Journal of Theoretical and Applied Information Technology
20th May 2014. Vol. 63 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPINION MINING: APPROACHES, RESOURCES AND CHALLENGES

[1] MAQBOOL AL-MAIMANI , [2] NAOMIE SALIM, [3] AHMED M. AL-NAAMANY

[1] IT Department, Oman Air, Muscat, Sultanate of Oman
[2] Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
[3] Modern College of Business and Science, Muscat, Sultanate of Oman
E-mail: [1]maqbool.almaimani@omanair.com , [2]naomie@utm.my , [3]naamany@mcbs.edu.om

ABSTRACT

Web2.0 has contributed tremendously towards the rapid growth of web contents which ensures that customers utilize existing blogs, discussion forums, e-commerce and many other sites to express their opinions and read reviews of other people on different products and services which they plan to procure. Such an online wealth of information over the web has helped customers, firms, manufacturers, service providers, social and government units to take proper decision to procure or enhance various products and services. This practice has triggered the need to enhance existing methods and techniques to extract and summarize opinions from different online reviews. This paper surveys recent and leading methods and techniques that are used for opinion mining and then outlines available resources which have been developed in this regard. The paper also presents few challenges and open issues that need to be addressed and researched in more depth in order to improve the way opinions are extracted and summarized to users and other interested groups.

Keywords: *Mining; Sentiment Analysis; NLP; Objective; Subjective; Polarity; Items; Features.*

1. BACKGROUND

In the era of Web 2.0, more and more people like to express their views and read online user reviews before they really buy a product / service. The web offers many sites (like blogs, discussion forums, e-commerce) to enable people to access available on line reviews to take proper decisions. In 2007 and 2008, statistics showed that 81% of Internet users did online research on a product at least once and between 73% and 87% report that reviews had a significant influence on their purchase [1][2]. It is estimated that 75,000 new blogs emerge daily with 1.2 million new posts each day covering many consumer opinions on products and services. Finding and extracting such opinions is very essential for various reasons [1][2]:

- To understand customers' feelings and opinions on a particular product/services in order improve the quality and delivery of such goods/services
- To scientifically record different opinions and positions of people on a specific event, accident, incident, occasion etc. This covers areas like economical changes, history, scientific explora--tions and many other day-to-day issues in order to take proper measures and required improvements.
- To improve social services provided to public by governments and social organizations by understanding people's demands and suggestions.

The paper consists of four main sections. Section II presents few key concepts used in Opinion Mining field. Section III discusses few leading approaches and techniques used in extracting and summarizing the opinions. Opinion Mining available resources are outlined in section IV. Challenges and few open issues of Opinion Mining are briefly analyzed in section V.

2. KEY CONCEPTS

Opinion mining or sentiment analysis is a process of a where computer tools and techniques are to search for results on a given item / service, generate a list of product attributes

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(quality, features, etc.) and aggregate opinions about each of them (poor, mixed, good) [2]. This section highlights few important concepts related Opinion Mining (OM).

- Objective and subjective information: Facts are the objective information representing truth details about different things (i.e science, geography, history, political science, space world etc). These types of information are retrieved using normal search engines like Google. Subjective information, on the other hand is people's opinions and thoughts which they express about a particular product, service, event, situation, incident etc. This type of information is difficult to find and extract as one word in one sentence can mean something while it means something else in another. For example, the sentence "This laptop has long battery life" is different than "This lecture is boring and it takes long time to finish."
- Opinions: an Opinion is a disputed view or stand of someone about something. Opinions can be of three types: 1) Explicit opinions which are direct opinions that are clearly expressed (e.g. "this food is delicious"). 2) Implicit opinions are opinions which can be implied from text (e.g. "The camera stopped working in two days"). 3) Emotional Opinions are related to people's emotions like happiness, sadness, humor, and anger etc. (e.g. I will never see movies of this director again.").
- Opinion Polarity: This refers to the direction of the opinion or subjective information and can be Positive, Negative or Neutral. Words like beautiful, wonderful, good and amazing are positive; whereas, words like bad, poor, terrible and loss are negative ones.
- Opinion mining process: Given an object and a collection of reviews on it, the task in opinion mining process usually consists in general view of the following tasks: 1) Identify and extract object features that have been commented on in each review, 2) Making hierarchy of features, 3) Grouping synonyms of features, 4) Sentiment Analysis: Determining the orientation of opinion is positive, negative, neutral, 5) Provide summary of opinion in textual or in a visualization way.
- Blogosphere: Blogosphere is the name associated to universe of all the blog sites. A blog is a website that allows people to write various topics and express their opinions about different products and services [2].

