A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation

Anaïs Collomb (anais.collomb@insa-lyon.fr)*
Crina Costea (crina.costea@insa-lyon.fr)*
Damien Joyeux (damien.joyeux@insa-lyon.fr)*
Omar Hasan (omar.hasan@insa-lyon.fr)*
Lionel Brunie (lionel.brunie@insa-lyon.fr)*

Abstract: The aim of this paper is to study and compare some of the methods used to evaluate the reputation of items using sentiment analysis. We explain the challenge of the increasing amount of data available on the Internet and the role of sentiment analysis in mining this information. We classify recent solutions into different categories based on techniques, document approach, and rating methods. We present six methods corresponding to different categories and analyze them based on the technique used, advances and results.

Keywords: Sentiment analysis, opinion, reputation, trust

1 Introduction

Sentiment analysis is a new kind of text analysis which aims at determining the opinion and subjectivity of reviewers. With the growing popularity of websites like `Amazon.com` and `Epinion.com` where people can state their opinion on different products and rate them, the internet is replete with reviews, comments and ratings. It is thus easy to find subjective reviews on specific products. The online reputation of an item is considered as the cumulative opinion of the online community regarding that item.

The challenge of sentiment analysis is that, contrary to simple text classification, using an intuitive lexical-based classification doesn't work well. The reason is that among the overwhelming number of reviews, there are reviews which don't contain any intuitively subjective words and however express a strong opinion. Other reviews contain highly pejorative words and express a positive opinion (and reciprocally).

Here is an example of a contradictory review on a film: "This film should be *brilliant*. It sounds like a *great* plot, the actors are *first grade*, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up." In these sentences, positive words dominate but they don't reflect the global opinion. These types of reviews deceive the lexical-based classification and reduce the accuracy of this method.

The role of sentiment analysis is to address this problem with new methods of classification. The goal could be a simple polarity classification (positive or negative), or

* University of Lyon, INSA-Lyon, F-69621 Villeurbanne, France

a multi-class one (like the 5 star classification). In reviews, people express opinions on several subjects and the opinion mining process must be adapted to the specific subject we are interested in. The analysis can be done on a global topic level: the result is the general opinion on the discussed product; or on a more specific level and extract opinion on certain aspects of the product.

2 Applications of sentiment analysis

When consumers have to make a decision or a choice regarding a product, an important information is the reputation of that product, which is derived from the opinion of others. Sentiment analysis can reveal what other people think about a product. The first application of sentiment analysis is thus giving indication and recommendation in the choice of products according to the wisdom of the crowd. When you choose a product, you are generally attracted to certain specific aspects of the product. A single global rating could be deceiving. Sentiment analysis can regroup the opinions of the reviewers and estimate ratings on certain aspects of the product.

Another utility of sentiment analysis is for companies that want to know the opinion of customers on their products. They can then improve the aspects that the customers found unsatisfying. Sentiment analysis can also determine which aspects are more important for the customers.

Finally, sentiment analysis has been proposed as a component of other technologies. One idea is to improve information mining in text analysis by excluding the most subjective section of a document or to automatically propose internet ads for products that fit the viewer's opinion (and removing the others). Knowing what people think gives numerous possibilities in the Human/Machine interface domain.

Sentiment analysis for determining the opinion of a customer on a product (and consequently the reputation of the product) is the main focus of this paper. In the following section, we will discuss solutions that allow to determine the expressed opinion on products.

3 Classification of existing solutions

The existing work on sentiment analysis can be classified from different points of views: technique used, view of the text, level of detail of text analysis, rating level, etc.

From a technical point of view, we identified *machine learning*, *lexicon-based*, *statistical* and *rule-based* approaches.

- The machine learning method uses several learning algorithms to determine the sentiment by training on a known dataset.

- The lexicon-based approach involves calculating sentiment polarity for a review using the semantic orientation of words or sentences in the review. The "semantic orientation" is a measure of subjectivity and opinion in text.

- The rule-based approach looks for opinion words in a text and then classifies it based on the number of positive and negative words. It considers different rules for classification such as dictionary polarity, negation words, booster words, idioms, emoticons, mixed opinions etc.

- Statistical models represent each review as a mixture of latent aspects and ratings. It is assumed that aspects and their ratings can be represented by multinomial distributions and try to cluster head terms into aspects and sentiments into ratings.

