by using eBay sellers' feedback comments.

Keke Cai *et al.*, [2] designed a sentiment analysis system that included a sentiment classification component and sentiment topic words recognition component. The authors found that sentiment classification alone was insufficient because sentiment classification summarizes customer’s opinions but does not reveal the implicit root cause of the sentiments. Such reasons were called the “sentiment topics” linked with the sentiments. Sentiment topic words recognition component detects the topics hidden behind the positive or negative opinions based on a combined PMI (Pointwise Mutual Information) and word support metrics. The sentiment classification process is based on a semantic based approach that mainly depends on opinion word collection in the form of knowledge base or sentiment dictionary. Using real-world usage cases the authors verified the effectiveness of their technique on reviews posted on blogs, message boards, etc.

Bing Liu *et al.*, [1] developed a feature-based opinion summarization system that mines and summarizes all the customer reviews of a product. This system differs from other text summarization systems in number of ways. It mines only the features of the product that customers have expressed their opinions on and the summary generated is structured rather than a free text document as generated by other text summarization systems. The summarization is performed by the system in three steps: (i) Finding the features on which the customers have expressed their opinions. (ii) Identifying opinion sentences in each review and analyzing the polarity of the sentence (iii) Producing a summary.

Rudy Prabowo and Mike Thelwall [11] evaluated the effectiveness of different classifiers, and shows that the use of multiple classifiers in a hybrid manner can improve the effectiveness of sentiment analysis in terms of micro and macro-averaged than any individual classifier. This paper combines rule-based classification, supervised learning and machine learning into a new combined method. This method reviews number of automatic document classification techniques. This also includes semi-automatic, complementary approach in which each classifier can contribute to other classifiers to achieve a good level of effectiveness.

Ronen Feldman *et al.*, [7] have presented a set of techniques for analyzing how consumers view product markets. They have extracted relative sentiment analysis and comparisons between products, to understand what attributes users compare products on, and which products they prefer on each dimension. This system also compares state-of-the-art machine learning methods against simple hand-coded patterns. The key steps in extraction of the product mentions are information extraction and snippet extraction. Snippet extraction involves pattern extraction and pattern filtering. They have used a fast algorithm that finds all matches of all patterns with complexity logarithmic in the number of patterns and linear in the corpus size. Finally summary of the results is generated.

Hang Cui, Vibhu Mittal and Mayur Datar [8] classified online product reviews into positive and negative classes. Three classifiers and n-grams are used as linguistic features to classify sentences. Classifier combined with high order n-grams as features can achieve better performance.

Jianshu Sun *et al.*, [3] proposed an automated system that could perform comparison and recommendation of products to the customers from both subjective and objective perspectives. They built an evolution tree that describes the evolutionary process of one type of product. The evolution tree can indirectly recommend customers about good products.

Jiaming Zhan *et al.*, [4] proposed a summarization process based on topical structure to automatically summarize multiple customer reviews. This method is different from other methods that are based on sentence ranking and clustering. From a set of online reviews this system extracts the prominent topics and ranks them according to their saliency. Based on the ranked topics a final summary is produced. The experimental study compared the summarization performance of the proposed method with the approaches of opinion mining and clustering-summarization. The experimental results showed that the proposed approach can achieve better summarization performance than other approaches.

A classification based approach was developed by Chin-Yew Lin *et al.*, [6] for detecting low quality product reviews. A set of specifications for judging the quality of reviews are defined. Based on the specifications the low-quality reviews are separated from high-quality ones then summarization is done.

Jeonghee Yi *et al.*, [9] designed and developed *Sentiment Analyzer* (*SA*) that *i)*extracts topic-specific features, *ii)* extracts sentiment of each sentiment-bearing phrase, *iii)* makes (topic| feature, sentiment) association. They extract only noun phrases from documents and apply two feature term selection algorithms such as Mixture Model and Likelihood Test. *SA* consistently demonstrated high quality results of 87% for review articles, 86 *to* 91% (precision) and 91 *to* 93% (accuracy) for the general web pages and news articles.

Prem Melville *et al.*, [10] used lexical information in terms of word-class associations combined with supervised learning. The two baseline approaches in classification of documents are used here i) Lexical classification ii) Feature supervision. They have proposed another method “Pooling Multinomials classifier” for text categorization. The generality of this approach on three, very different domains — blogs discussing enterprise-software products, political blogs discussing US Presidential candidates, and online movie reviews.

