A Study on Sentiment Analysis: Methods and Tools

Abhishek Kaushik¹, Anchal Kaushik², Sudhanshu Naithani³

¹Kiel University of Applied Sciences, Computer and Electrical Department, Sokratesplatz 1, 24149 Kiel, Germany
²Amity University, Department of Management Studies, Sector 125, Noida, India
³Kurukshetra University, Department of Computer Science, Thanesar Taluk, Kurukshetra, India

Abstract: The purpose of social media has created many chances for people to publicly voice their beliefs, simply when they are employed to deliver an opinion hit a vital problem. Sentiment Analysis is a case of natural language processing which could mark the mood of the people about any specific product by analysis. Sentiment Analysis is a process of automatic extraction of features by mode of notions of others about specific product, services or experience. The Sentiment Analysis tool is to function on a series of expressions for a given item based on the quality and features. Sentiment analysis is also called Opinion mining due to the significant volume of opinion. Analyzing customer opinion is very important to rate the product. To automate rate the opinions in the form of unstructured data is been a challenging problem today. Thus, this paper discusses about Sentiment analysis methods and tools used.

Keywords: Data Mining, Opinion Mining, Opinion Summarization, Sentiment Analysis, Text Mining, Web Mining.

Figure 1: Hierarchy of Data Mining

1. Introduction

The era of electronic information in every phase of life is evolving rapidly, which tends to produce a large number of data. As an outcome huge volumes of data are generated in field of technology, business, healthcare, tourism, e-marketing, etc. Automated analysis systems are meant for analysis, summarization and classification of data and number of efficient methods to store huge amount of data. Text mining is an approach used different fields like machine learning, information retrieval, statistics, and computational linguistics for opinion mining. Web mining is a subset of text mining used to mine the unstructured web data in the form Content mining, Web Structure mining and Web Usage mining. The aim of sentiment analysis is to make an automated machine able to recognize and categorize emotions [2]. A thought, view, or attitude based on emotion instead of reason is called sentiment. Figure 1 shows the different sub level of Data Mining and the branches of sentiment analysis.

2. Literature Overview


3. Information Source

User view is an important factor for the improvement of the quality of services. Blogs, review sites, data and micro blogs provide a good information of the products and services provided to clients.

Blogs: The name relates all the blog sites is called blogosphere [1]. People express about their thoughts they want to share with others on a blog. Blog pages [1] have become the popular platform to share ones personal views about specific products.

Review sites: The opinions of others is being an important factor while purchasing anything. A large number of users express their views on a particular product. These reviews are easily available on the Internet. The re-viewer’s data used in most of the opinion classification gather from the e-commerce websites [10] like www.flipkart.com.

Data Set: The dataset contains different types of product reviews (including Books, DVDs, Electronics and Kitchen appliances) and movie reviews extracted from Flip-kart and IMDB webpage.

4. Sentiment Analysis

Sentiment analysis is a technique which is used to extract the meaningful information in the documents [6]. In general, opinion mining tries to figure out the sentiment of a writer about some specific aspect and also the overall contextual polarity of a document. The sentiment may be a judgment, mood or evaluation of the writer [2]. A core issue in this field is an opinion classification, where a review is classified as a positive or negative evaluation of a subjected object (film, book, etc.). The assessment of sentiment can be done in two ways:

4.1 Direct opinions: It gives positive or negative sentiment about the product directly [12]. For example, “The food quality of this hotel is poor” expresses a direct opinion.

4.2 Comparison: It means to compare the subject with any other similar objects [12]. For example, “The food quality of the hotel-a is better than that of hotel-b.” expresses a comparison. Figure 2 had a workflow of Opinion Mining.

The views are being extracted from writers review over their comment. Opinion feature extraction is a sub-process of opinion mining [15]. Pre-processing In this process, raw data taken and is pre-processed for feature extraction.

