Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Sentiment Analysis Based on Opinion Classification Techniques: A Survey

Veena Dubey
M.Tech, Dept of CSED, K.N.I.T.
Sultanpur, Uttar Pradesh, India

Dharmendra Lal Gupta
Associate Professor: Dept of CSED, K.N.I.T
Sultanpur, Uttar Pradesh, India

Abstract— ***In present scenario the advancement of blogging services, social websites and electronic media play important role by measuring user's or client's messages (for example customer reviews, comments and opinions). Opinion Analysis is imperative term alluded to gathering data through a source using NLP, computational etymology and content investigation technique. It settles by choice using subjective data extraction and also by opinion examination. Opinion analysis recognizes negative reviews and positive reviews by measuring how positively and negatively an entity (public, organization and product) is involved. Sentiment analysis is the region of study to analyze people's reviews, states of mind, emotions and feelings from composed languages. In this paper, we have introduced a writing study on sentiment analysis survey and the methodologies adopted for it. We have focused on various opinion classification techniques, which may performed on any data set. Now a day's the feature extraction approach which is used by specialists to deal with sentiment analysis which is utilized by legislator, news groups, manufacturing organizations, movies, products etc.***

Keywords— ***Sentiments, customer reviews, data mining, feature extraction, text analysis.***

I. INTRODUCTION

Opinion mining involves making a system using reviews posted by user as to improve the product's features. Given an arrangement of reviews our task involves features identification of any item on which clients have communicated their assessment (called product features). Techniques like data mining and NLP in order to mine the features have been used [2]. Every component is partitioned into positive and negative reviews. To decide the sentiment orientation, basically there are the three subtasks which are mentioned below:

- Identify an arrangement of adjectives regularly used to express opinions using NLP method, which are called opinion words.
- For each opinion word semantic orientation is determined.
- Decide the opinion or sentiment orientation for each sentence by generating a summary out of the discovered information.

II. SENTIMENT ANALYSIS

Sentiment classification is a technique to concentrate on the sentiments or opinions expressed in an article or conveyed orally. The term sentiment includes emotions, conclusions, behaviors and others.

A. Opinion mining:

Opinion mining involves analysing opinions, sentiments or mentality of the writer from the written text. Opinion mining uses the concepts of NLP, machine learning and data mining to perform this task. Opinion mining is the science which combines techniques of computational linguistics [1] and information retrieval. It is concerned with the opinions expressed rather than topics mentioned in the text.

B. Why opinion mining?

Online opinions have indirect influence on the business of several e-commerce sites. Those sites market their products and the web users go through the reviews of the item before buying that product. Many organizations utilize opinion mining systems to track client audits of products sold online. Opinion mining is an incredible way of maintaining focus on several business trends related to deals administration, status management and also advertising. Pattern prediction is also done using the opinion of the customers [3].

III. THE PROBLEM OF OPINION MINING

Opinion mining enables to see a structure from the intimidating unstructured text and to give a unified system for the current research. The reflection comprises of two parts: opinion definition and opinion summarization.

A. Opinion definition

Audit portion on iPhone has been utilized to present the problem (an id number is connected with each sentence for simple reference) [5]:

"(1) I purchased an iPhone a couple days back.
(2) It was such a decent phone.
(3) The touch screen was truly cool.
(4) The voice quality was clear as well.
(5) Be that as it may, my mother was frantic with me as I didn't advise her before I bought it.
(6) She also thought the phone was too expensive, and wanted me to return it to the shop ..."

The inquiry is: The thing that the authors [5] need to mine or concentrate from this survey? The first thing that they noticed is that there were several opinions in this review.

- Some positive opinions are expressed by sentences (2), (3) and (4) express while sentences (5) and (6) express negative opinions or emotions.
- Later they have also noticed that the opinions all have some objectives. Sentence (2) having the objective of the opinion is the iPhone as a whole, and the objective of the opinions in sentences (4) and (3) are "voice quality" and "touch screen" of the iPhone respectively.
- Sentence (6) having the objective of the opinion is the price of the iPhone, but the objective of the opinion/emotion in sentence (5) is "me", not iPhone.
- Finally, they may also notice the holders of opinions. The holder of the opinions in sentences (2), (3) and (4) is the author of the review ("I"), but in sentences (5) and (6) it is "my mother".

