Opinion Mining and Sentimental Analysis Approaches: A Survey

Mahmoud Othman[1], Hesham Hassan[2], Ramadan Moawad[1] and Abeer El-Korany[2]

[1]Department of Computer Science, Faculty of Computers and Information Technology, Future University, Cairo, Egypt
[2]Department of Computer Science, Faculty of Computers and Information, Cairo University, Cairo, Egypt
msamy@fue.edu.eg

Abstract: The automatic extraction of information from unstructured sources has opened up new ways for querying, organizing, and analyzing data by building a clean semantics of structured databases from a huge number of unstructured data and the society became more data oriented with easy online access to both structured and unstructured data. New applications of structured extraction came around such as the paper topic opinion mining, which is a type of natural language processing for tracking the mood of the public about a particular topic. Opinion mining, which is also called sentiment analysis, involves building a system to collect and examine opinions about the product or topic made in blog posts, comments, reviews or tweets. Automated opinion mining often uses machine learning, which is a component of artificial intelligence (AI).

[Mahmoud Othman, Hesham Hassan, Ramadan Moawad and Abeer El-Korany. **Opinion Mining and Sentimental Analysis Approaches: A Survey.** *Life Sci J* 2014;11(4):321-326]. (ISSN:1097-8135). http://www.lifesciencesite.com. 44

Key Words: Opinion Mining, Sentiment Analysis, Subjectivity analysis.

1. Introduction

The internet presents a huge amount of useful information which is usually formatted for its users, which makes it difficult to extract relevant data from various sources. Therefore, the availability of robust, flexible Information Extraction (IE) systems that transform the web pages into program-friendly structures such as a relational database will become a great necessity, and whenever we need to make a decision, we often seek out the opinions of others. Individuals can get opinions from friends and family and organizations use surveys, focus groups, opinion polls and consultants.

Having an access to large quantities of data through internet and its transformation into a social web is no longer an issue, as there are terabytes of new information produced on the web every day that are available to any individual. Even more importantly, it has changed the way we share information. The receivers of the information do not only consume the available content on web, but in turn, actively annotate this content and generate new pieces of information.

Also, today people not only comment on the existing information, bookmark pages and provide ratings but they also share their ideas, news and knowledge with the community at large. In this way, the entire community becomes a writer, in addition to being a reader. The existing mediums like blogs, wikis, forums and social networks where users can post information, give opinions and get feedback from other users on different topics, ranging from politics and health to product reviews and travelling. The increasing popularity of personal publishing services of different kinds suggests that opinionated information will become an important aspect of the textual data on the web [18, 19]. Recently, many researchers have focused on this area. They are trying to fetch opinion information to analyze and summarize the opinions expressed automatically with computers.

This new research domain is usually called Opinion Mining and Sentiment Analysis. Until now, researchers have evolved several techniques to the solution of the problem. Current-day Opinion Mining and Sentiment Analysis is a field of study at the crossroad of Information Retrieval (IR) and Natural Language Processing (NLP) and share some characteristics with other disciplines such as text mining and Information Extraction.

This paper surveys the major opinion mining approaches and compares them in three dimensions: the dataset used, the techniques used, and the system domain.

The following sections are basic definitions related to the paper topic, then the approaches of opinion mining followed by the challenges of opinion mining, the related works and significance of the opinion mining systems then the evaluation method and finally our conclusion.

2. Basic Definitions

2.1. Opinion Definition

Opinion is an emotion about an **Entity** or an **Aspect** of the entity from an **Opinion Holder** [31]. The entity may be a product, person, event, organization, or topic, and can be a hierarchy of components, sub-components, and so on; each node represents a component and is associated with a set of attributes of the component. For example in figure 1 the entity is apple organization and the components are iPhone5, battery and camera but the attributes are the size and the battery life attached to the battery and the quality attached to the camera. An opinion can be expressed on any node or attribute of the node. For simplicity, the researcher use the term **aspects** (features) to represent both components and attributes.

