Sentiment Analysis-Towards Harvesting Opinions from the Net

Ashwini Rao

Mukesh Patel School of Technology Management & Engineering
Mumbai, Maharashtra, India

Abstract

Sentiment analysis also called as Opinion mining classifies various opinions in text into categories like positive, negative as well as an implicit category of neutral. The data for this classification comes from Web (reviews, blogs, social network, discussion forums etc.). This user generated content is now regarded as a true source for exploring factual and subjective information. Sentiment analysis application involves competitive and marketing analysis as well as detection of unfavorable rumors for risk management, thereby helping companies to improve customer service, enhance their products and check the vulnerability of competitors. Opinions which are classified as positive often mean profits and fame for individuals and customers but, the system unfortunately has a loop hole where fake opinions or reviews are posted to discredit some individuals or products without disclosing their true identity. The accuracy of a sentiment analysis system in principle is to find out how well it agrees with human judgment. This paper presents a survey of various Challenges, Data Store and Levels that appear in the field of sentiment analysis.

Keywords: *Sentiment Analysis, Opinion Holder, Web 2.0, Supervised Learning, Spammer group detection, Subjectivity Classification, Machine Learning.*

1. Introduction

The area of sentiment analysis has recently enjoyed a huge burst of research activity. There has been a steady undercurrent of interest for quite a while. The year 2001 or so seems to mark the beginning of widespread awareness of the research problems and opportunities that sentiment analysis and opinion mining raise, and subsequently there have been literally hundreds of papers published on the subject. The sudden rush in this direction is mainly because of

- Increase in the number of machine learning and Information retrieval methods.
- Availability of huge data sets because of popularity of World Wide Web using which the machine learning algorithms can be trained on.
- Realization of the fascinating intellectual challenges and commercial intelligence applications that the area offers.

Sentiment analysis aims to find the Opinions/Sentiments about various topics and reviews. The current trend in opinion mining is to mine online discussion as these focuses on current events and issues which work as motivators for many startup companies where the users only want evaluative opinions.

Due to the popularity and reach of Internet, the receivers of the information have gone far away from just consuming the content, to commenting and contributing to it. Many researches are being done in Information Retrieval, Text Mining, Natural Language Processing and Machine Learning Techniques .These techniques are used to process vast amounts of user generated text content. Text categorization differs from Sentiment in terms of criterion used for classification which is features of frequent text rather than views expressed. Given the rise of online commerce, it is surprising to find that the polarity of the reviews available online have a measurable and a significant influence on the actual customer preference. In the years since the dawn of the Internet, information access systems have been at the core of the user experience. They have empowered the Internet user to navigate it effectively to satisfy their information needs. Recently, real time social content is more and more becoming part of our perception of the real world. Yet, it is unclear how to develop systems to best enable users to explore this information. Also, the role played by subjectivity in real-time information systems is largely unknown.

Some reviews are also fake which are written for the purpose of promoting a product. This is called as opinion spamming. Opinions spamming about social and political issues can be more frightening as they tend to wrap opinions and mislead the masses into situations which may be difficult to handle. Research work is also being done in this direction to find out such spams, so that reviews can be believed upon. Review spam is however harder to detect because it can't be identified by just manually reading them. This is a critical problem for opinion mining. Spammers are usually hired by companies to promote their products and/to distract customers from their competitors. One more factor that some studies point out is that the number of reviews, positive or negative, may simply reflect "word of mouth," so that in some cases, what the

underlying correlative of economic impact is really not the amount of positive feedback but, is merely the amount of feedback in total. This explains reasons as to why in some cases, negative feedback is seen to "increase" sales as it leads to increased "buzz "which brings more attention to the product.

2. Challenges Faced in Sentiment Analysis

Creating systems that can process subjective information effectively requires overcoming a number of novel challenges. Research in this field started with classification which were either sentiment or subjectivity classification. Classifying a document with views within it or checking whether a sentence expresses positive, negative or neutral opinion is sentiment classification. Determining whether a given sentence is subjective or objective is subjectivity classification. But in real life applications a user often wants to find out the opinions which require a detailed analysis.

