A Survey of Opinion Mining and Sentiment Analysis

Vishakha Patel
Dwarkadas J. Sanghvi College of Engineering
Mumbai-400056
Maharashtra, India

Gayatri Prabhu
Dwarkadas J. Sanghvi College of Engineering
Mumbai-400056
Maharashtra, India

Kiran Bhowmick
Dwarkadas J. Sanghvi College of Engineering
Mumbai-400056
Maharashtra, India

ABSTRACT

A huge amount of online information, rich web resources are highly unstructured and such natural language are not solvable by machine directly. The increased demand to capture opinions of general public about social events, campaigns and sales of the product has led to study of the field opinion mining and sentiment analysis. Opinion refers to extraction of lines in raw data which expresses an opinion. Sentiment analysis identifies polarity of extracted opinions. The major challenge lies in analyzing the sentiments and identifying emotions expressed in texts. This paper presents a survey which covers a problem of sentiment analysis, techniques and methods used for the same.

General Terms

Sentiment classification, opinion mining, spam detection

Keywords

Opinion Mining, Sentiment Analysis, Naive Bayes, SVM

1. INTRODUCTION

Since widespread of World Wide Web, internet and extensive growth of social media, organizations feel need to study public opinions for decision making. However to analyze polarity of opinions the exact intelligent information needs to be filtered. Hence automated opinion mining and sentiment analysis systems are needed [1]. Opinion mining techniques are used to extract reviews, opinions, political issues, brand perception automatically from web [2]. And sentiment analysis tracks, examines and evaluates public mood by using natural language processing [3]. Opinion mining and sentiment analysis can be used for business intelligence systems so as to analyze the opinions of public towards their brand and accordingly implement market strategies [4]. This paper illustrates: Section 2. 3 levels of sentiment classification are explained, Section 3. Different types of opinions and opinion mining process is described, Section 4. Opinion spam detection and rules for identifying opinion spam is summarized, Section 5. Evaluation of reviews, Section 6. Conclusion, Section 7. Acknowledgements and Section 8. References.

2. SENTIMENT CLASSIFICATION

2.1 Document level Sentiment Analysis

It aims to classify an opinion or a single review where a single topic is to be studied. The basic source of information is whole documented text [5]. The challenge is that all sentences in document may not contribute to opinion about a specific entity. Hence subjectivity/objectivity classification is very important, so as to discard irrelevant sentences or text. In this context, focus is on supervised learning methods which are used for document level classification. E.g.: Naïve Bayes classification and Support vector machines.

The features that can be used for machine learning are individual words and frequency counts, parts of speech such as adjectives, opinion words to indicate polarity of sentiment (e.g.: good, wonderful are positive opinions and rubbish, cheap, terrible are negative opinions), opinion phrases and idioms, negations and words dependency based features[6].

A combination of the above mentioned features and techniques based on polarity of words are used to further improve classification. Another interesting method used is domain adaptation, as sentiment analysis is highly sensitive to the domain from which data source is used. As language dependencies and their context are different from domain to domain, hence expressing opinions may also vary. Hence domain adaptation is very useful in document level sentiment analysis [6].

2.2 Sentence level Sentiment analysis

In this method, polarity of every sentence in document is determined. Document level classification methods can be applied to individual sentences. In this technique, we classify a sentence as subjective or objective, and resulting sentences are further classified as positive or negative opinions [5, 6]. This can be classified as: 1) Subjectivity classification 2) Objectivity classification.

The above 2 subtasks are very useful as it filters out sentences which contains no opinions and classifies aspects, through which polarity of opinions is determined. Also sentence-level classification may not determine the exact opinion for complex and compound sentences, as it may contain multiple opinions. For e.g.: the bag pack is very light and is of best use for students however colors aren’t very attractive. Here there are 2 positive opinions and 1 negative opinion, but sentence overall is positive. Also not all subjective sentences forms opinions, objective sentences can also imply opinions. Hence we have to focus on both sentence classifications.