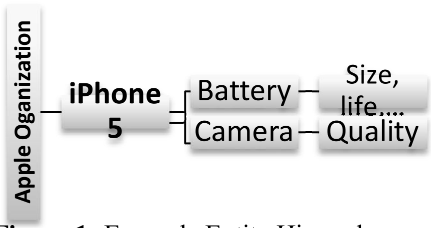


Figure 1: Example Entity Hierarchy

There are two main types of opinion [31]; the first one is regular opinions, which the Sentiment/opinion expressed on some target entities e.g. "*The touch screen is really is good*". The second one is comparative opinions which the opinion expressed as comparison of more than one entity. e.g. "*iPhone is better than Blackberry*".

2.1.1. Opinion Representation

An opinion can be represented as a quintuple (Ej, Ajk, SOijkl, Hi, Tl) where [31]; Ej is a target entity which can be a product, service, individual, organization, event, or topic, Ajk is an aspect/feature of the entity Ej. An object usually has two types of attributes; components, e.g. "battery, keypad/touch screen" and properties, e.g. "size, weight, color, voice quality", SOijkl is the sentiment value of the opinion from the opinion holder Hi on feature Ajk of entity Ej at time Tl, it also called opinion orientation (polarity) which can be positive, negative, or neutral, Hi is an opinion holder who expresses the opinion and Tl is the time when the opinion is expressed.

2.2. Opinion Mining Definition

Opinion mining [31], is a type of natural language processing for tracking the mood of the public about a particular topic. Opinion mining, which is also called sentiment analysis, involves building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets.

2.3. Subjectivity Analysis

Subjectivity analysis involves various methods and techniques that originate from information retrieval, artificial intelligence and natural language processing. This confluence of different approaches is explained by the nature of the data being processed and application requirements. Moreover, opinion mining originates from the information retrieval community, and aims at extracting and further processing users' opinions about products, movies or other entities as we defined in previous sections. Sentiment analysis, on the other hand, was initially formulated as the natural language processing task of retrieval of sentiments expressed in texts [19, 27]. However these two problems are similar in their own essence and fall under the scope of subjectivity analysis.

2.4. Sentimental Analysis Levels

There are several levels of sentimental analysis such as [31,8] ; **Document-level** which identify if the document (e.g. product reviews, blogs, and forum posts) expresses opinions and whether the opinions are positive, negative, or neutral; **Sentence-level** which identify if a sentence is opinionated and whether the opinion is positive, negative, or neutral, and **Attribute-level** which extract the object attributes (e.g. image quality, zoom size) that is a subject of an opinion and the opinion orientations.

3. Opinion Mining Approaches

3.1. Machine Learning Approaches

In general, sentiment analysis is concerned with analyzing direction based text, determining whether a text is objective or subjective and whether a subjective text contains positive or negative sentiments is a common two-class problem that involves classifying sentiments as positive or negative. Additional variations include classifying sentiments as opinionated/subjective or factual/ objective. Some studies have attempted to classify emotions (such as happiness, sadness, anger, or horror) instead of sentiments. The machine-learning approach [3], treats the sentiment-classification problem as a topic-based text classification problem. Any text classification algorithm can be employed, such as Naïve Bayes or support vector machines (SVMs)

3.2. Lexicon Based Approach

The Lexicon based approach performs classification based on positive and negative sentiment words and phrases contained in each evaluation text and mining the data requires no prior training. Two types of techniques have been used in previous semantic orientation approach based sentiment classification research: corpus-based and dictionary-based.

3.2.1. Corpus-based Approach

The corpus-based approach aims to find co-occurrence patterns of words to determine their sentiments. Researchers have proposed different strategies to determine sentiments; for example, Peter

Turney [26], calculated a phrase's semantic orientation to be the mutual information between the phrase and the word "excellent" (as the positive polarity) minus the mutual information between the phrase and the word "poor" (as the negative polarity). Ellen Riloff and Janyce Wiebe [7], used a bootstrapping process to learn linguistically rich patterns of subjective expressions to distinguish subjective expressions from objective expressions.