3. OM APPROACHES AND TECHNIQUES

The mining process can be as simple as learning polarity (positive or negative) and sentiment of the words, or as complicated as performing deep parsing of data to identify grammar and structure of the sentences. Approaches and techniques for opinion mining and sentiment analysis can be classified into different ways. One classification is based on manual and automated techniques. In the manual-based approaches dictionaries and lexicon based approaches are implemented. On the dictionary-based methods, resources like WordNet are used to find opinions from words Synsets and hierarchies. Seeds are used to search for synonyms and antonyms in WordNet [1][3]. The manual approach tends to find sentences, phrases, words and patterns that express subjectivity and the orientation of the opinionated text. Words like "honest, important, mature, large, patient" are used to express positive opinions; whereas words like "harmful, hypocritical, inefficient, insecure" are used to express negative ones. Although the dictionary-based approach will help to find many of such words, but it does not help to find context dependent opinion words like long, big, fat, small etc. In the lexicon approaches, rule-based methods are researched and implemented using corpus and various available datasets. Such approaches depend on syntactic rules and co-occurrence patterns to be extracted from large corpora. These approaches are used to find domain dependant opinions. However, many improvements are still required to advance such rule-based approaches [4]. The other category of OM techniques are machine-based methods. Studies showed that standard machine learning techniques outperform human-based approaches. In the machine-based approaches, systems are built and trained to categorize opinionated text into positive, negative or neutral opinions using supervised and unsupervised learning techniques [2].

In a different classification, Yee et. al, presented OM and sentiment analysis techniques by

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

opinion mining components. The techniques are listed for different components of OM, covering Items extraction methods, Features extraction methods, Sentiment Classification techniques, Strength of sentiments and Summarization of opinion techniques. We believe the above is more structured approach for presenting various methods and techniques used for opinion mining and hence it is followed in this paper.

Item is a subject matter on which the opinion is expressed. Kobayashi [5] and Gamon [6] proposed two frameworks for item extractions which show significant improvements over other methods used for item extractions. Kobayashi and his team researched on a method of opinion extraction which was based on a structured form. In their study they concentrated on structuring the opinions of the customers in an effective way especially in connection to web documents where the main focus was on extraction of the subject/aspect evaluation relations, and extracting subject/aspect-aspect relations, using a machine learning-based method, which is portable across domains. Their study addressed the main area of opinion extraction where they combined the contextual clues and the context independent statistical clues with the help of a machine-learning technique ('boosting-based algorithm'). Experiments were carried out and evaluation was conducted using 5 fold cross validation on all data in the aspects of recall and precision [5]. The developed algorithm had its own limitations and could not be used especially when it comes to clustering techniques and machine-learned sentiment classifier.

Feature extraction is a process that can be performed after item extraction. This procedure involves recognizing the features of products that clients have indicated their opinions on through evaluations and comments. Consider a digital camera as an example. The camera has numerous features, as highlighted in [26]; these features include image quality, battery life, move, dimension, and weight. For example, a digital camera with poor image quality may have an extremely long battery life or may be very light.

Feature sentiment can refer to an opinion on a certain item based on its features. After features are recognized, a feature sentiment can be indicated for each feature, thus providing information regarding the strengths and weaknesses of the features of an item, such as battery life, dimensions, and colors.