2.3 Entity level Sentiment analysis

Document level and sentence level classification may not be useful in all applications, as it fails to review opinion about a specific entity. Aspect based Sentiment analysis uses a set of problems which follows natural language processing techniques and gives a better opinion set [6].

Hence, the context to which opinion is formed is extracted for every aspect in the sentence. However to solve for complex and compound sentences, lexicon-based approach is used which works as:

1. We mark all words and phrases containing aspects and assign score of +1 and -1 for positive and negative words respectively.
2. We classify or focus on those words that can change the orientation of a sentence. e.g.: Negation words: neither, never.
3. Handling contrary words like ‘but’. E.g.: Phone x is great phone but Phone y is better in terms of processor performance.

4. Finally, after classifying opinions, we aggregate all similar opinions to determine final orientation of the sentence.

3. OPINION MINING

3.1 Types of opinion

Sentiment can be expressed as regular opinions and comparative opinions [6].

1) Regular opinions: Regular opinions can be direct or indirect. E.g.: The display quality is crisp.' Here it refers to aspect of "picture quality" directly, hence direct opinion. 'After applying the cream, my skin broke out completely''. Here the entity "cream" is indirectly expressed as negative opinion on aspect "skin". Direct opinions are simpler to process. To identify polarity of indirect opinions, one needs to have knowledge of the given data source domain [6]. For e.g.: In the above example of review of the cream, skin broke out expresses negative opinion.

2) Comparative opinions: Unlike regular opinions, they may express different opinions for same entity [7]. E.g.: The processor speed and screen resolution of S6 is better than IPhone 6; however the metal body of IPhone 6 is more attractive than S6. Here S6 has two positive opinions and 1 negative opinion in same sentence. Comparative opinions can be explicit and implicit. Explicit comparisons are easier to evaluate as they express single opinion, either positive or negative. E.g.: The processor speed of One Plus is better than Yureka. Here the aspect of processor speed is explicitly compared, which is positive for entity One Plus. Implicit comparison is objective statement where opinions are indirect and implied. E.g.: Program x execution time is greater than program y. Here greater execution time implies program x performance is not good as compared to program y. Hence negative opinion is expressed for program x.

3.2 Opinion mining process

The data set can be collected from various social media network, website reviews and it is processed to have precise oriented sets. Then we can apply classification techniques at document, sentence or aspect level as per requirements, to classify opinion as positive, negative or neutral. Here main focus is on supervised learning. For supervised learning method, Naïve Bayes, SVM can be used to classify positive, negative opinions [8, 9]. Using document level sentiment analysis, following techniques can be applied to implement the current features of the above analysis [7]:

1) N-gram model for extracting opinion words and phrases: The n-gram model is a contiguous sequence of n items from given data set [7]. Each sentence is spitted into words and its frequency is noted as tf (w, d)=|{w E d}| where tf(w,d) is number of times word occurs in text d. The term tp(w,d) only checks if word w is present, hence its function can be written as tp(w,d)=1 if w E d, else tp(w,d)=0. However if frequent words are uniformly distributed then its power will be low. The E.g.: The perfume smells good. For n=2 (bigrams) "this perfume", "perfume smells", "smells good".

2) Frequency-inverse document frequency measure for listing most important words with respect to context of the text: the tf-idf is a statistic that specifies how frequently the word is used and hence states its importance in the given text [7]. The inverse document frequency is used to measure the rareness of the word in the given text. Hence the greater the value of tf-idf, the lesser the words importance. Idf (w,D) of word w in document D is idf(w,D) =|D|/log(df(w,D)). [7]

3) Part of speech Tagger for identifying opinions words in text: Parts of speech as noted earlier are a feature which can be easily used to indicate opinions [7]. Hence POS tagger helps us to identify POS and associate it to the word. E.g.: The design is exquisite. Here "exquisite" is an adjective which is a positive opinion. Hence in a given sentence, there are very few words which indicate sentiment. In English, POS are noun, verb, adjective, and adverb. However adjective are given major focus in identifying sentiments.

Here focus is on supervised learning methods applied to sentiment classification - Naive Bayes classification and Support Vector machine. Using unigrams as classification features, the efficiency was to be good in Bayes and SVM.