The foremost problem is to discover the object or a target entity that has been commented on. Unless this is known, the opinion is of little use. So the first need given any piece of text is to separate relevant and irrelevant objects. The other issue is to find out the features which are being commented on. The problem aggravates as people tend to use synonym features, i.e. different words or phrases to describe the same feature. The next challenge is to finally give an opinion using any of the supervised or unsupervised method but this requires one to find out the opinion words which are unlimited in number. The problem here is that the people use different expressions in different domains as well as same domain sometimes. The challenge also comes when we have a sentence that may not explicitly comment on any product but are implied due to pronouns, language conventions and the context.

Next issue is to deal with Negation which plays a very important role in Sentiment Analysis. It is a very common linguistic construction, conveyed by common words like nor, neither not etc. Research in the field has shown that these words invert the polarity of an opinion expressed such as "I find this mobile phone great, but fail to understand why it is so heavy". This is an example which shows the effect of connectives having an impact on opinion expressed.

Next is a situation when the sentences have no opinions at all e.g. "If I can find a good camera I will buy it". Handling such sentences is a tricky job as is the case when we have a co reference resolution problem. This set of interesting sentences occur when we have a have reviewer giving his review in comparison with a similar item as well as ,on various features of the product, e.g. "Try out working with Google chrome, as Firefox keeps crashing". This statement clearly compares Google Chrome with Firefox, and it gives a negative feedback about Firefox, but opinion about Google chrome is not given.
The other challenge is to find out the various spam behavior models which may be targeting groups or may be targeting products. Identifying Spammers is also a critical job. These spammers are either individual spammer who does not work with anyone. He/she just writes fake reviews using a single user-id e.g. Author of a book. The other type of spammer is Group Spammers who work in groups and a single person registers multiple user-ids called as sock puppetting.

3. Data Store for Sentiment Analysis

The recent growth in the volume of data in the real-time web, specifically on Micro blogging sites like Twitter [1], is staggering. At least one website has recently measured the rate of Twitter posts, or tweets, being published as 2 billion per month, or 64 million per day[1]. The improvements in Smartphone and Tablet technology, combined with affordable pricing, mean that the barriers to access the social web have been considerably eroded. User-generated content can now be created and consumed instantaneously, wherever the user is. For example, if a user has a thought about a product they are using or has captured an interesting photograph concerning a breaking news story, they can instantly upload this to the web for others to see. Rich and diverse data is provided by Web 2.0 applications. Different dimensions to the data have been provided by social network sites, blogs, forums, review sites etc. Review sites are web sites whose purpose is to appraise a specific object. These sites are visited by users who post/give critical opinion about people, business, products or services. Blogs are simple web pages which consist of brief paragraphs of opinions, personal entries called as posts which are arranged according to the time they have been posted. The blogs have different styles of presentation and are updated by bloggers on an hourly/daily/weekly basis. Forums unlike blogs allow user to post on a specific/dedicated topic thereby restricting users to a single domain and thereby facilitates easy analysis. Social network sites like Twitter and Face book try to emulate relationships among people who know each other or share a common interest. The users of these sites can create their profiles as well as view profiles of other users who are added as their friends and can exchange views on various topics.

4. Levels of Sentiment Analysis

Sentiment analysis can be done at various levels, namely word level, phrase or sentence level, document level and feature level.

4.1 Document level Sentiment Analysis

This considers the whole document as the basic unit and determines its sentiment. In this level it's presumed that the opinion is given by a single opinion holder and on a single object. Various machine learning approaches exist for this task. Pang et al. [4] used traditional machine learning methods to classify reviews as positive and negative. They experimented with three classifiers (Naive Bayes, maximum entropy, and support vector machines) and features like unigrams, bigrams, term frequency, term presence and parts-of-speech. Document level sentiment analysis has also been formulated as a regression problem by Pang and Lee [4]. The difficulty over here lies in the fact that there could be mixed opinions in a document and sometimes the opinions are expressed without even using the opinion words. This limitation has led to the next analysis which is extracting useful information from subjective sentences.