Under this platform, item features are believed to be provided or determined before an opinion is identified as beneficial or damaging. Previous studies suggested using sentiment analysis methods to categorize web forum opinions in several languages. Feature sentiment approach has two important steps: removing an original group of features and executing selected features. These measures are employed to perform sentiment categorization of newsgroup communications. The test creates an effect similar to that of standard film review datasets. The procedure targets file amount categorization of sentiments.

Item Sentiment term denotes the total sentiments being expressed about an object. For example, a camera has positive and negative suggestions from customers online. This item is extremely useful when a prevalent opinion must be recognized immediately. The majority of literature is focused on discovering merchandise sentiment. Scientists demonstrate significant interest on this topic. In particular, Turney [32] offered a commonly employed paradigm that provides a foundation to remove the opinion on an item. The conditions employed for the evaluation are summarized and may be one of three organizations, namely, positive, negative, or neutral. Most studies have focused on evaluating the subjectivity of negative and positive terms to determine evaluation sentiments while overlooking neutral terms. Nevertheless, other scientists have asserted that impartial (or objective) conditions should be considered because such conditions may enhance the precision of outcomes.

Feature comparison represents better granularity of an opinion between two different entities, such as AB cameras vs. XY cameras. AB cameras generally have more favorable evaluations than XY Cameras. This factor is necessary for shoppers who may have enough time to search for the best cost and merchandise to purchase. Consumers can save much time in decision-making processes with regard to everyday concerns, thus creating a dual economy involving cash and time. Grams, Wang, and Araki [30] practiced a book method that graphically shows comparisons. Several improvements that occurred in this field include

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

graphically exhibiting each feature of an item. This improvement obviously enhances purchasing experience by exhibiting the pros and cons of each feature. Moreover, manufacturer can readily determine features that do not satisfy customer expectations and demands [30][31].

Summarization Techniques

Summarization is the final and a very important stage of the opinion mining process. Summarization is required because one opinion does not represent all opinions. There are many systems used to summarize the extracted and mined opinion. FBS and OPINS are two excellent product review summarization system [11][17]. In addition, there are many techniques and approaches which were developed to summarize customer reviews and opinions and present them in textual and graphical formats to enable customers and supplier take proper actions.

Summarization technique was proposed in the year 2009 by Hu and Wu [18] especially to summarize the various reviews given by customers for specific group of products. This summarization technique was based on a set of lists which indicated the pros and cons which could be easily understood just by having a glance over the lists. The main intention of the researcher was to classify these pros and cons into a list. In previous studies which typically make use of classical summarization techniques where phrases were used as a primary component during the whole summarization process. The main reason to use such technique was to help the readers to understand, remember and recognize the opinions. Apart from that the summarization technique will help readers and also the manufacturers to gather information of the main flaws and improvement suggestions related to their products which could be utilized in the future.

With regard to the classification process, one should make use of word weightage technique to calculate the strength of each word towards both opinions and word score in order to show the word strength of expressing the sentiments accordingly. Chi-square analysis technique can be deployed in order to calculate the correlation among the terms used in the review.

Apart from the chi-square analysis one may also use frequency statistical analysis value for the word-stem in order to weigh the strength of each word towards the sentiment in positive orientation and negative orientation. Using stem but not the word itself is to make the result more precise and comprehensive. The Word Weight is not enough for describing the strength of a word's orientation, since the word's linguistic type itself plays an important role in expressing the sentiment [2][4]. Other research found that adverb and adjective are the core types to empress human sentiment although verb can show some polarized intention [20]. Based on these findings, they used a score algorithm to combine the linguistic feature of a word to its weight value.

A significant component of feature-based strategies is the ability to distinguish features of various services and products. As opposed to the related function presented in Hu et al. 26], Dork, Lawrence, and Pennock [28] completed a significant improvement in the sentiment categorization of merchandise reviews in a document. The objective of their research was to categorize each evaluation document as indicating a favorable or unfavorable sentiment regarding a particular item.