3.2.2. Dictionary-based Approach

Use synonyms, antonyms, and hierarchies in WordNet (or other lexicons with sentiment information) to determine word sentiments [2]. Building upon WordNet, SentiWordNet is a lexical resource for sentiment analysis that has more sentiment-related features. It assigns to each synset of WordNet three sentiment scores regarding positivity, negativity, and objectivity, respectively. SentiWordNet has been used as the lexicon in recent sentiment classification studies.

The corpus-based techniques, however, often rely on a large corpus to calculate the statistical information needed to decide the sentiment orientation for each word or phrase. Therefore, they might not be as efficient as the dictionary-based techniques. Still, a good lexicon is critical for the dictionary-based techniques [9].

3.2.3. Example Lexicon APIs

There are several lexicon APIs such as, the general inquirer lexicon which has 1915 positive words and 2291 negative word [13], LIWC (Linguistic Inquiry and Word Count) which contain 2300 word with more than 70 class for example bad, weird, hate, problem and tough as negative emotion and love, nice and sweet as positive emotion [12], also opinion lexicon contains 2006 positive word and 4783 negative word [11] , finally SentiWordNe which all WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality [10].

4. Opinion Mining Challenges

An opinion mining system is often built using software that is capable of extracting knowledge from examples in a database and incorporating new data to improve performance over time. The process can be as simple as learning a list of positive and negative words, or as complicated as conducting deep parsing of the data in order to understand the grammar and sentence structure used.

There are several challenges in opinion mining. The first is that a word that is considered to be positive in one situation may be considered negative in another situation. Take the word "long" for instance. If a customer said a laptop's battery life was long, that would be a positive opinion. If the customer said that the laptop's start-up time was long, however, that would be is a negative opinion. These differences mean that an opinion system trained to gather opinions on one type of product or product feature may not perform very well on another.

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. In opinion mining, however, "عظيم كان الاستفتاء" is very different from "عظيم يكن لم الاستفتاء".

A Third challenge is that people can be contradictory in their statements. Most reviews will have both positive and negative comments, which is somewhat manageable by analyzing sentences one at a time. However, the more informal the medium (twitter or blogs for example), the more likely people are to combine different opinions in the same sentence. 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 to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context. For example, "That movie was as good as his last one" is entirely dependent on what the person expressing the opinion thought of the previous film.

Finally, people can express their opinion in many languages such as Arabic; English….etc. so it is difficult to the computer parses all sentences belonging to the same topic with the same application considering the following example:

Entity: iPhone 5

User 1: iphone5 is the most likely phone to buy

User 2: هذا الجهاز امكاناتة جميلة جدا

User 3: 3na ba7eb 7za el mobile

5. Pervious Works

In this paper we classify the related works by the approaches used and comparing them in three dimensions: the dataset used, the techniques used and its domain. Some of the previous works use machine learning approach and implement the algorithm based on different techniques. Some of them use Support Vector Machine (SVM) [30, 16, 21, 22] and some use Naïve Bayes [32, 1, 4, 22, 14, 23, 24]. Also some of them use Maximum Entropy [1, 4], and some use K-nearest Neighbor [22, 23]. Go and *et al.* [1] use both

Naïve Bayes and K-nearest Neighbor and compare the result between them. Also Samsudian [22] and *et al.* use K-nearest Neighbor and SVM and compare the result. Another approach was lexicon based approach, and as we discussed before in the previous sections that it may be corpus-based or dictionary based. Sobkowicz and *et al.* [25], use corpus based technique and the domain was politics, they use tweets as a dataset. Hamouda and *et al.* [2], use dictionary based technique. Also Fei and *et al.* [9] and Dang and *et al.* [30] use dictionary based as a technique.