4.2 Sentence Level Sentiment Analysis

Here the task is to find out subjective sentences from a collection of objective and subjective sentences and then determining their sentiment orientation. This can be done by using a good sentiment lexicon but the problem crops up as objective sentences can also have opinion words. Yu and Hazivassiloglou [5] try to classify subjective sentences and also determine their opinion orientations. For subjective or opinion sentence identification, it uses supervised learning. For sentiment classification of each identified subjective sentence, it used a method similiar to Turney [7], but with many more seed words, and log-likelihood ratio as the score function. A simple method used by Liu et al. [8], was to aggregate the orientations of the words in the sentence to get over all polarity of the opinion sentence.

4.3 Word Level Sentiment Analysis

Here the polarity of words and phrases at sentence and document level is used for classification. Most works use the prior polarity [8] of words and phrases for sentiment classification at sentence and document levels. Word sentiment classification use mostly adjectives as features but adverbs, and some verbs and nouns are also used by researchers [10][11]. This classification uses various parts of speech to perform manual or semiautomatic construction of word lexicon. Dictionary based method which is used in this level uses a small seed list of words with known prior polarity. This list is further extended by extracting synonyms and autonyms from various online dictionaries. Kim and Hovy [12] manually created two seed lists consisting of positive and negative verbs and adjectives. They then expanded these lists by extracting from Word Net, the synonyms and antonyms of the words of the seed list and assigning them to appropriate list. Based on Word Net lexical relation, Kamps [13] measured the semantic orientation of words.. They collected words and all their synonyms in Word Net, i.e. words of the same synset. Then a graph was created with edges connecting pairs of synonymous words. The semantic orientation of a word was calculated by its relative distance from the two seed terms good and bad. The distance was the length of a shortest path between two words wi and wj. The issue in this method is to handle classifications which are not domain specific.

Corpus based method rely on the techniques like co-occurrence of words with another. Hatzivassiloglou and McKeown [14] predicted the orientation of adjectives by assuming that pairs of conjoined adjectives have same orientation (if conjoined by and) and opposite orientation (if conjoined by but). Thus they used conjunctions such as ―corrupt and brutal or ―simplistic but well-received to form clusters of similarly and oppositely-oriented words using a log linear regression model.

They intuitively assigned the cluster that contained terms of higher average frequency as the positive list. In all the above level of analysis like document, sentence and word level , the feature on which the opinion holder is giving his opinion is not known. So we have feature based analysis.

4.4 Feature Based Sentiment Analysis

Feature based analysis involves extracting features of the product and then finding out the opinions about it. Yi et al.[10] restricted the candidate words further by extracting only base noun phrases, definite base noun phrase(noun phrases preceded by a definite article ―the) and beginning definite base noun phrases(definite base noun phrase at the beginning of a sentence followed by a verb phrase). For each sentiment phrase detected, its target and final polarity is determined based on a sentiment pattern database. Hu and Lui[11] extract the feature that people are most interested in and thus extract the most frequent noun or noun phrase using association mining.

Khairullah Khan [15] developed a method to find features of product from user review in an efficient way from text through auxiliary verbs (AV) {is, was, are, were, has, have, had}. From the results of the experiments, they found that 82% of features and 85% of opinion-oriented sentences include AVs. Most of existing methods utilize a rule-based mechanism or statistics to extract opinion features, but they ignore the structure characteristics of reviews. The performance has hence not been promising. Yongyong Zhail [17] proposed a approach of Opinion Feature Extraction based on Sentiment Patterns, which takes into account the structure characteristics of reviews for higher values of precision and recall.

With a self constructed database of sentiment patterns, sentiment pattern matches each review sentence to obtain its features, and then filters redundant features regarding relevance of the domain, statistics and semantic similarity.

5. Conclusion

The proliferation of various micro blogging sites offers an unprecedented opportunity to create and employ theories & technologies that search and mine for sentiments. The research community has been focusing on various mining aspects but reported solutions are still far from perfect. Most business intelligence solutions based on Sentiment analysis are extremely powerful but require a base knowledge of the underlying data in order to leverage them effectively.