1) Naive-Bayes Classifier

It is very popular algorithm as it is simple, efficient and shows better performance for real world problems. "Naive" assumes that features are fully independent. In spite of real world not following the above conditions, this algorithm solves the problems suited to normal distribution. This technique is supervised learning and statistical method. It assumes a probabilistic model and allows the capture of uncertain aspects in the text, by calculating probabilities of the outcomes [10].

Bayes theorem specifies mathematically the relation between probability of 2 events A and B. Let P(A) be conditional probability of event A conditioned by B and P(B) be conditional probability of event B conditioned by A.

Bayes Formula [10]

$P(A/B)=P(B/A).P(A) / P(B)$

This formula helps us to find conditional probability of contrary events and independent probabilities of events. Hence we estimate probability of a document is positive or negative, or likelihood that an event will take place is positive or negative. So we estimate the probability of word with positive or negative meaning by analysing examples of positive and negative series using

P(sentiment|sentence)= P(sentiment).P(sentence)/P(sentence) [10]

Using above relation, we estimate P(word|sentiment) for all words as:

P(word|sentiment)=(no. of words occurrences in a class +1)/(no. of words belonging to a class + total no. of words) [10]

Hence using Naive Bayes we can find the polarity of a document by estimating the polarity associated with an opinion word. The major advantage is that we can train this model even by using a relatively small training set.

2) Support Vector Machine

SVM is a supervised classifier [10, 11] which exists in linear and non-linear forms.

Using SVM, ideally datasets i.e. classes should be linearly separable. So that a line is found which divide the two classes perfectly in 2 regions. In real world problems classes cannot be perfectly linearly separable. Hence a function of higher order is applied which maps points in non-linear data to linear data. For example, consider an instance which belongs to either class employed or unemployed. There is a separating line which defines a boundary. At the right side of boundary all instances are employed and at the left side all instances are unemployed.

For training data set D, a set of n points can be written as:

D {(x , c) x R , c { 1,1}} ....(1)

Where, xi is a p-dimensional real vector [10].

We find the maximum-margin hyper plane i.e. splits the points having ci = 1 from those having ci = -1. Any hyper plane can be written as the set of points satisfying [10]: w • x - b =1 ....(2)

The distance between two hyper planes is w b and therefore w needs to be minimized. The minimized w in w, b subject to

ci(w.xi − b) ≥ 1 for any i = 1… n

SVM outperforms Naive Bayes by attaining maximum accuracy of approximately 80% using unigram data model.

4. OPINION SPAM DETECTION

Since there is versatile data available on web, it is generally referred by public to read reviews of the product that he will buy. Positive opinions can result in economic gain to company and hence it gave rise to a concept called fake/spam reviews [2]. Organizations deliberately post positive reviews to promote the brand. In the world where companies emerge every hour of the day, also gives rise to competitors [5]. They can put up fake reviews by giving unjust reasons to damage reputation [12]. A spammer can be individual or may work as groups by using different identities.

Manually classifying data as spam and non-spam is tiresome and not feasible as spammer can review the product as any other genuine user. Based on supervised learning, 2 sets of features were identified as [2]:

1) Review centric features include the rating each reviewer gives, the total reviews written and the number of reviews where only single review was written.

2) Product centric features include average rating of the product and the standard deviations in ratings.

Based on above pattern, behavior model assigns a spamming score and sums the total for reviewer.

Experimental results found 4 types of unexpected rules for identifying unusual behavior patterns of reviewer [2].

1) Using confidence unexpectedness measure, reviewers can be found who only gives positive reviews even to a product having overall low ratings.

2) Using support unexpectedness measure, reviewers can be identified who writes more than one review for a product where others have written only one.

3) Using attribute unexpectedness measure, most positive reviews from single reviewer can be found, where there are also many other people who have reviewed.

4) Using attribute unexpectedness measure, reviewers who review positive opinions for a particular brand and negative reviews to others can be identified.

