

# Opinion Mining Survey

Trung Huynh - trunghlt@gmail.com

## 1 Abstract

With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. The sudden eruption of activity in the area of opinion mining and sentiment analysis, which deals with the computational treatment of opinion, sentiment, and subjectivity in text, has thus occurred at least in part as a direct response to the surge of interest in new systems that deal directly with opinions as a first-class object.

This survey covers techniques and approaches that promise to directly enable opinion-oriented information-seeking systems. It includes material on summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact that the development of opinion-oriented information-access services gives rise to. To facilitate future work, a discussion of available resources, benchmark datasets, and evaluation campaigns is also provided.

**Acknowledgement:** Much of this work is cited from [1].

## 2 Introduction

Even before the internet, we frequently ask our friends for recommendation or review about stuff we are interested in. The web has enabled collecting experience and opinions from a vast pool of users. The demand of these information interpretation and summarization has been increasing for both individual customers and corporations.

From the surveys [2], [3], 81% of Internet users (60% of Americans) have done online research on a product at least once; 20% (15% of all Americans)

do so on a typical day; Among readers of online reviews of restaurants, hotels, and various services (e.g., travel agencies or doctors), between 73% and 87% report that reviews had a significant influence on their purchases; Consumers report being willing to pay from 20% to 99% more for a 5-star-rated item than a 4-star-rated item (the variance stems from what type of item or service is considered); 32% have provided a rating on a product, service, or person via an online rating system, and 30% (including 18% of online senior citizens) have posted an online comment or review regarding a product or service.

For political information, [4] studied the 31% of Americans - over 60 million people that were 2006 campaign internet users: 28% said that a major reason for these online activities was to get perspectives from within their community; 34% said that a major reason was to get perspectives from outside their community; 27% had looked online for the endorsements or ratings of external organizations.

[5] notes 75,000 new blogs are created daily; 1.2 million new posts each day. Therefore businesses require a new tool to replace traditional methods in order to keep track of consumer opinions.

### 2.1 Applications to review-related websites

The current search engine can retrieve documents which contain search keywords but can't answer for questions like "How do people think about iPad 2?". In order to answer for these option questions there needs a different treatment to analyse sentiment in clause, sentence and document levels.

Sentiment analysis can also be used to create an automated review and opinion-aggregation machine. Online customers usually do careful research before buying expensive products. Such an opinion-aggregation machine helps customers save

time from searching and acknowledges or recommends them about products with good or bad reviews from others.

## 2.2 Applications as a sub-component technology

Sentiment analysis helps advertisers identify opinion in articles and distribute proper ads to readers. If the author is expressing positive opinion about a product, advertisers can distribute advertisement about that product or related accessories. If the author is expressing negative opinion about that product, advertisers can sell advertisement about its competitors. The effect of improper advertisement can be very low and counter-productive.

## 2.3 Applications in businesses and government intelligence

Applications in this area include reputation and public relationship management. Companies have been paying a lot of money for manually tracking reputation of their products online. Instead, an automated sentiment analysis machine can save them a significant amount of resource. In addition, these data can also help with trend prediction and proper reaction for corporates to optimise their customer satisfaction.

## 2.4 Applications across different domains

In politics, understanding what voters are thinking - clarification of politician positions, such as what public figures support or oppose, enhances the quality of information that voters have access to. Automatic analysis of the opinions also helps politicians with the most updated opinions people submit about pending policy or government-regulation proposals ( ? ], ? ).

# 3 Challenges

Consider the following sentence from Mark Twain: “Jane Austen’s books madden me so that I can’t conceal my frenzy from reader”. Regular

lexicon methods can’t help in this case because “madden” and “frenzy” suggests negative sentiment, while the whole sentiment should be positive. In order to solve the problem of sentiment analysis completely, it is necessary to consider the whole context in an article rather than purely key words.

Lets consider another example from ? ]:

- (1) I bought an iPhone 2 days ago.
- (2) It was such a nice phone. (3) The touch screen was really cool.
- (4) The voice quality was clear too.
- (5) However, my mother was mad with me as I did not tell her before I bought it. (6) She also thought the phone was too expensive, and wanted me to return it to the shop...