Tak-Lam Wong and Wai Lam [12] developed a technique for mining and summarizing hot items from multiple auction Web sites. In the auction Web sites, the items with high number of bids are automatically regarded as hot items. This approach is a two phase framework for mining and summarizing hot items. The first phase is to extract the product features and product feature values of the items and the second phase is to mine and summarize hot items. Hidden Markov model (HMM) is used to reliably extract the product features since the format of the description is greatly different ranging from regular format such as tables to unstructured free texts.

3. Existing Method

Existing works are mostly context insensitive and provide inaccurate results.

Example: *"The time taken to focus is long"*

In the above example, the word 'long' is considered to be a positive opinion word. Hence the sentence would be scored as a positive sentence. But the sentence expresses a negative opinion indirectly about the camera's zoom feature. Most of the existing algorithms do not analyze the context in which the words occur and perform scoring blindly.

Even the few existing context-sensitive algorithms are not capable of handling various review formats like the Pros-cons format and the freeform format. Different formats may need different techniques to perform the feature and opinion extraction task.

Analysis based on comparative study of reviews is often biased and inaccurate. A comparison is used to state that one object has more of a certain quantity than another object. Hence one can obtain the comparative information with respect to two or more products rather than the outright opinion about an individual product.

Also, most of the algorithms perform 'stemming' – the process for reducing inflected words to their root form, which often leads to losing the much important detail of the opinion word.

Example: *"iPhone is **light**, but MotoRAZR is even **lighter**."*

In the above example, if stemming was performed, the word 'lighter' would be reduced to 'light'. The word 'lighter' is the higher degree superlative form of the adjective 'light'. Hence the word 'lighter' carries higher value than the word 'light'. But most of the existing algorithms would give the same score for both of these words if stemming is performed.

4. Proposed Method

Crawled user-review from online review web sites is obtained as input through the user interface. Such reviews are called as unstructured reviews, since they are of free form expressed by users. Extraction involves splitting the review into units of sentences by considering the punctuation marks and conjunctions present in the review. In pre-processing, the non-informative words called stop words and special characters present in the review sentences are removed. Opinion mining identifies the feature words, negations, intensifiers, opinion words, neutral words in the review sentences. An appropriate score is assigned to each word. Each sentence is scored and the overall scores for all the sentences in the review are calculated.

The main advantages of this method are listed below:

- This method is context-sensitive.
- It promises more accurate results.
- It is capable of handling various review formats.
- Comparisons are not made and hence unbiased.
- Stemming is intentionally not performed, so that the details are retained.
- Values human psychology and scores depend on length of sentences.

5. Conclusion and Future Work

In this paper, we have proposed an automated system to analyze the user-review. This system recommends the product to the user only when the calculated recommendation percentage is above 50%. Apart from the product recommendation verdict, it also displays the feature scores for each feature. This method is domain independent and can be applied to any domain. Moreover this method is context sensitive and can be applied to any review format.

In the future work, we will first classify the sentences as subjective (opinions) or objective (facts) and then we will analyze only the subjective sentences thereby improving performance. We will try adding a spelling correction component in the pre-processing of the reviews. Also, we would add a smart crawler component so that all the relevant information from various web pages in a website is automatically crawled and extracted upon providing a seed URL and certain conditions.

Acknowledgements

The authors would like to express our special thanks to insightful suggestions for the anonymous reviewers. The first author would extend the support rendered by the management for carrying out the research work successfully.

References

[1] M. Hu and B. Liu, “Mining and summarizing customer reviews”, Proceedings of the Tenth ACM SIGKDD International conference on Knowledge discovery and Data mining, **(2004)**, pp. 168-177.

[2] K. Cai, S. Spangler, Y. Chen and L. Zhang, “Leveraging Sentiment Analysis for Topic Detection”, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, **(2008)**, pp. 265-271.

[3] J. Sun, C. Long, X. Zhu and M. Huang, “Mining Reviews for Product Comparison and Recommendation” Research journal on Computer Science and computer engineering with applications, no. 39, **(2009)**, pp. 33-40.