Mishne [27] required a completely distinctive strategy, which became a pioneer on OM books. The study of Mishne was more concentrated on categorizing blog articles according to diverse emotions. Simply stated, these studies did not address the task of feature extraction, but continue using omitted features to help classify blogs by emotions. These researchers considered disposition classification to be beneficial to numerous applications, including enhancing physician–patient interaction and helping behavioral scientists. Their objective was determining the probable state of mind of the reviewer once a post was composed by using a developed learning method to recognize several features that would be applied to the learning procedure. Ding et al. [25] discussed the issue of identifying the semantics of opinions on item features in client reviews as opposed to on the goods (things) mentioned in the reviews such as in the studies of Esuli et al. [29]. The objective of the work of Ding was using language rules.

4. BUILDING DATASETS, CORPUS AND DICTIONARIES FOR OPINION MINING

Corpus is the plural form for corpora which is basically a collection of linguistic data in an electronic format (in a Computer-readable text).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

There are many efforts made by different researches to build or refine corpus for Opinion Mining. The following are few corpus and Data sets in an Alphabetical order[2]:

- Cornell-movie-review-datasets: http://www.cs.cornell.edu/people/pabo/movie-review-data/. This dataset contains [2]: 1) document-level: polarity dataset v2.0: 1000 positive and 1000 negative processed reviews 2)sentence-level: sentence polarity dataset v1.0: 5331 positive and 5331 negative processed sentences/snippets. 3) Subjectivity dataset v1.0: 5000 subjective and 5000 objective processed sentences.
- Customer-review-datasets: http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip. "This dataset consists of reviews of five electronics products downloaded from Amazon and Cnet. The sentences have been manually labeled as to whether an opinion is expressed, and if so, what feature from a pre-defined list is being evaluated "[20].
- Multiple-aspect-restaurant-reviews: http://people.csail.mit.edu/bsnyder/naacl07. "The corpus, introduced in Snyder and Barzilay, consists of 4,488 reviews, both in raw-text and in feature-vector form. Each review gives an explicit 1-to-5 rating for five different aspects—food, ambiance, service, value, and overall experience "[2]
- NTCIR multilingual corpus: "The corpus for the NTCIR 6 pilot task consists of news articles in Japanese, Chinese, and English and formed the basis of the Opinion Analysis Task at NTCIR6. The training data contains annotations regarding opinion holders, the opinions held by opinion holder, and sentiment polarity, as well as relevance information for a set of predetermined topics. The corpus of the NTCIR Multilingual Opinion-Analysis Task (MOAT) is drawn from Japanese, Chinese, and English blogs" [2].

5. CHALLENGES

Until today the field of opinion mining is not well developed to provide user with a powerful opinion and sentiment mining systems. This section lists few challenges and open issues that need to be addressed and researched in depth [2]

- The fact that product reviews are written in different languages has created a challenge in opinion mining. The main problem becomes the time spent in reviewing all available data and resolves the language barrier. We need a language independent method that automatically analyze, extract and assign values for a given product or service [16].
- There is a need to develop powerful techniques in mining emotional related opinions like happiness, sadness, humor, anger etc
- OM experiences numerous challenges, including determining which section of a text is an opinion, pinpointing the opinion writer, determining the beneficial or unfavorable power of an opinion, and so on.
- Phrase file intricacy, contextual emotions, heterogeneous files, benchmark quality, and modal workers remain difficult issues in this field.
- There is a need to develop opinion mining search engines which extracts subjective details from different reviews of consumers. Furuse came up with a search engine which extracts opinion sentences based on open-domain query from Japanese blog pages [22]. But this is just very initial efforts. Existing search engines are for fact searching. How such a search engine would be developed? Can we search for opinions as conveniently as general Web search? How the opinion search queries should be formed? Finding the opinion of a person or organization (opinion holder) on a particular object or a feature of an object is yet a challenge [2].