Also some of previous researches merge between the two approaches, such as Yan Dang and *et al.* [30] , use SVM as a technique and merge it with lexicon based approach , they use SVM as the classifier and or each test, they randomly choose 90 percent of the reviews as training data and the remaining 10 percent as testing data. They also used 10-fold cross validation to conduct the evaluation. They use a product review as a data set and there general domain was marketing; in this paper we summarize all of them in the following table.

Approach	Technique	Paper, Year	dataset	Accuracy %	Domain
Machine Learning	Support Vector Machines (SVM)	[30],2010	reviews	82.7	Marketing
		[16],2013	reviews	No evaluation	General
		[21],2011	reviews	91.5	
		[22],2013	reviews	82.9	
	Naïve Bayes	[32],2011	reviews	84.5	Marketing
		[1],2009	tweets	82.7	General
		[4],2013	tweets	No evaluation	
		[22],2013	reviews	84.9	
		[14],2012	reviews	81.4	
		[23],2013	online messages	91.4	
		[24],2013	reviews	No evaluation	
	Maximum Entropy	[1],2009	tweets	83.0	
		[14],2008	reviews	77.1	
	K-nearest Neighbor	[22],2013	reviews	64.1	
		[23],2013	online messages	79.8	
Lexicons-Based	Dictionary-based	[2],2011	reviews	67-68.6	General
		[9],2012	tweets	81.2	General
		[30],2010	reviews	82.7	Marketing
	Corpus-based	[25],2010	reviews	76.8	Politics
		[7],2003	reviews	71% - 85%	General
		[26],2002	reviews	74.39	General

6. Significance of the Opinion Mining Systems

Opinion mining can be useful in several ways. If you are in marketing, for example, it can help you judge the success of an ad campaign or new product launch, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features. For example, a review might be broadly positive about a digital camera, but be specifically negative about how heavy it is 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 data is created by the customer.

Also opinion mining systems is very important for extracting a particular opinion about any hot topic belongs to the government such as: the referendum, elections and the extent of dissatisfaction with the performance of the government. In addition to, it could be used for user classification.

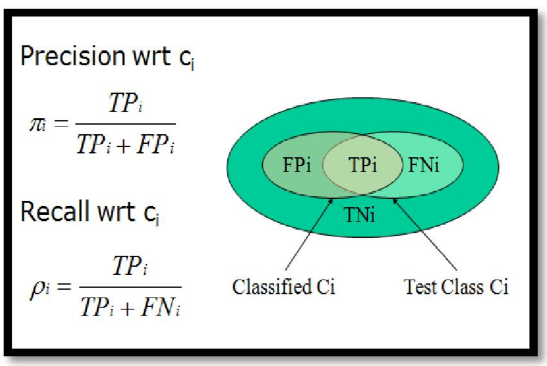
7. Evaluation Methods

7.1. Precision , Recall and F-Measure

The two most frequent and basic measures for information retrieval effectiveness are precision and recall [15, 29]. From these measures we can measure the accuracy or the retrieval quality which called F-Measure which calculated as the "*precision multiplied by the recall multiplied by two and the result of the multiplication is divided by the summation of the precision and the recall*".

After a classifier is constructed using a training set, the effectiveness is evaluated using a test set the following counts are computed for each category i:

- **TPi** true positives w.r.t. category ci which is the set of documents (opinions) that both the classifier and the previous judgments (as recorded in the test set) classify under ci.
- **FPi:** false positives w.r.t. category ci which is the set of documents (opinions) that the classifier classifies under ci, but the test set indicates that they do not belong to ci.
- **TNi:** true negatives w.r.t. ci which both the classifier and the test set agree that the documents (opinions) in TNi do not belong to ci.
- **FNi:** false negatives w.r.t. ci which the classifier do not classify the documents (opinions) in FNi under ci, but the test set indicates that they should be classified under ci



This approach [28,29] in evalation use the previous values to estimate the value of the precision and the recall, recall is a measure of the ability of the system to present all relevant items but the precision is a measure of the ability of the system to present only relevant items, the behind figure illustrate the how to compute precision and recall then we can calculate the F-Measure.