[4] J. Zhan, H. Tong Loh and Y. Liu, “Gather customer concerns from online product reviews-A text summarization approach”, Expert Systems with Applications, vol. 36, no. 2, Part 1, **(2009)**, pp. 2107-2115.

[5] Y. Lu, C. Xiang Zhai and N. Sundaresan, “Rated Aspect Summarization of Short Comments”, Proceedings of the 18th International Conference on World Wide Web**,** Madrid, Spain, **(2009)**, pp. 131-140.

[6] C.-Y. Lin, J. Liu, Y. Cao, Y. Huang and M. Zhou, “Low-Quality Product Review Detection in Opinion Summarization “, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, **(2007)**, pp. 334-342.

[7] R. Feldman, M. Fresko, J. Goldenberg, O. Netzer and L. Ungar, “Using Text Mining to Analyze User Forums”, International Conference on Service Systems and Service Management, **(2008)**, pp. 1-5.

[8] H. Cui, V. Mittal and M. Datar, “Comparative Experiments on Sentiment Classification for Online Product Reviews” , Proceedings of the 21st National Conference on Artificial Intelligence, vol. 2, **(2006)**, pp. 1265-1270.

[9] J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, “Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques”, Third IEEE International Conference on Data Mining (ICDM'03), **(2003)**, pp. 427.

[10] P. Melville, W. Gryc and R. D. Lawrence, “Sentiment analysis of blogs by combining lexical knowledge with text classification”, International Conference on Knowledge Discovery and Data Mining, **(2009)**, pp. 1275-1284.

[11] R. Prabowo and M. Thelwall, “Sentiment Analysis: A Combined Approach”, Journal of Informetrics, vol. 3, no. 2, **(2009)**, pp. 143-157.

[12] T.-L. Wong and W. Lam, “Learning to extract and summarize hot item features from multiple auction web sites”, Knowledge and Information Systems, vol. 14, no. 2, **(2008)**, pp. 143-160.

Authors

Periakaruppan Sudhakaran was born in 1979, Trichy, Tamilnadu, India. He completed his B.E. (comp. science) at Shri Angalamman College of Engineering and Technology, Bharathidasan University, Tiruchirappalli. He awarded M.E. (Computer Science) degree from JJ College of Engineering and Technology, Anna University, Tiruchirappalli, Tamilnadu. He has nearly 13 years of experience in Teaching both UG and PG program. He is presently working as an Assistant. Professor Grade- III & Head in Department of Information Technology in Oxford Engineering College, Tiruchirappalli. So far nearly 10 P.G students are guided by his guidance and also guided many project and he attended totally more than 30 conferences, workshop and seminars.

Dr. S. Hariharan received his B.E degree specialized in Computer Science and Engineering from Madurai Kammaraj University, Madurai, India in 2002, M.E degree specialized in the field of Computer Science and Engineering from Anna University, Chennai, India in 2004. He holds his Ph.D degree in the area of Information Retrieval from Anna University, Chennai, India. He is a member of IAENG, IACSIT, ISTE, CSTA and has 9 years of experience in teaching. Currently he is working as Associate Professor in Department of Computer Science and Engineering, TRP Engineering College, India. His research interests include Information Retrieval, Data mining, Opinion Mining, Web mining. He has to his credit several papers in referred journals and conferences. He also serves as editorial board member and as program committee member for several international journals and conferences.

Professor Lu's extensive research covers XML technology, Information retrieval and knowledge engineering, Internet Computing, wire/wireless distributing systems and cloud computing across disciplines such as science and technology, education and environmental industry, etc. Except her regular journal, chapters and conference publications, she has published 5 academic books in the topics that her research activities devote. She has been the Founder and a Programme Chair for International XML Technology workshop, XMLTech since 2003 to 2011 in USA. She has been editor and associate editor for Internet computing conference from 2003-2011. She is the founder and Editor in chief for the International journal of information retrieval research, USA. She published academic books in the areas of spatial pattern mining, and information retrieval research from 2009 to 2013. She was principle investigator for three EU projects: Edumecca, DO-IT and DONE-IT (2008-2012: 143545-2008-LLP-NO-KA3-KA3MP and 511485-LLP-1-2010-NO-KA3-KA3MP) specialized in XML and web services, Internet, Smartphone and wireless response technologies with seven EU countries involving physics, history, math, languages, mechanical engineering, materials science.