6. SUMMARY

Opinions are essential to any person who is likely to make a choice. OM is effective for individuals who need to purchase an item and are able to choose which item to purchase by studying opinions instead of lengthy product reviews and producing outline their faces. OM is equally essential to businesses by helping them understand how clients regard their merchandise. Consequently, businesses may make decisions regarding their products based on the opinions of clients. Companies may also modify their

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

merchandise in accordance with the opinions of clients through an efficient and rapid method. Therefore, businesses may develop better customer relationships by providing their requests and satisfying their demands. Businesses can find, attract, and maintain clients, thus possibly saving on manufacturing costs by adopting insights from consumer demands.

To date, excellent advancements have been achieved in this field, and the challenges presented in this paper have been addressed by numerous strategies. Approaches that can handle the challenges of OM simultaneously are in demand.

Moreover, the paper highlighted few important resources and corpuses that are built for data mining. The paper also presented many opening areas that need to be addressed and researched to enhance this field.

REFERENCES:

[1] Esuli A. & Sebastiani F., "Determining the Semantic Orientation of Terms through Gloss Classification", in Proceedings of 14th ACM International Conference on Information and Knowledge Management, CIKM 05, Bremen, DE, pp. 617-624, 2005.

[2] Pang B. and Lee L., "Using very simple statistics for review search: An exploration," in Proceedings of the International Conference on Computational Linguistics (COLING), (Poster paper), 2008.

[3] Hu, M. & Liu, B. (2004), 'Mining and Summarizing Customer Reviews', in Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining, KDD-04, Seattle, WA, pp. 168-177, 2004.

[4] Kanayama H. and Nasukawa T.,"Fully automatic lexicon expansion for domain-oriented sentiment analysis," EMNLP-06, 2006.

[5] Kobayashi N., Inui, K. & Matsumoto, Y., 'Opinion Mining from Web Documents: Extraction and Structurization', in Proceedings of the Transactions of the Japanese Society for Artificial Intelligence 22, JSAI07, pp. 326-337, 2007.

[6] Gamon, M., Aue, A., Corston-Oliver, S. & Ringger, E.,"Pulse: Mining Customer Opinions from Free Text", in Proceedings of the Natural Language Processing, Microsoft Research, IDA, Redmond, WA, pp. 121-132, 2005.

[7] Turney P., "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," Procs. of the 40th Annual Meeting of the Association for Computational Linguistics, 2002.

[8] Kobayashi N., Iida, R., Inui, K. & Matsumoto, Y. "Opinion Mining as Extraction of Attribute-Value Relations", in Proceedings of the Nara Instutute of Science and Technology, JSAI -2005, Takayama, Ikoma, Japan, pp. 470-481, 2005.

[9] Mishne G.,"Experiments with Mood Classification in Blog Posts", in Proceedings of the Stylistic Analysis of Text for Information Access, Style, Amsterdam, The Netherlands, 2005.

[10] Liu, B., Hu, M. & Cheng, J., "Opinion Observer: Analyzing and Comparing Opinions on the Web", in Proceedings of the International World Wide Web Conference Committee, WWW05, Chiba, Japan, 2005.

[11] Sheng Feng, Ming Zhang, Yanxing Zhang, Zhihong Deng, "Recommended or Not Recommended? Review Classification through Opinion Extraction", 12th International Asia-Pacific Web Conference, pg: 350-352, 2010.

[12] Matthew Whitehead, Larry Yaeger, "Building a General Purpose Cross-Domain Sentiment Mining Model", World Congress on Computer Science and Information Engineering, pg: 472-476, 2009.

[13] Pang B. and Lee L., "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in ACL,pp. 271–278, 2004.

[14] Ding X. & Liu, B., "The Utility of Linguistic Rules in Opinion Mining", in Proceedings of the SIGIR 2007, SIGIR, Amsterdam, The Netherlands, 2007.