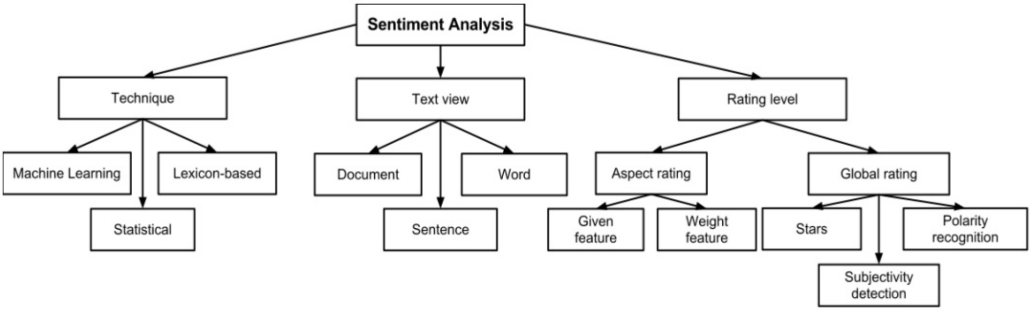

Fig. 1: Classification

Another classification is oriented more on the structure of the text: *document* level, *sentence* level or *word/feature* level classification. Document-level classification aims to find a sentiment polarity for the whole review, whereas sentence level or word-level classification can express a sentiment polarity for each sentence of a review and even for each word. Our study shows that most of the methods tend to focus on a document-level classification.

We can also distinguish methods which measure sentiment strength for different aspects of a product and methods which attempt to rate a review on a global level. Most of the solutions focusing on global review classification consider only the polarity of the review (positive/negative) and rely on machine learning techniques. Solutions that aim a more detailed classification of reviews (e.g., three or five star ratings) use more linguistic features including intensification, negation, modality and discourse structure [dAPGD11]. Figure 1 presents a detailed classification of existing methods. This classification is not exclusive. One solution can fit into more than one category.

4 Analysis of some existing solutions

This section presents six solutions that are representative for the categories mentioned earlier.

4.1 Sentiment Classification from Online Customer reviews Using Lexical Contextual Sentence Structure

The solution presented in the article by Khan et al. [KBK11] is a new method for sentiment analysis and classification: a domain independent rule based method for semantically classifying sentiment from customer reviews. As it is neither a learning-based, nor a lexicon-based approach, it is interesting to compare this method with others.

This method operates in three parts. First, the reviews are preprocessed: they are split into sentences which are corrected and the method “Part Of Speech” (POS) is used to tag and store each word of the sentence. Second there is the opinion word extraction phase which allows determining the polarity of a sentence based on the contextual information and sentence's structure. The "aspects" of a product are identified as noun phrases of the sentences. The last part consists of classifying the sentences into objective or subjective using a rule based module. Each opinion word has a semantic score extracted from the SentiWordNet dictionary containing the semantic score of more than 117,662 words. With these scores, by rating each term, the sentence can be assigned a weight (global score at the sentence level) in order to decide whether the review is positive or negative.

For the evaluation, three types of online customer reviews datasets were collected by the authors to check the system's performance (movie, airlines and hotel reviews). The data set consists of an average of 1,000 movie reviews, 1,000 airlines reviews and 2,600 hotel reviews. This is not much considering that some other methods have more than 10,0000 reviews as datasets. The performance has been assessed with an accuracy of 91% at the review level and 86% at the sentence level. Moreover, the sentence level sentiment classification performs better than the word level. The accuracy seems better than the average results of other methods (70-75%) but there is no comparison with other lexicon-based methods, neither with learning based methods.

4.2 Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis

Another solution is discussed in the article by Mudinas et al. [MZL12]: a concept-level sentiment analysis system called pSenti which combines lexicon based and learning based approaches. It measures and reports the overall sentiment of a review through a score that can be positive, negative or neutral or 1–5 stars classification. The main advantages and main interests of this article are the lexicon/learning symbiosis, the detection and measurement of sentiments at the concept level and the lesser sensitivity to changes in topic domain.