5. EVALUATION OF REVIEWS

After application of various highly optimized techniques for classification of sentiments, orientation of opinion sets and spam detection, the determination of usefulness of review should be relevant after retrieving the above instances [6,9]. For e.g.: Most shopping sites like snapdeal.com prompts the user if review was useful. The results are then aggregated and displayed. To determine the helpfulness of review, we can use regression problem where model assigns a utility value to each review [13]. This can be used for training and testing the review ranking. Researchers have used features like review length, ratings, POS tags, opinion words and their polarity, tf-idf weighting scores and also subjectivity classification [14]. Finally it should be noted that low quality reviews may not always be spam. However it is still a valid review and hence must not be neglected and discarded as it is helpful for users.

6. CONCLUSION

Sentiment and Opinion mining are research fields in machine learning problems that are still progressing and improving on performance measures. In this survey, techniques and approaches to sentiment analysis are studied. We have tried to highlight the main aspects on which opinions are expressed. Since collective intelligence on fields related to commerce, tourism, education, health, business and corporate world has its data all over the web, relevant solutions to solve this problem are becoming research interest in recent years [15]. Also a new feature to sentiment analysis is spam detection which is studied in paper. However it still remains a challenge in machine learning field because of the unstructured nature of the natural language. The coming generation requires highly efficient reasoning methods to process, connect and relate unstructured data into machine process able data.

7. ACKNOWLEDGMENTS

We would like to thank the Department of Computer Engineering and Dwarkadas J. Sanghvi College of Engineering, Mumbai, India.

8. REFERENCES

[1] Bo Pang and Lillian Lee, “Opinion Mining and Sentiment Analysis”, 2008

[2] Anand Mahendran and Anjali Duraiswany, “Opinion Mining for text classification”, International Journal of Scientific Engineering and Technology (2277-1581),Vol No.2,2013

[3] Deepali Virmani, Vikrant malhotra and Ridhi tyagi, ” Sentiment Analysis Using Collaborated Opinion Mining”, 2014

[4] G.Vinodhini and RM.Chnadrasekaran, “ Sentiment Analysis and Opinion Mining: A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering (2277-128X), Vol No.2, 2012

[5] Raisa Varghese and Jayashree M, “A Survey on Sentiment Analysis and Opinion Mining ”, International Journal of Research in Engineering and Technology(2319-1163)

[6] Bing Liu and Lie Zhang, “A Survey on Sentiment Analysis and Opinion Mining ”, 2012

[7] Murthy Canapathibholta and Bing Liu, "Minings in comparative sentences", 2008

[8] Pravesh Kumar Singh and Mohd Shahid Husain, "Methodological Study Of Opinion Mining And Sentiment Analysis Techniques", International Journal On Soft Computing (IJSC) Vol. 5,2014

[9] Ion Smeurean, "Applying Supervised Opinion Mining Techniues on Online User Reviews", Informatica Economics, Vol No. 16, 2012

[10] K. Saraswati and Dr. A. Tamilrasai, "Investigation of SVM for Opinion Mining", Journal of theoretical and Apllied Information Technology, Vol. 59 No. 2, 2014

[11] M. Rushdi Saleh, M.T. Martin-Valdiva, A.Montejo-Raez and L.A. Urena Lopez, "Experiments with SVM to classify opinions in different domains", Elsevier (14799-14804), 2011

[12] I. Smeureanu, M. Zurini, " Spam Filtering for Optimization in Internet Promotions using

[13] Methods, Vol. 5, Issue.2, pp. 198-211, 2010. Mrs. Vijaylaxmi, Mrs. Shalu Chopra, Mrs Sangeeta Oswal, Mrs Deepshika Chaturvedi, "The How, When, and Why of Sentiment Analysis", IJCTA,Vol 4

[14] B.L. Minqing Hu, "Mining and Summarizing Customer Reviews," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, USA, 2004

[15] Smeureanu, A. Diosteanu, C. Delcea, L.A. Cotfas. "Busines Ontology for Evaluating Corporate Social Responsibility," (Ontologii de afaceri pentru evaluarea responsabilităţii, corporaţiilor), Amfiteatru Economic, vol. 29, 2011, pp. 28-42