7.2. Area Under Curve (AUC)

Another method may be use is area under curve which is commonly used evaluation method for binary choice problems, which involve classifying an instance as either positive or negative [20,29]. Its main advantages over other evaluation methods, such as the simpler misclassification error, are:

- It's insensitive to unbalanced datasets (datasets that have more installed than not-installed or vice versa).
- For other evaluation methods, a user has to choose a cut-off point above which the target variable is part of the positive class (e.g. a logistic regression model returns any real number between 0 and 1 - the modeler might decide that predictions greater than 0.5 mean a positive class prediction while a prediction of less than 0.5 mean a negative class prediction). AUC evaluates entries at all cut-off points, giving better insight into how well the classifier is able to separate the two classes.

The true positive rate, or recall, is calculated as the number of true positives divided by the total number of positives. When identifying aircraft from radar signals, it is proportion that are correctly identified.

The false positive rate, is calculated as the number of false positives divided by the total number of negatives. When identifying aircraft from radar signals, it is the rate of false alarms.

Conclusion

This paper surveys the major opinion mining approaches and compares them in three dimensions: the dataset used, the techniques used and the system domain. The criteria of the first dimension explain which dataset used in the experiments. The criteria of the second dimension classify opinion mining systems based on the techniques used. The third dimension is the domain of opinion mining systems. In addtion to introduce and survey the approaches of sentiment analysis and opinion mining, we tried to showcase from basic definitions, different techniques, various evaluation methods, Finally, this paper concludes that all the sentiment analysis tasks are very challenging and it has been a very active research area in recent years.

References

1. Go A., R. Bhayani, L. Huang, "*Twitter sentiment classification using distant supervision*", CS224N Project Report , 2009
2. Hamouda A., M. Rohaim," *Reviews Classification Using SentiWordNet Lexicon*" , The Online Journal on Computer Science and Information Technology, Vol. 2, 2011
3. Jebaseel A., E. Kirubakaran, "*M-Learning Sentiment Analysis with Data Mining Techniques*", International Journal of Computer Science and Telecommunications, Vol. 3, 2012
4. Mahendran A., A. Duraiswamy, A. Reddy, C. Gonsalves, "*Opinion Mining for Text Classification*",