There are several different opinions here. The sentence (2) expresses a positive sentiment about the iphone as general. The sentences (3) and (4) also express positive opinions but about the phone’s screen and voice quality. The sentence (5) and (6) express negative opinions of the author’s mother on the author and the phone’s price. As we can see, sentiment detection is more complicating as in order to detect the opinions of the author on iphone, in each phrase which contains opinions, we need to detect if the opinion holder is the author himself and if he is talking about either the iphone or its features. Therefore the problem of sentiment detection potentially contains difficult subproblems which are entity detection, feature extraction and subjectivity and objectivity classification of opinions.

## 3.1 Named Entity Recognition (NER)

This is a hard problem that has been investigated widely by researchers around the world. State-of-the-art NER systems for English produce near-human performance ? ]. The best system entering MUC-7 scored 93.39% of f-measure while human annotator scored 97.60% and 96.95% ? ]. However most of NER systems work well in some specific domains and need to be retrained for other domains.

This problem also contains the problems of feature extraction and synonym grouping [?]. This part of the problem has remained a major challenge even though there are many attempts to solve it.

## 3.2 Opinion orientation classification

The most popular approach to this problem is to use opinion words such as *good*, *bad*, *poor*, *brilliant* to predict the sentiment in the context. However this approach is brittle because we need different sets of opinion words for different domain for it to work efficiently. [?] gives an example about a phone battery: “The battery is long”, and its camera: “This camera takes a long time to focus”. Even though the word “long” is used in both sentences, the former implies a positive sentiment while the latter implies a negative sentiment.

Another the problem with this approach is that it is not easy to build a lexicon that works efficiently. There are unlimited number of expressions that people use to express opinions. [?] show that there are words that we might not think they indicate any sentiment but they appear more frequently in a certain class, i.e “still” surprisingly mainly appears in the positive class.

We also need to build different lexicons for different domains because a bag of words in this domain is positive but can be negative in another domain. For example “go read the book” most likely indicates positive sentiment for book reviews, but negative sentiment for movie reviews.

## 4 The approaches

### 4.1 Sentence/phrase-level sentiment analysis

#### 4.1.1 Lexicon based

A typical approach to sentiment analysis is to use a lexicon of positive and negative words and phrases to identify the overall sentiment in a sentence or an article.

#### Unsupervised approaches

[?], [?], [?] built a lexicon using some function based on positive, negative polarity and subjectivity with in it.

[?] use combination of this method and the sentiment polarity of the previous sentence if the scoring function does not indicate a deterministic classification of a given sentence.

[?] use a semi-supervised method to build such a lexicon. Their idea is to classify words by searching for their relationship by looking at conjunctions such as “elegant but over-priced”, or “clever and informative”. In [?], after clustering, they simply assign the highest average frequency cluster as “positive cluster”. The classification precision is more than 90%. In some other work, some *seed words* which are known as positive or negative are used to predict sentiment orientation of other words in the same clusters or using Wordnet-defined relations.

[?] propose a spectral feature aligning (**SFA**) algorithm to align domain specific words from different domains into unified clusters, with the help of domain-independent words as a bridge. Compared to previous approaches, **SFA** can discover a robust representation for cross-domain data by fully exploiting the relationship between the domain-specific and domain-independent words via simultaneously co-clustering them in a common latent space.

#### Semi-supervised approaches

The idea is to use the output of an initial simple classifier to feed labeled data into a supervised trainer. The trained models learn certain patterns from the input data and hence can classify a wider set of inputs.

[?] use an initial classifier to learn patterns of subjective expressions (an interesting example is the noun “fact”, as in “The fact is...”, exhibits high correlation with subjectivity).

[?] use a similar method to learn patterns of HTML documents with polarity labels.

[?] identify relations that link opinion words and targets using a dependency parser and then utilise to expand the initial opinion lexicon and to extract targets based on bootstrapping (**double propagation**). The results show this approach outperforms these existing methods significantly.

Similar work can be found at [?], [?].

#### Supervised approaches

However lexicon based approaches are not that trivial. [?] built a lexicon applying machine learn-

ing techniques based on unigram models that can achieve over 80% in accuracy.

A comparison of supervised and unsupervised methods can be found in [? ].