The preprocessing phase [1] has been further divided into a number of sub phases as follows:

Tokenization is the process to split up into tokens by removing white spaces, commas and other symbols, etc. Stop word Removal removes words (like “a, an, the, of, for,). Stemming reduce the relevant tokens into a single type. Normalization is a process that has English texts to be published in both higher and lower case characters and turns the entire document or sentences into lowercase/uppercase.

Feature extraction phase deals with feature types [3] (which identifies the type of features used for opinion mining), feature selection (used to select good features for opinion classification), feature weighting mechanism (weights each feature for good recommendation) reduction mechanisms (features for optimizing the classification process).

Types of features used for opinion mining could be:
1)Term frequency (The presence of the term in a document carries a weight age).
2)Term co-occurrence (features which occurs together like uni-gram, bi-gram or n-gram),
3)Part of speech information (POS tagger is used to separate POS tokens).
4)Opinion words (Opinion words are words which express positive (good) or negative (bad) emotions) [3].
5)Negations (Negation words (not, not only) shift sentiment

Figure 2: Work flow of Opinion Mining
5. Opinion Summarization

Summarization of opinion is a major character in the opinion mining process. Summary of reviews provided should be based on features or subtopics that are mentioned in the reviews. Many works have been done on summarization of product reviews [9].

The opinion summarization process mainly involves the following two approaches. Feature based summarization is a type summarization involves the finding of frequent terms (features) that are appearing in many reviews. The summary is submitted by selecting sentences that contain particular feature information. Characteristics present in review text can be identified using Latent Semantic Analysis (LSA) method.

Term frequency is a count of term occurrences in a document. If a term has higher frequency it means that the condition is more import for summary presentation. In many product reviews certain product features come out frequently and associated with user opinions about it. Fig. 3 has the architecture of Opinion Mining which says how the input is being classified on the various steps to summarize the reviews.

5. Standard Structure of Sentimental Analysis

Opinion Mining also called sentiment analysis is a process of finding user’s opinion towards a topic or a product. Opinion mining concludes whether the user’s view is positive, minus, or neutral about a product, issue, event, etc. Opinion mining and summarization process involve three primary steps, first is Opinion Retrieval, Opinion Classification and Opinion Summarization. Review Text is retrieved from review websites. Opinion text in blog, reviews, comments, etc. contains subjective information about the topic.

Reviews classified as positive or negative review. Opinion summary is generated based on features opinion sentences by considering frequent features about a matter.

5.1 Opinion Retrieval

It is the procedure of collecting review text from review sites. Different review websites contain reviews for products, movies, hotels and news.

5.2 Information retrieval

Techniques such as web crawler can be employed to collect the review text data from many sources and store them in a database. This step involves retrieval of reviews, micro-blogs and comments by user.

5.3 Opinion Classification

Primary steps in sentiment analysis are a classification of review text. Given a review document $M = \{M_1, \ldots, M_l\}$ and a predefined category set $K = \{\text{positive, negative}\}$, sentiment classification is to classify each day in $M$, with a label expressed in K. The approach involves classifying review text into two forms namely positive and negative [9]. Machine learning and dictionary based approach is more popular [3].

5.4 Opinion Summarization

Summarization of opinion is a major character in the opinion mining process. Summary of reviews provided should be based on features or subtopics that are mentioned in the reviews. Many works have been done on summarization of product reviews [9].

The opinion summarization process mainly involves the following two approaches. Feature based summarization a type summarization involves the finding of frequent terms (features) that are appearing in many reviews. The summary is submitted by selecting sentences that contain particular feature information. Characteristics present in review text can be identified using Latent Semantic Analysis (LSA) method.

Term frequency is a count of term occurrences in a document. If a term has higher frequency it means that the condition is more import for summary presentation. In many product reviews certain product features come out frequently and associated with user opinions about it. Fig. 3 has the architecture of Opinion Mining which says how the input is being classified on the various steps to summarize the reviews.

5.3 Opinion Classification

Primary steps in sentiment analysis are a classification of review text. Given a review document $M = \{M_1, \ldots, M_l\}$ and a predefined category set $K = \{\text{positive, negative}\}$, sentiment classification is to classify each day in $M$, with a label expressed in K. The approach involves classifying review text into two forms namely positive and negative [9]. Machine learning and dictionary based approach is more popular [3].

