A Survey on Detection of Reviews Using Sentiment Classification of Methods

[1]Rekha V.Bandakkanavar, [2]Ramesh Medar , [3]Geeta Hegde

[1,2,3]Dept. of Computer
[1,3]SGBIT, Belgaum. India
[2]GIT, Belgaum. India
rekha35vb@gmail.com, rameshcs.git@gmail.com, geeta.gappati@gmail.com

Abstract - Merchants selling products on the Web often ask their customers to review the products that they have purchased and the associated services. As e-commerce is becoming more and more popular, the number of customer reviews that a product receives grows rapidly. For a popular product, the number of reviews can be in hundreds or even thousands. This makes it difficult for a potential customer to read them to make an informed decision on whether to purchase the product. It also makes it difficult for the manufacturer of the product to keep track and to manage customer opinions. As the numbers of customers are growing, reviews received by products are also growing in large amount. Thus, mining opinions from product reviews is an important research topic. In the fast decade considerable research has been done in academia. However, existing research is more focused towards categorization and summary of such online opinions. In this paper we survey various techniques to classify opinion as positive or negative and also detection of reviews as spam or non-spam.

Keywords: *Opinion mining, POS tagging, Semantic Orientation, Spam Detection*

*****

1. INTRODUCTION

Online shopping has become very popular to purchase all things without leaving our home, and it is a convenient way to buy things like electronic appliances, furniture, cosmetics, and many more. With the rapid expansion of e-commerce, more and more products are sold on the Web, and more and more people are also buying products online. To enhance the customer satisfaction, merchants and product manufacturers allow customers to review or express their opinions on the products or services. The customers can now post a review of products at merchant sites. These online customer reviews, thereafter, become a cognitive source of information which is very useful for both potential customers and product manufacturers [1]. With more and more common users becoming comfortable with the Web, an increasing number of people are writing reviews. As a result, the number of reviews that a product receives grows rapidly. Some popular products can get hundreds of reviews at some large merchant sites.[1] Furthermore, many reviews are long and have only a few sentences containing opinions on the product. This makes it hard for a potential customer to read them to make an informed decision on whether to purchase the product. If he/she only reads a few reviews, he/she may get a biased view.

Product re-views exist in a variety of forms on the web [3]. For product manufacturer perspective, understanding the preferences of customers is highly valuable for product development, marketing and consumer relationship management. But this practice of asking customer for their reviews, gives good chances for "review spam" as anyone can write anything on web [3]. Review spam refers to the fraud spam written by spammer to hype the product features or defame them. Though these reviews are important source of information there is no quality control on this user generated data, anyone can write anything on web which leads to many low quality reviews still worse review spam which mislead customers affecting their buying decisions. Though this is the case in past few years there is growing interest in mining opinion from these reviews by academicians and industries; Detecting spam reviews is very critical task for opinion mining [4].

Textual information in the world can be broadly categorized into two main types: facts and opinions. Facts are objective expressions about entities, events and their properties. Opinions are usually subjective expressions that describe people's sentiments, appraisals or feelings toward entities, events and their properties. The concept of opinion is very broad. In this paper, we focus on opinion expressions that convey people's positive or negative sentiments and also focus on detection of review as spam or non-spam.

2. SENTIMENT ANALYSIS TECHNIQUES

Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text. We use the following review segment on iPhone to introduce the problem.[2] (An number is associated with each sentence for easy reference): "(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) Although the battery life was not long, that is ok for me. (6) However, my mother was mad with me as I did not tell her before I bought it. (7) She also thought the phone was too expensive, and wanted me to return it to the shop. … "

Sentences (2), (3) and (4) express positive opinions, while sentences (5), (6) and (7) express negative opinions or emotions. The opinion in sentence (2) is on the iPhone as a whole, and the opinions in sentences (3), (4) and (5) are on the "touch screen", "voice quality" and "battery life" of the iPhone respectively. In general, opinions can be expressed on anything, e.g., a product, a service, an individual, an organization, an event, or a topic.

The above opinion can be classified into positive or negative using the following approaches.