It operates in four parts. First, the pre-processing of the review where the noise (idioms and emoticons) is removed and each word is tagged and stored by the method Part Of Speech (POS). Second, the aspects and views are extracted to generate a list of top 100 aspect groups and top 100 views. The aspects are identified as nouns and noun phrases, and the views as sentiment words, adjectives and known sentiment words which occur near an aspect. Then the lexicon-based approach is used to give a "sentiment value" to any sentiment word and generates features for the supervised machine learning algorithm. Finally, this algorithm generates a "feature vector" for each aspect which is either the sum of the sentiment value for a sentiment word or the number of occurrences of this word in relation with other adjectives.

To evaluate this method, experiments were conducted on two datasets: software reviews (more than 10,000) and movie reviews (7,000). Software reviews were separated into two categories: software editor reviews and customer software reviews. As a result, pSenti's accuracy was proved close to the pure learning-based system and higher than the pure lexicon-based method. It was also shown that the performance was not as good on customer software reviews as on software editor reviews because customer software reviews are usually much "noisier" (with comments that are irrelevant for the subject) than professional software editor reviews. Its accuracy was also affected by a large number of reviews for which it failed to detect any sentiment or assigned neutral score. However, the sentiment separability in movie reviews was much lower than in software reviews. One of the reasons is that many movie reviews contain plots description and many quotes from the movie where words are identified as sentiments by the system.

4.3 Interdependent Latent Dirichlet Allocation

Interdependent Latent Dirichlet Allocation (ILDA) was introduced in 2011 by Moghaddam and Ester [ME11]. The main contribution of this paper is introducing the probabilistic assumption that there is interdependency between an aspect and its corresponding rating. An aspect (or feature) is an attribute or component of the product that has been commented on in a review. For example, 'battery life' in the opinion phrase 'The battery life of this camera is too short'. A rating is an intended interpretation of the user satisfaction in terms of numerical values. Most of the reviewing websites use ratings (number of stars) in the range from 1 to 5. A review is an assessment of the quality of a product posted online.

ILDA is a probabilistic graphical model which represents each review as a mixture of latent aspects and ratings. It assumes that aspects and their ratings can be represented by multinomial distributions and try to cluster head terms into aspects and sentiments into ratings. ILDA relies on a concept introduced in 2003 by Blei et al.: Latent Dirichlet Allocation (LDA). It is a generative probabilistic model for collections of discrete data such as text corpora [BNJ03]. The basic idea is that each item of a collection is modeled as a finite mixture over an underlying set of latent variables.

The experimental results show notably improved results for ILDA compared to the other two graphical models described in the paper (PLSI [Hof99] and LDA [BNJ03]), gaining in average almost 20% for the accuracy of rating prediction. They obtain in average 83% accuracy in aspect identification and 73% accuracy in aspect rating.

4.4 A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating

This solution was introduced in 2011 by de Albornoz et al. [dAPGD11]. The authors propose a method that globally rates a product review into three categories by measuring the polarity and strength of the expressed opinion. This solution was chosen as a representative of the global rating solutions, as it goes further than other solutions: they try to identify the strength of the opinion as well as the relevance of the feature the opinion is about. The mechanism of this method is straightforward: (i) the important features are identified; (ii) sentences containing opinions on those features are identified in the body of the review; (iii) polarity and strength are computed; (iv) a global score is computed. The main contribution of this paper is the fact that they do not rely on any previous knowledge about the importance of the features to the customer, contrary to Hu and Liu in [HL04], but learn it from a set of reviews using an unsupervised model. Another contribution is that each feature is automatically weighted.

Feature importance and opinion extraction, as well as opinion rating (the 3rd step) rely on the WordNet lexical database for English. This can be an important disadvantage of the method, as it cannot be applied on reviews written in other languages. The fourth step – predicting rating review – reveals the technical contribution of this paper. They propose the use of a Vector Feature Intensity (VFI) graph to structure the reviews. It is constructed using the strength of the opinion and the relevance of the feature the opinion is about. This graph is fed as input to any machine learning algorithm that will classify the review into different rating categories.

This solution offers a great flexibility when it comes to choosing the best machine learning method for classifying the reviews. The advancement stands in the creation of the VFI graph, which takes into consideration the relevance of the expressed opinion.