[15] Liu B., "Web data mining; Exploring hyperlinks, contents, and usage data," Opinion Mining. Springer,2006.

[16] Alexandra BALAHUR, Andrés MONTOYO, "A Feature Dependent Method for Opinion Mining and Classification", 978-1-4244-2780-2/08, IEEE, 2008

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[17] Zhuang L., Jing F., Zhu X., and Zhang L., "Movie review mining and summarization," in CIKM, pp. 43–50, 2006.

[18] Xinghua Hu, Bin Wu, "Classification and Summarization of Pros and Cons for Customer Reviews", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops, 2009.

[19] Hu. X. and Wu B., "Automatic keyword extraction using linguistic features," Proceedings of the 6th annual international IEEE conference on Data Mining, 2006.

[20] Hatzivassiloglou V. and McKeown K.,"Predicting the semantic orientation of adjectives," ACL-1997.

[21] Mohsen Jafari Asbagh, Mohsen Sayyadi, and Hassan Abolhassani,"Blog Summarization for Blog Mining", Software Engineering, Artificial Intelligence, SCI 209, pp. 157–167, Springer-Verlag Berlin Heidelberg, Germany, 2009.

[22] O. Furuse, N. Hiroshima, S. Yamada and R. Kataoka, "Opinion sentence search engine on open-domain blog," IJCAI, 2007, 2760□2765.

[23] Kobayashi, N., Inui, K. & Matsumoto, Y., 'Opinion Mining from Web Documents: Extraction and Structurization', in Proceedings Of the Transactions of the Japanese Society for Artificial Intelligence 22, JSAI07, 2007, pp. 326-337.

[24] G. Wang and K. Araki, "An Unsupervised Opinion Mining Approach for Japanese Weblog Reputation Information Using an Improved SO-PMI Algorithm,"
IEICE TRANS. INF. & SYST, vol. VOL.E91–D, pp. 1032-1041, 2008.

[25] Ding, X. & Liu, B., 'The Utility of Linguistic Rules in Opinion Mining', in Proceedings of the SIGIR 2007, SIGIR, 2007, Amsterdam, The Netherlands.

[26] Hu, M. & Liu, B., 'Mining and Summarizing Customer Reviews', In Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining, 2004, KDD-04, Seattle, WA, pp. 168- 177.

[27] Mishne, G., 'Experiments with Mood Classification in Blog Posts', in Proceedings of the Stylistic Analysis of Text for Information Access, Style, 2005, Amsterdam, The Netherlands

[28] Dave, K., Lawrence, S. & Pennock, D.M., 'Mining the Peanut Gallery: Opinion extraction and Semantic Classification of Product reviews', in Proceedings of
the 13th International WorldWide Web Conference, 2003, WWW03.

[29] Esuli, A. & Sebastiani, F., 'Determining Term Subjectivity and Term Orientation for Opinion Mining', in Proceedings of the ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, CL-06, 2006, Madrid, ES, pp. 174- 181

[30] G. Wang and K. Araki, "An Unsupervised Opinion Mining Approach for Japanese Weblog Reputation Information Using an Improved SO-PMI Algorithm,"
IEICE TRANS. INF. & SYST, vol. VOL.E91–D, pp. 1032-1041, 2008

[31] Liu, B., Hu, M. & Cheng, J., 'Opinion Observer: Analyzing and Comparing Opinions on the Web', in Proceedings of the International World Wide Web Conference Committee,WWW05, 2005, Chiba, Japan

[32] Turney P, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.," presented at In Proc. of the Meeting of the Association for ComputationalLinguistics (ACL'02), 2002.

[33] Haji Binali, Vidyasagar Potdar, Chen Wu, "A State Of The Art Opinion Mining And Its Application Domains", Digital Ecosystems and Business Intelligence Institute, 2008.

[34] Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Trans. Inform. Syst. 26(3), Article 12, 34 pages.