1. Document Sentiment Classification
2. Feature-Based Sentiment Analysis
3. Sentiment Analysis of Comparative Sentences

2.1 Document Sentiment Classification

This approach determines whether each document expresses a positive or negative opinion (or sentiment) on an object. The existing research assumes that the document is known to be opinionated. Naturally the same sentiment classification can also be applied to individual sentences. However, here each sentence is not assumed to be opinionated in the literature. The task of classifying a sentence as opinionated or not opinionated is called subjectivity classification. The resulting opinionated sentences are also classified as expressing positive or negative opinions, which is called the sentence-level sentiment classification. Given a set of opinionated documents D, it determines whether each document d∈D expresses a positive or negative opinion (or sentiment) on an object. Existing research on sentiment classification makes the following assumption:

Sentiment classification assumes that the opinion document d (e.g., a product review) expresses opinions on a single entity and the opinions are from a single opinion holder h. This assumption holds for customer reviews of products and services because each such review usually focuses on a single product and is written by a single reviewer [5].

Most existing techniques for document-level sentiment classification are based on supervised learning, such as naive Bayesian classification, and support vector machines (SVM).

2.2.1 Naive Bayes classifier : It is a commonly used supervised machine learning algorithm. This approach pre supposes all sentences in opinion or factual articles as opinion or fact sentences.

Naive Bayes uses the sentences in opinion and fact documents as the examples of the two categories. The features include words, bigrams, and trigrams, as well as the part of speech in each sentence. In addition, the presence of semantically oriented (positive and negative) words in a sentence is an indicator that the sentence is subjective. Therefore, it can include the counts of positive and negative words in the sentence, as well as counts of the polarities of sequences of semantically oriented words (e.g., ''++" for two consecutive positively oriented words). It also include the counts of parts of speech combined with polarity information (e.g., ''JJ+" for positive adjectives), as well as features encoding the polarity (if any) of the head verb, the main subject, and their immediate moodier. Naive Bayes assigns a document to the class c_i that maximizes $P(c_i /d^*_j)$ by applying Bayes' rule as follow,

$$P(c_i|d_j^*) = \frac{P(c_i)P(d_j^*|c_i)}{P(d_j^*)}$$

Where $P(d_j^*)$ is the probability that a randomly picked document d has vector d_i^* as its representation, and P(c) is the probability that a randomly picked document belongs to class c.

To estimate the term $P(d_j^*|c)$, Naive Bayes decomposes it by assuming all the features in d_j^* are conditionally independent, i.e.,

$$P(c_i|d_j^*) = \frac{P(c_i)\left(\prod_{i=1}^{m} P(f_i|c_i)\right)}{P(d_j^*)}$$

2.2.2 Support vector machines (SVM): It is a discriminative classifier is considered the best text classification method (Rui Xia, 2011; Ziqiong, 2011; Songho tan, 2008 and Rudy Prabowo, 2009).The support vector machine is a statistical classification method proposed by Vapnik. Based on the structural risk minimization principle from the computational learning theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set. Multiple variants of SVM have been developed in which Multi class SVM is used for Sentiment classification(Kaiquan Xu, 2011).

2.2.3 Unsupervised learning

Unsupervised method works in 3 steps. Firstly perform Part-of-speech tagging, later extracting two consecutive words (two-wordphrases) from reviews if their tags conform to some given patterns, e.g., (1) JJ, (2) NN. Secondly Estimate the semantic orientation (SO) of the extracted phrases using Point wise mutual information

$$PMI(term_1, term_2) = \log_2\left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1)\Pr(term_2)}\right). \quad (1)$$

Here $\Pr(term_1 \wedge term_2)$ is the co-occurrence probability of $term_1$ and $term_2$, and $\Pr(term_1)\Pr(term_2)$ gives the probability that the two terms co-occur if they are statistically independent.

The opinion orientation (oo) of a phrase is computed based on its association with the positive reference word "excellent" and its association with the negative reference word "poor":

oo(phrase) = PMI(phrase, "excellent") − PMI(phrase, "poor"). (2)

The probabilities are calculated by issuing queries to a search engine and collecting the number of hits. For each search query, a search engine usually gives the number of relevant documents to the query, which is the number of hits. Thus, by searching the two terms together and separately, we can estimate the probabilities in Equation 1.

Equation 2 can be rewritten as:

$$oo(phrase) = \log_2\left(\frac{hits(\text{phrase } NEAR \text{ "excellent"})hits(\text{"poor"})}{hits(\text{phrase } NEAR \text{ "poor"})hits(\text{"excellent"})}\right).$$

Thirdly the algorithm computes the average oo of all phrases in the review, and classifies

the review as recommended if the average oo is positive, not recommended otherwise.