## 4.2 Sentiment summarization

### 4.2.1 Supervised approaches

[?] try to detect the summary quotation of the sentiment in a review from [www.rottentomatoes.com](http://www.rottentomatoes.com). They experimented on two learning methods e.g. Naive Bayes and Regularized Logistic Regression to detect the summary quotation basing on some information like the length and location of sentences, dictionary words in the sentences and some other combinations. However the result is not good and the correct percentages of both methods are around or lower than 25%. It does not take into account that the sentiment in a review is expressed in not only one sentence.

[?] present a technique named **Delta TF-IDF** which is an intuitive general purpose technique to efficiently weight bag of word scores before classification. The authors then use SVM to train their models for and show that their method improve accuracy for three well known data sets.

[?] introduce two new feature selection methods and three new feature weighting methods with Naive Bayes and Support Vector Machine. Their methods prove that it is possible to maintain a state-of-the-art classification accuracy of 87.15% while using less than 36% of the features.

[?] develop a hybridised genetic algorithm (Entropy Weighted Genetic Algorithm - **EWGA**) that incorporates the information gain heuristic for feature selection. **EWGA** is designed to improve performance and get a better assessment of the key features. The experimental results using **EWGA** with SVM indicate high performance levels, with accuracy over 95% on the benchmark data set and over 93% for both the U.S. and Middle Eastern forums.

### 4.2.2 Classification based on relationship information

**Relationships between discourse participants**

[?] observe on manual examination of 100 responses in newsgroup that the relationship between two individuals in the “responded-to” network is more likely to be antagonistic. Assuming “respond-to” links imply disagreement, they effectively classify users into opposite camps via graph partitioning, outperforming methods that depend solely on the textual information within a particular document.

### Relationships between product features

[?] utilize agreement information in a task where one must predict ratings for multiple aspects of the same item. They then construct a linear classifier to predict whether all aspects of a product are given the same rating, and combine this prediction with that of individual-aspect classifiers so as to minimize a certain loss function. [? ]

[?] study the problems of *entity discovery* and *entity assignment*. *Entity discovery* is the problem of detecting entities explicitly mentioned in sentences while *Entity assignment* is the problem of detecting entity mention by pronouns and language conventions. *Entity discovery* is based on pattern discovery and *entity assignment* is based on mining of comparative sentences. Experimental results demonstrate the effectiveness on both forum posts and a commercial setting.

### Sentiment analysis of conditional sentences

[?] study sentiment analysis of conditional sentences. Sentiment in conditional sentences are different from normal sentences and hard to determine. For example, in the sentence, *if your Nokia phone is not good, buy this great Samsung phone*, the author is positive about *Samsung phone* but does not express an opinion on *Nokia phone*. Conditional sentences can start with also *unless, even if, until, as long as, assuming, supposing, in case, only if*. The authors use SVM to train data from 5 diverse domains and achieve remarkable results.

## 4.3 Multi-class classification

### 4.3.1 Relationships between classes

Standard multi-class categorization focuses on capturing the distinct features present in each class, and ignore the fact that “5 stars” is much more like

“4 stars” than “2 stars”. [?] observe that one-vs-all multi-class categorization scheme can outperform regression for a three-class classification problem (positive, neutral, and negative), perhaps due to each class exhibiting a sufficiently distinct vocabulary, but for more fine-grained classification, regression emerges as the better of the two.

[?] formulate rating inference as a metric label problem ([?]), so that a natural notion of distance between classes (“2 stars” and “3 stars” are more similar to each other than “1 star” and “4 stars” are) is captured explicitly. More specifically, an optimal labelling is computed that balances the output of classifiers that considers items in isolation with the importance of assigned similar labels to similar items.

## 5 Conclusion

Even though an emerge of interest in sentiment analysis have attracted worldwide researchers from both institutions and corporates, a state-of-the-art sentiment analysis machine’s performance is still far away from what a human can do, especially for sophisticated documents which include sarcasm, conditional sentences, implications and comparisons. Many sub-problems have been addressed and attempted. The scope of each of these sub-problems is huge and, thus, sentiment analysis still remains a very challenging problem to researchers. However I am positive about the future of sentiment analysis as demanded from industry and its commercial potential. The availability of online user-generated data also enables statistical approaches and, hence, fosters the research of sentiment analysis.