4.5 Opinion Digger

The solution Opinion Digger was introduced in 2010 by Moghaddam and Ester [ME10]. This method is a good and accurate example of a completely unsupervised machine-learning method. The particularity of this solution is to use as input a set of known aspects on a product and a ratings guideline (correspondence between ratings and adjectives i.e. 5 means "excellent" 4 means "good", ...). With these elements, it determines in output a set of other aspects and the ratings in each aspect according to the guideline. The idea was that many reviewing websites like amazon.com provide these input elements but there was no method that used them.

Opinion digger operates in two steps. First, it determines the set of aspects. After the pre-processing, each sentence is tagged with POS. It assumes that aspects are nouns so it first isolates the frequent nouns as potential aspects. With the sentences matching the known aspects, they determine opinion patterns, sequence of POS-tags that expressed opinion on an aspect. The frequent patterns used with known aspects are considered opinion patterns. If reviews with a "potential aspect" noun match at least two different opinion aspects, Opinion digger considers the noun as an aspect.

The second phase is rating the aspects. For each sentence containing an aspect, Opinion Digger associates the closest adjective to the opinion. It searches two synonyms from the guideline in the WordNet synonymy graph. The estimated rating of the aspect is the weighted average of the corresponding rating in the guideline. Weight is calculated by the inverse of the minimum path distance between the opinion adjective and the guideline's adjective in the WordNet hierarchy

The experiments show good performance in aspect determination and an excellent accuracy in ratings. The evaluation of aspect ratings was made using only the known set of aspects and compared to 3 other unsupervised methods. Opinion Digger performs an average ranking loss of only 0.49, meaning the difference between estimated and actual ratings. By incorporating new current information in the machine learning process, Opinion Digger increases the accuracy of unsupervised machine-learning method.

4.6 Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach

The last solution that we discuss treats a special problem called Latent Aspect Rating Analysis with a Model-based method. The model is called the Latent Rating Regression (LRR) model [WLZ10] and was created by Wang et al. in 2010 [WLZ10]. It estimates ratings on different aspects in a review but also determines the emphasis of the author on each aspect. It uses a given set of aspects and the overall ratings of the review.

It starts with an aspect-segmentation step. By recursively associating words with aspects, it can build an aspect dictionary and link each phrase of a review to the corresponding aspect. Then it applies the model:"Our assumption of reviewer's rating behavior is as follows: to generate an opinionated review, the reviewer first decides the aspects she or he wants to comment on; and then for each aspect, the reviewer carefully chooses the words to express her or his opinions. The reviewer then forms a rating on each aspect based on the sentiments of words she used to discuss that aspect. Finally the reviewer assigns an overall rating depending on a weighted sum of all the aspect ratings, where the weights reflect the relative emphasis she has placed on each aspect." So the overall rating is not directly determined by the words used in the review but rather by latent ratings on different aspects which are determined by the words.

With a probabilistic regression approach, it converts this model into a Bayesian regression problem, and then determines the aspect ratings and weight with consideration to the author's intent. The overall rating r is assumed to be a sample drawn from a Gaussian distribution with variance delta square and mean the weighted sum of the aspect ratings S. S is the result of the weighted sum of the words W in the reviews. A multivariate Gaussian distribution is employed as the prior for aspect weight's alpha.

The experimentation shows an average performance compared to other unsupervised methods in aspect ratings. However, it achieves to estimate an aspect's weight which was the initial goal. Since the publishing of this article, Wang et al. wrote a new article in which they proposed an improved version of this method [WLZ11] in which they determine the latent aspect rather than using a given aspect list.

5 Comparison

Tables 1 through 6 present a comparison of the solutions discussed.

Tab. 1: Sentiment Classification Using Lexical Contextual Sentence Structure.

Technique	Rule based method
Text approach	Sentence level
Classification	Rating per aspect and global rating
Accuracy in rating	91% at the document level and 86% at the sentence level
Advancements	Said to be independent of the domain of the review (it does not depend on the subject) and rule-based method
Drawbacks	Based on WordNet database

6 Discussion

Most of the future work detailed in each article consists of improving certain algorithms used in the solutions. For example, for the rule based methods, the knowledge base for semantic scores is a dictionary containing an average of 118,000 words. One of the improvements is to expand this knowledge base to a bigger dictionary. The ability to extract the acute sense of a sentence is also desired for an efficient semantic orientation.

For the combination of learning based and lexicon based approaches, the main improvement to be made is to improve the objectivity/subjectivity detection of a sentence by developing or using a more effective subjectivity detection algorithm.