2.2 Feature based sentiment classification

Due to the increasing amount of opinions and reviews on the internet, Sentiment analysis has become a hot topic in data mining, in which extracting opinion features is a key step. Sentiment analysis at both the document level and sentence level has been too coarse to determine precisely what users like or dislike. In order to address this problem, sentiment Analysis at the attribute level is aimed at extracting opinions on products' specific attributes from reviews. Hu" s work in (Hu, 2005) can be considered as the pioneer work on feature-based opinion summarization. Their feature extraction algorithm is based on heuristics that depend on feature terms" respective occurrence counts. They use association rule mining based on the Apriori algorithm to extract frequent itemsets as explicit product features. Popescu et al (2005) developed an unsupervised information extraction system called OPINE, which extracted product features and opinions from reviews [6]. OPINE first extracts noun phrases from reviews and retains those with frequency greater than an experimentally set threshold and then assesses those by OPINE" s feature assessor for extracting explicit features. The assessor evaluates a noun phrase by computing a Point-wise Mutual Information score between the phrase and meronymy discriminators associated with the product class. Popescu et a apply manual extraction rules in order to find the opinion words.

Kunpeng Zhang (2009), proposed a work which used a keyword matching strategy to identify and tag product features in sentences. Bing xu (2010) , presented a Conditional Random Fields model based Chinese product features identification approach, integrating the chunk features and heuristic position information in addition to the word features, part-of-speech features and context features.

At feature level, main focus on two key mining tasks.

1. Identify object features that have been commented on. For instance, in the sentence, "The picture quality of this camera is amazing," the object feature is "picture quality".
2. Determine whether the opinions on the features are positive, negative or neutral. In the above sentence, the opinion on the feature "picture quality" is positive.

Current research on object feature extraction is mainly carried out in online product reviews.Different formats may need different techniques to perform the feature extraction task [9, 10].

Format 1 − Pros, cons and the detailed review: An example of such a review is given in Figure 1.

My SLR is on the shelf
by camerafun4. Aug 11 '04
Pros: Great photos, easy to use, very small
Cons: Battery usage; included memory is stingy.
I had never used a digital camera prior to purchasing this Canon A70. I have always used a SLR … Read the full review.

Figure 1: An example review of Format 1

Pros in Figure 1 can be separated into three segments:

great photos ⟨photo⟩
easy to use ⟨use⟩
very small ⟨small⟩ ⇒ ⟨size⟩.

Cons in Figure 1 can be separated into two segments:
Battery usage ⟨battery⟩
Included memory is stingy ⟨memory⟩

To extract the features label sequential rules (LSR), which are generated from sequential patterns in data mining. Label sequential rules working process is as follows.
Each segment is first converted to a sequence. Each sequence element is a word, which is represented by both the word itself and its POS tag in a set [11]. In the training data, all object features are manually labeled and replaced by the label $feature. An object feature can be expressed with a noun, adjective, verb or adverb. Thus, they represent both explicit features and implicit feature indicators. The labels and their POS tags used in mining LSRs are: {$feature, NN}, {$feature, JJ}, {$feature, VB} and {$feature, RB}, where $feature denotes a feature to be extracted, and NN stands for noun, VB for verb, JJ for adjective, and RB for adverb.
Feature extraction is performed by matching the patterns with each sentence segment in a new review to extract object features. That is, the word in the sentence segment that matches $feature in a pattern is extracted.

Format 2 − Free format: The reviewer can write freely, i.e., no separation of Pros and Cons. An Example of such a review is given in Figure 2. The reviews of Format 2 usually use complete sentences. To extract features from such reviews, some unsupervised methods for finding explicit features that are nouns and noun phrases are used.

GREAT Camera. Jun 3, 2010
Reviewer: jprice174 from Atlanta, Ga.
I did a lot of research last year before I bought this camera... It kind a hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.
The pictures coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out. …

Figure 2. An example review of Format 2

The method requires a large number of reviews, and consists of two steps:
1. Finding frequent nouns and noun phrases. Nouns and noun phrases (or groups) are identified by using a POS tagger.
2. Finding infrequent features by making use of opinion words. Opinion words are usually adjectives and adverbs that express positive or negative opinions.