With this illustration in mind one can now formally characterize the sentiment mining issue. In simple way, opinions can be define about anything, e.g., a product, a service, an individual, an organization, an event, or a topic, by any person or organization.

Entity

An entity or substance is a product, service, person, event, organization, or topic. It is connected with a couple, e: (T, W), where T is a chain of components (or parts), sub-components, and so on, and W is a set of attributes of e. Each component or subcomponent additionally has its own set of attributes.

In a specific brand of phone is an entity, e.g., iPhone. It has a set of components, e.g., battery and screen, and also a set of attributes, e.g., voice quality, size, and weight. The battery component likewise has its own set of attributes, e.g., battery life, and battery size.

Based on this definition , an entity or element is represented as a tree or hierarchy. The root node of the tree is the name of the element or entity. The components or sub-components of the entity is considered as non-root node. Relation is shown as a link. Each node is connected with a set of attributes. An opinion can be expressed on any node and any quality of the node.

There are two primary types of opinions: regular opinions and comparative opinions. Regular opinions are often referred to simply as opinions in the examination writing. A comparative opinion expresses a relation of likenesses or differences between two or more entities, and/or an inclination of the opinion holder based on some of the shared aspects of the entities.

An opinion (or regular opinion) is just a positive sentiment or negative sentiment, attitude, emotion or appraisal around an element or a part of the element from a holder of opinion. Positive, negative and neutral are called opinion orientations (additionally called sentiment orientations, semantic orientations, or polarities) [4].

An opinion (or regular opinion) is a quintuple $e_{i,}$ $a_{i\,j}$, oo_{ijkl} ,h_k ,t_l where e_i is the name of an entity, $a_{i\,j}$ is an aspect of $e_{i\,,}$ oo_{ijkl} is the orientation of the opinion about aspect $a_{i\,j}$ of entity $e_{i,,}h_k$ is the opinion holder and t_l is the time when the opinion is expressed by h_k . The opinion orientation oo_{ijkl} can be positive, negative or neutral, or be expressed with different strength/intensity levels. When an opinion is on the entity itself as a whole, we use the special aspect GENERAL to signify it.

These five components are essential. Without any of them, it can be problematic in general. For instance, if one says "The picture quality is great", and we don't know whose picture quality, the opinion is of little use. However, we don't mean that every bit of information is needed in every application. For example, it is not important to know each opinion holder, if we need to summarize opinions from a countless people. Similarly, by the authors [4] don't guarantee that nothing else can be added to the quintuple. An important contribution of this definition is that it provides a basis for transforming unstructured text to structured data. The quintuple used to give the essential information for a rich set of qualitative and quantitative analysis of opinions. Specifically, the quintuple can be regarded as a schema for a database table.

B. Document Level Sentiment

Document Sentiment Classification classifies an opinion document (e.g., a product review) as expressing a positive or negative opinion or sentiment. The assignment is also commonly known as the document-level sentiment classification because it considers the whole document as the basic information unit.

Most existing procedures for document-level sentiment classification are depend on supervised learning, however there are some unsupervised methods too.

1.Classification based on Supervised Learning

Sentiment classification obviously can be defined as a supervised learning problem with three classes, positive, negative and neutral. Training and testing data used in the existing research are mostly product reviews, which is not surprising because of the above assumption. Since each review already has a reviewer-assigned rating (e.g., 1-5 stars), training and testing data are readily available. For example, a review with 4 or 5 stars is taken as a positive review, a review with 1 or 2 stars is taken as a negative review and a review with 3 stars is taken as a neutral review.

Any current supervised learning strategies can be applied to sentiment classification, e.g., naive Bayesian classification, and support vector machines (SVM).

This approach has been adopted by Pang et al. to classify movie reviews into two classes, positive and negative. It was shown that Naïve Bayesian or SVM, using unigrams (a bag of individual words) as features in classification performed well. Subsequent research used many more features and techniques in learning.

The main task of sentiment classification is to engineer an effective set of features. Some existing features are listed below [6].