Feature extractions and synonym grouping remain to be very challenging. The main question that crops up is the "sentiment analysis accuracy of the current state of the art algorithms". This question is difficult to answer as there are so many individual sub problems which do not have annotated data for benchmark testing. One point that is worth mentioning is about the applications that Sentiment analysis can be used but requires to work on data about people's preferences which may trigger concerns about privacy violations.

Sentiment-analysis technologies allow users to consult many people who are unknown to them, but this means precisely that it is harder for users to evaluate the trustworthiness of those people they are consulting. Thus, opinion-mining systems might potentially make it easier for users to be mis-led by malicious entities, a problem on which more research can be carried on. However, the huge practical need for such opinions will keep this field of sentiment analysis vibrant and lively for years to come.

References

[1] Bermingham, A. and Smeaton, A. F. (2011). On using Twitter to monitor political sentiment and predict election results. In SAAIP - Sentiment Analysis where AI meets Psychology workshop at the International Joint Conference on Natural Language Pro-cessing (IJCNLP) November 13, 2011, Chiang Mai, Thailand.

[2] Phelan, O., McCarthy, K., Bennett, M., and Smyth, B. (2011). Terms of a feather: content-based news recommendation and discovery using Twitter. In Proceedings of the 33rd European conference on Advances in information retrieval, ECIR'11, pages 448–459, Berlin, Heidelberg. Springer-Verlag.

[3] Y. Wilks and J. Bien, "Beliefs, points of view and multiple environments,"in Proceedings of the international NATO symposium on artificial and human intelligence, pp. 147–171, USA, New York, NY: Elsevier North-Holland, Inc.,1984.

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

[5] Pang B. and Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, Proceedings of the Association for Computational Linguistics (ACL),2005:115–124.

[6] Yu H. and Hatzivassiloglou V., Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences.In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

[7] Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Association for Computational Linguistics (ACL), 2005: 417–424.

[8] Liu, B., Hu, M., & Cheng, J. Opinion observer: Analyzing and comparing opinions on the web. In Proceedings of the 14th international World Wide Web conference (WWW-2005). ACM Press: 10–14.

[9] Wilson T., Wiebe J., and Hoffmann P.,Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)2005: 347–354.

[10] Esuli, A., & Sebastiani, F. Determining the semantic orientation of terms through gloss classification.In Proceedings of CIKM-05, the ACM SIGIR conference on information and knowledge management, Bremen, DE, 2005.

[11] Aue, A. and Gamon, M., Customizing sentiment classifiers to new domains: A case study. Proceedings of Recent Advances in Natural Language Processing (RANLP),2005.

[12] Kim, S. and Hovy, E., Determining the sentiment of opinions.In Proceedings of the International Conference on Computational Linguistics (COLING) ,2004.

[13] Kamps, J., Marx, M., Mokken, R.J., de Rijke, M., Using WordNet to measure semantic orientation of adjectives. In Language Resources and Evaluation (LREC), 2004.

[14] Hatzivassiloglou, V. and McKeown, K., Predicting the semantic orientation of adjectives. In Proceedings of the Joint ACL/EACL Conference,2004: 174–181.

[15] Khairullah Khan, Baharum B. Baharudin, Aurangzeb Khan, and Fazal_e_Malik, "Automatic Extraction of Features and Opinion Oriented Sentences from Customer Reviews", World Academy of Science, Engineering and Technology 62 2010.

[16] Pingdom (2010). Twitter, now 2 billion tweets per month. http://royal.pingdom.com/2010/06/08/twitter-now-2-billion-tweets-per-month/.

[17] Yongyong Zhail, Yanxiang Chenl, Xuegang Hu, "Extracting Opinion Features in Sentiment Patterns" , International Conference on Information, Networking and Automation (ICINA),2010.

Ms. Ashwini Rao is pursuing PhD in Computer Engineering from NMIMS University. She has done her ME (Master of Engineering) and BE (Bachelor of Engineering) degrees in Computer Science. She is currently working as a Assistant Professor in Dept. of Information Technology at Mukesh Patel School of Technology Management and Engineering, Mumbai, India. She is currently working as a research scholar in the field of Sentiment Analysis.