The idea is as follows: The same opinion word can be used to describe different object features. Opinion words that modify frequent features can also modify infrequent features, and thus can be used to extract infrequent features. For example, “picture” is found to be a frequent feature, and we have the sentence, “The pictures are absolutely amazing.” If we know that “amazing” is a positive opinion word, then “software” can also be extracted as a feature from the following sentence, “The software is amazing.” because the two sentences follow the same pattern and “software” in the sentence is also a noun. It evaluates each noun phrase by computing a point wise mutual information (PMI) score between the phrase and meronymy discriminators associated with the product class, e.g., a scanner class. The meronymy discriminators for the scanner class are, “of scanner”, “scanner has”, “scanner comes with”, etc., which are used to find components or parts of scanners by searching on the Web. The PMI measure is a simplified version of the measure in [8]

$$PMI(f,d) = \frac{hits(f \wedge d)}{hits(f)hits(d)},$$

Where f is a candidate feature identified in step 1 and d is a discriminator. Web search is used to find the number of hits of individuals and also their co-occurrences. If the PMI value of a candidate feature is too low, it may not be a component of the product because f and d do not occur frequently. The algorithm also distinguishes components/parts from attributes/properties using WordNet’s is-a hierarchy (which enumerates different kinds of properties) and morphological cues (e.g., “-iness”, “-ity” suffixes).

2.3 Mining Comparative Opinions

In general, a comparative sentence expresses a relation based on similarities or differences of more than one entity. The comparison is usually conveyed using the comparative or superlative form of an adjective or adverb. A comparative sentence typically states that one entity has more or less of a certain attribute than another entity [7]. A superlative sentence typically states that one entity has the most or least of a certain attribute among a set of similar entities. Comparatives are usually formed by adding the suffix -er and superlatives are formed by adding the suffix -est to their base adjectives and adverbs. For example, in “The battery life of Camera-x is longer than that of Camera-y", “longer" is the comparative form of the adjective “long". In “The battery life of this camera is the longest", \longest" is the superlative form of the adjective “long". We call this type of comparatives and superlatives as Type 1 comparatives and superlatives.

ISSN: 2321-8169
310 – 314

Adjectives and adverbs with two syllables or more and not ending in y do not form comparatives or superlatives by adding -er or -est. Instead, more, most, less and least are used before such words, e.g., more beautiful. We call this type of comparatives and superlatives as Type 2 comparatives and superlatives. Both Type 1 and Type 2 are called regular comparatives and superlatives.

Most comparative sentences contain comparative adjectives and comparative adverbs, e.g., better, and longer, many sentences that contain such words are not comparatives, e.g., "I cannot agree with you more". Similarly, many sentences that do not contain such indicators are comparative sentences (usually non-gradable), e.g., "Cellphone-x has Bluetooth, but Cellphone-y does not have."

To extract objects and object features being compared, many information extraction methods can be applied, e.g., Conditional Random Fields (CRF), Hidden Markov Models (HMM), and others. Jindal and Liu used label sequential rules (LSR) and CRF to perform the extraction [12]. The algorithm makes the following assumptions:
1. There is only one comparative relation in a sentence.
2. Objects or their features are nouns (includes nouns, plural nouns and proper nouns) and pronouns.

3. Conclusion

In this paper we surveyed existing techniques to detect positive and negative opinions. Thus there is a real and huge need in the industry for such services because every company wants to know how consumers perceive their products and services and those of their competitors. These practical needs and the technical challenges will keep the field vibrant and lively for years to come.

4. References

[1] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 168–177, 2004.

[2] Bing Liu and Lei Zhang " A survey of opinion mining and. sentiment analysis." In Chicago, 2010.

[3] Liu, B. Web Data Mining. Springer, 2007.

[4] N. Jindal and B. Liu, "Opinion spam and analysis", in WSDM, 2008.

[5] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86, 2002.

[6] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in NLP (HLT/EMNLP), 2005.

[7] S. Sarawagi, "Information extraction," to appear in Foundations and Trends in Information Retrieval, 2009.

[8] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424, 2002.

[9] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, 2006.

[10]. B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," Proceedings of WWW, 2005.

[11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proceedings of ICML, pp. 282–289, 2001.

[12] N. Jindal and B. Liu, "Mining comparative sentences and relations," Proceedings of AAAI, 2006.