- Terms and their frequency: These features are individual words or word n-grams and their frequency counts. In some cases, word positions may also be considered. The TF-IDF weighting scheme from information retrieval may be applied too. These features have been shown quite effective in sentiment classification.
- Part of speech: It was found in many researches that adjectives are important indicators of opinions. Thus, adjectives have been treated as special features.
- Opinion words and phrases: Opinion words are words that are mostly used to express positive or negative sentiments. For instance, beautiful, great, good, and amazing are positive opinion words, and awful, poor, and horrible are negative opinion words. However many opinion words are noun, adjectives and adverbs (e.g., refuse, junk, and crap) and verbs (e.g., like and hate) can also indicate opinions. Aside from individual words, there are additionally conclusion expressions and figures of speech, e.g., cost someone an arm and a leg. Opinion words and expressions are instrumental to conclusion examination for evident reasons.
- Negations: Obviously nullification words are imperative in light of the fact that their appearances regularly change the opinion orientation. For example, the sentence “I don’t like this camera” is negative. However, nullification words must be maneuvered carefully in light of the fact that not all events of such words mean nullification. For example, “not” in “not only but also” does not change the orientation direction.
- Syntactic dependency: Word dependency based features generated from parsing or dependency trees are also tried by several researchers.

2.Classification as per Unsupervised

It is not hard to imagine that phrases and opinion words are the dominating indicators for sentiment classification. Thus, using unsupervised learning based on such phrases and words would be quite natural.

IV. SENTENCE SUBJECTIVITY AND SENTIMENT CLASSIFICATION

Naturally the same document-level sentiment classification techniques can likewise be applied to individual sentences. The job of classifying a sentence as subjective or objective is often called subjectivity classification in the existing literature.

In sentence-level sentiment classification [8], the subsequent subjective sentences are also classified as expressing positive or negative opinions.

A. Opinion Lexicon Expansion

In the exploration writing, opinion words are also known as opinion-bearing words or sentiment words. Positive opinion words are used to show some desired states while negative opinion words are used to show some undesired states. Positive opinion words examples are: beautiful, wonderful, good, and amazing.

Negative opinion words examples are bad, poor, and terrible. Aside from individual words the opinion lexicon is also there, defined as opinion phrases and idioms, e.g. cost someone an arm and a leg. They are instrumental for opinion mining for evident reasons.

To arrange or gather the opinion word list, fundamental methodologies have been examined: corpus-based approach, dictionary-based approach, and manual approach. The manual approach is extremely tedious and thus it is not typically utilized alone, but combined with automated approaches as the final check because automated methods make mistakes.

1.Dictionary based approach

One of the straightforward techniques in this approach is based on bootstrapping using a small set of seed opinion words and an online dictionary, e.g., WordNet or thesaurus. The methodology is to first gather a small set of opinion words manually with known orientations, and then to grow this set by searching in the WordNet or thesaurus for their synonyms and antonyms. The recently discovered words are inserted to the seed list. The next iteration begins. The iterative procedure stops when not any more new words are found. After the procedure finishes, manual inspection can be carried out to remove and/or correct errors. Analysts have also used additional information (e.g., glosses) in WordNet and additional techniques (e.g., machine learning) to generate better lists.

B. Corpus-based approch and sentiment consistency

The strategies in the corpus-based approach depend on syntactic or co-occurrence patterns and also a seed list of opinion words to find other opinion words in a vast corpus. One of the key thoughts is the one proposed by

Hazivassiloglou and McKeown [9]. The method begins with a list of seed opinion adjectives, and uses them and a set of linguistic constraints or conventions on traditions on connectives to recognize extra adjective opinion words and their orientations. One of the constraints is about the conjunction AND, which says that conjoined adjectives usually have the same orientation.

This idea is called opinion consistency. Obviously, in practice it is not generally consistent. Learning is connected to a large corpus to figure out whether two conjoined adjectives are of the same or different orientations. Different and same orientation joins between adjectives form a graph. At last, clustering is performed on the graph to produce two sets of words: positive and negative.

C. Aspect –based Sentiment analysis

In spite of the fact that classifying opinionated texts at the document level or at the sentence level is useful in many cases, it doesn't give the important point of interest expected to numerous different applications. A positive opinionated document about a specific entity does not imply that the author has positive opinions on all parts of the element. In like manner, a negative opinionated document does not imply that the author dislikes everything. In a regular opinionated document, the writer composes both positive and negative parts of the element, in spite of the fact that the general sentiment on the entity may be positive or negative. Document and sentence sentiment classification does not provide such information. To get these points of interest, we have to go to the aspect level.

Instead of treating opinion mining simply as a classification of sentiments, analysis of aspect-based sentiment introduces a suite of problems which require deeper natural language processing capabilities, and also produce a richer set of results.