International Journal of Scientific Engineering and Technology, Vol. 2, 2013
5. Valarmathi B. , V. Palanisamy , "*Opinion Mining based on Recombined Parallel Neural Network*", European Journal of Scientific Research, Vol. 55, 2011
6. Consoli D., "*Analysing Customer Opinions with Text Mining Algorithms Computational Methock in Science and Engineering*", Advances in Computational Science Journal, Vol. 2, 2009
7. Riloff E. , J. Wiebe, "*Learning Extraction Patterns for Subjective Expressions*", in: proceeding of Empirical Methods on Natural Language Processing Conference , pg. 105-112,2003
8. Cruz F., J. Troyano, F. Ortega, C. Vallejom, "Long autonomy or long delay? The importance of domain in opinion mining", Elsevier : Expert Systems with Applications, Vol. 40, 2013
9. Fei G. .B. Liu, M. Hsu , M. Castellanos , R. Ghosh , *"A Dictionary-Based Approach to Identifying Aspects Implied by Adjectives for Opinion Mining*", Proceedings of COLING: pg. 309–318, 2012
10. Baccianella S., Esuli A., and Sebastiani F. ,*"An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*", http://sentiwordnet.isti.cnr.it/
11. Liu B. , "Sentiment Analysis and Opinion Mining : An Introduction", http://www.cs.uic.edu/~liub/
12. James W. Pennebaker, Roger J. Booth, and Martha E. Francis, "*Linguistic Inquiry and Word Count* , http://www.liwc.net/
13. Kimberly N. , " General Inquirer ", http://www.wjh.harvard.edu/~inquirer
14. Smeureanu I, C. Bucur, "*Applying Supervised Opinion Mining Techniques on Online User Reviews*", Informatica Economică, Vol. 16, 2012
15. Davis J., M. Goadrich, "*The Relationship between Precision-Recall and ROC Curves*", in: Proceedings of the 23rd International Conference on Machine Learning, pp.233-240, 2006
16. Khairnar J., M. Kinikar, "Machine *Learning Algorithms for Opinion Mining and Sentiment Classification*", International Journal of Scientific and Research Publications, Vol. 3, pp. 153-161, 2013
17. Cai K. , S. Spangler, Y. Chen, L. Zhang , "*a Leveraging Sentiment Analysis for Topic Detection*", Web Intelligence and Agent Systems: An International Journal, pp. 80-91, Vol. 8, 2010
18. Shimada K., T. Endo, "*Seeing Several Stars: a Rating Inference Task for a Document Containing Several Evaluation Criteria*", in: Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pg. 1006–1014., 2008
19. Wong K., Y. Xia, R. Xu, M. Wu, W. Li , "*Pattern-Based Opinion Mining for Stock Market Trend Prediction*", International Journal of Computer Processing of Languages, Vol. 21, 2008
20. Zuva K., T. Zuva, "*Evaluation Of Information Retrieval Systems*", International Journal of Computer Science & Information Technology, Vol. 4, pp. 113-120, 2012
21. Saleh M., M. Valdivia, A. Ráez, L. López, "*Experiments with SVM to Classify Opinions in Different Domains*", Expert Systems with Applications Journal, Vol. 38, pp. 90-105, 2011
22. Samsudin N., M. Puteh, A. Hamdan, M. Zakree , A. Nazri," *Immune Based Feature Selection for Opinion Mining*", in: Proceedings of the World Congress on Engineering, Vol. 3, pp. 120-127, 2013
23. Samsudin N., A. Hamdan, M. Puteh, M. Zakree, A. Nazri , "*Mining Opinion in Online Messages*", International Journal of Advanced Computer Science and Applications, Vol. 4, pp. 52-59, 2013
24. Kishore P., K. Naidu, "*Sentiment Analysis Using Semi-Supervised Naïve Bayes Classifier* ", International Journal of Innovative Technology and Research, Vol. 1, 2013
25. Sobkowicz P., M. Kaschesky, G. Bouchard, "*Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web*", Elsevier : Government Information Quarterly , Vol.29, 2012
26. Turney P., "*Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 417-424, 2002.
27. Abe S., K. Inui, K. Hara, H. Morita, C. Sao, M. Eguchi, A. Sumita, K. Murakami, S. Matsuyoshi, "*Mining personal experiences and opinions from Web document*" , Web Intelligence and Agent Systems: An International Journal, Vol. 9, 2011
28. Landgrebe T., "*Precision-recall operating characteristic (P-ROC) curves in imprecise environments*", in: The 18th International Conference on Pattern Recognition, pp. 123-127, 2006.
29. Mandl T., "*Recent Developments in the Evaluation of Information Retrieval System: Moving Towards Diversity and Practical Relevance*", Informatica, vol. 32, 2008.
30. Dang Y., Y. Zhang, H. Chen, "*A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews*", IEEE Intelligent Systems, Vol. 25, 2010
31. Lei Z., L. Bing, "*Sentiment Analysis and Opinion Mining: Introduction and Survey book*", publisher: Morgan & Claypool ,2011
32. Zhang Z., Q. Ye, Y. Li, "*Sentiment classification of Internet restaurant Reviews written in Cantonese*", Expert Systems with Applications, Vol. 31 , 2011

3/1/2014