Tab. 2: Combining Lexicon and Learning based Approaches.

Technique	Combination of lexicon based and learning based approaches
Text approach	Document level
Classification	Rating per aspect and global rating
Accuracy in rating	Not specified
Advancements	The main advantages are the lexicon/learning symbiosis, the detection and measurement of sentiment at the concept level and the lesser sensitivity to changes in topic domain
Drawbacks	Reviews with a lot of noise (irrelevant words for the subject of the review) are often assigned a neutral score because the method fails to detect any sentiment

Tab. 3: ILDA.

Technique	Probabilistic graphical model
Text approach	Document level – each review is considered as a mixture of latent aspects and ratings
Classification	Rating per aspect
Accuracy in rating	73%
Advancements	Probabilistic assumption that a sentiment and its rating are inter-dependent. Overcomes a well-known problem of other methods where a positive sentiment expressed with a word having a negative connotation is misinterpreted (e.g., low price).
Drawbacks	The correspondence between identified clusters and the actual aspects or ratings is not explicit (general drawback for unsupervised models)

Tab. 4: Feature mining and Sentiment Analysis.

Technique	Machine learning
Text approach	Document level
Classification	Global rating
Accuracy in rating	71.7% – 3 categories; 46.9% – 5 categories
Advancements	They take into consideration both the strength of an opinion as well as the the relevance of the feature the opinion is about for the general customer
Drawbacks	Based on WordNet database. Cannot be applied on reviews written in other languages than English.

Tab. 5: Opinion Digger.

Technique	Unsupervised machine learning
Text approach	Sentence level
Classification	Rating per aspect
Accuracy in rating	Ranking Loss of 0.49
Advancements	Usage of ratings guideline and seed-aspect to determine all aspects
Drawbacks	Needs a guideline and known aspects to work. Also based on Wordnet database.

Tab. 6: LRR.

Technique	Regression model
Text approach	Sentence level / reviewers level – the latent aspect rating is determined for each review
Classification	Rating per aspects, latent aspect ratings
Accuracy in rating	Not specified
Advancements	A new model: The global rating is not the direct result of the words in the review but is drawn from latent aspect ratings. Introduction and resolution of the Latent Aspect Ratings Analysis problem.
Drawbacks	Only average performance in simple aspect ratings. Still needs to know the aspects (treated in future works).

As for probabilistic models, the correspondence between generated clusters and latent variables has not yet been rigorously explained. It is a path that must be explored in order to improve the performance.

Last but not least, an important question that is raised is regarding the language that the review is written in.

7 Conclusion

This study shows that the field of sentiment analysis has been well studied by researchers in the past few years. Many different methods have been developed and tested. However, a lot of work is yet to be done. The most common approach is machine learning, a method that needs a significant data set for training and learning the aspects and sentiments associated. Also, models tend to target a simple global classification of reviews, rather than rating individual aspects of the reviewed product. Only a few of the methods are able to reach a somewhat high level of accuracy. Thus, the solutions for sentiment analysis still have a long way to go before reaching the confidence level demanded by practical applications.

References

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[dAPGD11] Jorge Carrillo de Albornoz, Laura Plaza, Pablo Gervás, and Alberto Díaz. A joint model of feature mining and sentiment analysis for product review rating. In *Advances in Information Retrieval*, pages 55–66. Springer, 2011.

[HL04] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *Proceedings of the National Conference on Artificial Intelligence*, pages 755–760. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[Hof99] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[KBK11] Aurangzeb Khan, Baharum Baharudin, and Khairullah Khan. Sentiment classification from online customer reviews using lexical contextual sentence structure. In *Software Engineering and Computer Systems*, pages 317–331. Springer, 2011.

[ME10] Samaneh Moghaddam and Martin Ester. Opinion digger: an unsupervised opinion miner from unstructured product reviews. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1825–1828. ACM, 2010.

[ME11] Samaneh Moghaddam and Martin Ester. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 665–674. ACM, 2011.

[MZL12] Andrius Mudinas, Dell Zhang, and Mark Levene. Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 5. ACM, 2012.

[WLZ10] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792. ACM, 2010.

[WLZ11] Hongning Wang, Yue Lu, and ChengXiang Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 618–626. ACM, 2011.