- **Aspect extraction:** Extract aspects that have been evaluated. For example, in the sentence, “The picture quality of this camera is amazing,” the aspect is “picture quality” of the entity represented by “this camera”. Note that “this camera” does not indicate the GENERAL aspect because the evaluation is not about the camera as a whole, but about its picture quality. However, the sentence “I love this camera” evaluates the camera as a whole, i.e., the GENERAL aspect of the entity represented by “this camera”. Bear in mind whenever one talk about an aspect, he/she must know which entity it belongs to.
- **Aspect sentiment classification:** Determine whether the opinions on different aspects are positive, negative or neutral. In the first example above, the opinion on the “picture quality” aspect is positive, and in the second example, the opinion on the GENERAL aspect is also positive.

V. OPINION SPAM DETECTION

It has become a common practice for people to find and to read opinions on the Web for many purposes. For instance, on the off chance that one needs to purchase an item, one typically goes to a merchant or review site (e.g., amazon.com) to read some reviews of existing users of the product. If one sees many positive reviews of the product, one is very likely to buy the product. However, if one sees many negative reviews, he/she will most likely choose another product. Positive opinions can bring about in significant financial gains and/or fames for organizations and individuals.

This tragically, gives great motivating forces for opinion spam, which alludes to human activities (e.g., write spam reviews) that attempt to deliberately mislead readers or automated opinion mining systems by giving undeserving positive opinions to some target entities in order to promote the entities and/or by giving unjust or false negative opinions to some other entities in order to harm their reputation. Such opinions are likewise called fake opinions, bogus opinions, or false reviews.

A. Singular Spammers and Group Spammers:

A spammer may act individually (e.g., the author of a book) or as a member of a group (e.g., a group of employees of a company) [5] .

Individual spammers: In this case, a spammer, who does not work with anyone else, writes spam reviews. The spammer may register at a review site as a single user, or as many fake users using different user-ids. He/she can also register at multiple review sites and write spam reviews.

Group spammers: A group of spammers works together to promote a target entity and/or to damage the reputation of another. They may also register at multiple sites and spam on these sites. Group spam can be very damaging because they may take control of the sentiment on a product and completely mislead potential customers.

B. Spam Detection Based on Supervised Learning

In general, detection of spam can be formulated as a classification problem with two classes, spam and non-spam. However, manually labeling the training data for learning is very hard, if not impossible. The issue is that identifying spam reviews by simply reading the reviews is extremely difficult in light of the fact that a spammer can carefully craft a spam review that is just like any innocent review. Since physically marking training data is hard, different courses must be explored in order to find training examples for recognizing possible fake reviews.

Certain sorts of copy and close copy reviews were regarded as spam reviews, and the rest of the reviews as non-spam reviews. Three sets of features were identified for learning:

- Review centric features: These are features about the content of reviews. Example features include actual words in a review, the number of times that brand names are mentioned, the percentage of opinion words, the review length, and the number of accommodating feedbacks.

- Reviewer centric features: These are features about each reviewer. Example features include the normal rating given by the reviewer, the standard deviation in rating, the proportion of the quantity of reviews that the reviewer wrote which were the first reviews of the products to the total number of reviews that he/she wrote, and the proportion of the number of cases in which he/she was the only reviewer.
- Product centric features: These are features about each product. Example features include the product price, the sales rank of the product (amazon.com assigns sales rank to 'now selling products' according to their sales volumes), the average review rating of the product, and the standard deviation in ratings of the reviews for the product.

C. Abnormal Behaviors Spam Detection

Because of the difficulty of physically marking training data, treating opinion spam detection as a supervised learning problem is problematic because many non-duplicated reviews can be spam too. Here, the various authors have discussed two techniques that try to identify a typical behaviors of reviewers for detecting spammers. For instance, if a commentator thought of all negative surveys for a brand yet different analysts were all positive about the brand, then this analyst is actually a spam suspect [8]. Below, a brief introduction using a case conduct for every startling quality is mentioned:

- Confidence Unexpectedness: Utilizing this measure can find reviewers who give every high ratings to products of a brand, however most different commentators are generally negative about the brand.
- Unexpectedness Support: By this measure, one can find reviewers who write multiple reviews for a single product, while other reviewers only write one review.
- Unexpectedness Attribute Distribution: By this measure, one can find that most positive reviews for a brand of products are from only one reviewer in spite of there are an expansive number of reviewers who have reviewed the products of the brand.
- Unexpectedness Attribute: By this measure, one can find reviewers who write only positive reviews to one brand, and only negative reviews to another brand.

D. Group Spam Detection

A group spam detection algorithm finds groups of spammers who work together to promote or demote some products. The procedure works in two steps:

Frequent pattern mining: First, it extracts the review data to produce a set of transactions. Each transaction represents a unique product and comprises of all the reviewers (their ids) who have checked on that product. Utilizing all the transactions, it performs frequent pattern mining. The patterns thus give us a set of candidate groups who might have spammed together. The reason for using frequent pattern mining is as: If a group of reviewers who only worked together once to promote or to demote a single product, it can be hard to detect based on their collective or group behavior.

E. Utility of Reviews

Determining the utility of reviews is typically defined as a regression problem. The learned model assigns a utility value to each review, which can be used in review ranking. In this area of research, the ground truth data used for both training and testing are usually the user-helpfulness feedback given to each review, which discussed above is provided for each review at many review sites. So unlike fake review detection, the training and testing data here is not an issue. Researchers have used numerous sorts of features for model building.

Example features include review length, review rating (the number of stars), counts of some specific POS tags, opinion words, weighting scores, words, product attribute mentions, comparison with product specifications, comparison with editorial reviews, and many more [7].

Finally, it is to be noted that review utility regression/classification and review spam detections are different concepts. Low quality or Not-helpful reviews are not necessarily fake reviews or spam, and helpful reviews may not be non-spam. A user often determines whether a review is helpful or not based on whether the review shows opinions on many aspects of the product.

A spammer can satisfy this requirement by carefully crafting a review that is just like a normal helpful review. Using the number of helpful feedbacks to define review quality is also problematic because user feedbacks can be spammed too. Feedback spam is a sub-issue of click fraud in search advertising, where a person or robot clicks on some online advertisements to give the impression of genuine customer clicks. Here, a robot or a human spammer can also click on helpfulness feedback button to increase the helpfulness of a review. Another important point is that a low quality survey is still a valid review and should not be discarded, but a spam review is untruthful and/or malicious and should be removed once detected.

VI. CONCLUSION

This paper introduced and surveyed the field of sentiment analysis and opinion mining. Due to many challenging research issues and a wide variety of practical applications, it has been a very active research area in recent years. In fact, it has spread from computer science to management science. In the paper we first presented the concept of sentiment analysis, which formulated the problem and provided a common framework to unify different research directions.

It then discussed the most broadly studied topic of sentiment and subjectivity classification, which determines whether a document or sentence is opinionated, and if so whether it carries a positive or negative opinion. We then described aspect-based sentiment analysis. After that we briefly introduced the problem of analyzing comparative sentences. Last but not least, we discussed opinion spam, which is increasingly becoming an important issue as more and more people are relying on opinions on the Web for decision making. Several initial algorithms have been described. Finally, we concluded that all the sentiment analysis tasks are very challenging. Our understanding and knowledge of the problem and its solution are still limited. The main reason is that it is a NLP task.

REFERENCES

[1] Andreevskaia, A. and S. Bergler. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In Proceedings of Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), 2006.

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

[3] Banea, C., R. Mihalcea and J. Wiebe. Multilingual subjectivity: are more languages better? In Proceedings of International Conference on Computational Linguistics (COLING-2010), 2010.

[4] Zhang, L. and B. Liu. Identifying noun product features that imply opinions. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2011), 2011.

[5] Zhai, Z., B. Liu., L. Zhang., H. Xu., and P. Jia. Identifying evaluative sentences in online discussions. In Proceedings of National Conf. on Artificial Intelligence (AAAI-2011), 2011.

[6] Qiu, L., W. Zhang., C. Hu., and K. Zhao. SELC: A self-supervised model for sentiment classification. In Proceedings of ACM International Conference on Information and knowledge management (CIKM-2009), 2009.

[7] Paul, M., C. Zhai, and R. Girju. Summarizing contrastive viewpoints in opinionated text. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010), 2010.

[8] Pantel, P., E. Crestan, A. Borkovsky, A. Popescu, and V. Vyas. Web-scale distributional similarity and entity set expansion. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2009), 2009.

[9] Vasileios Hatzivassiloglou and Kathleen R. McKeown Predicting the semantic orientation of adjectives. In Proc. of the 35th ACL/8th EACL, pages 174–181, 1997.