The opinion summarization process mainly involves the following two approaches. Feature based summarization a type summarization involves the finding of frequent terms (features) that are appearing in many reviews. The summary is submitted by selecting sentences that contain particular feature information. Characteristics present in review text can be identified using Latent Semantic Analysis (LSA) method.

Term frequency is a count of term occurrences in a document. If a term has higher frequency it means that the condition is more import for summary presentation. In many product reviews certain product features come out frequently and associated with user opinions about it. Fig. 3 has the architecture of Opinion Mining which says how the input is being classified on the various steps to summarize the reviews.

5. Standard Structure of Sentimental Analysis

Opinion Mining also called sentiment analysis is a process of finding user’s opinion towards a topic or a product. Opinion mining concludes whether the user’s view is positive, minus, or neutral about a product, issue, event, etc. Opinion mining and summarization process involve three primary steps, first is Opinion Retrieval, Opinion Classification and Opinion Summarization. Review Text is retrieved from review websites. Opinion text in blog, reviews, comments, etc. contains subjective information about the topic.

Reviews classified as positive or negative review. Opinion summary is generated based on features opinion sentences by considering frequent features about a matter.

5.1 Opinion Retrieval

It is the procedure of collecting review text from review sites. Different review websites contain reviews for products, movies, hotels and news.

5.2 Information retrieval

Techniques such as web crawler can be employed to collect the review text data from many sources and store them in a database. This step involves retrieval of reviews, micro-blogs and comments by user.

5.3 Opinion Classification

Primary steps in sentiment analysis are a classification of review text. Given a review document $M = \{M_1, \ldots, M_l\}$ and a predefined category set $K = \{\text{positive, negative}\}$, sentiment classification is to classify each day in $M$, with a label expressed in K. The approach involves classifying review text into two forms namely positive and negative [9]. Machine learning and dictionary based approach is more popular [3].

5.4 Opinion Summarization

Summarization of opinion is a major character in the opinion mining process. Summary of reviews provided should be based on features or subtopics that are mentioned in the reviews. Many works have been done on summarization of product reviews [9].

The opinion summarization process mainly involves the following two approaches. Feature based summarization a type summarization involves the finding of frequent terms (features) that are appearing in many reviews. The summary is submitted by selecting sentences that contain particular feature information. Characteristics present in review text can be identified using Latent Semantic Analysis (LSA) method.

Term frequency is a count of term occurrences in a document. If a term has higher frequency it means that the condition is more import for summary presentation. In many product reviews certain product features come out frequently and associated with user opinions about it. Fig. 3 has the architecture of Opinion Mining which says how the input is being classified on the various steps to summarize the reviews.
6. Hierarchy of Opinion Mining

- Document level Opinion Mining: A single document of opinionated text works as a basic data unit in this level [7]. Here the document level classification is a single review about a topic is viewed. Merely in the forums or blog scenario, there are possibilities for comparative sentences to appear and for clients to compare one product with another that has alike characteristics and that’s how document level analysis is not suitable for forums and web logs. Therefore the subjectivity/objectivity arrangement is very vital in this type of classification.

- Sentence level Opinion Mining: The calculated polarity of each sentence is considered in the case of sentence level Opinion Mining. The same classification approach as applied in document level, can be reactive to the sentence level classification problem also, but Objective and subjective sentences [12] necessarily be localized. Opinion words are carried by subjective sentences. These sentiment words aid to determine the sentiments related to that entity. After which the polarity classification takes place into positive and negative classes.

- Phrase level Opinion Mining: This level of classification is much more pinpointed approach to opinion mining. Here phrases containing opinion words are observed and the phrase level class is completed. Only in some special cases, where contextual polarity also matters, the effect may not be fully precise.

7. Techniques

Major data mining techniques used to dig the knowledge and information are: generalization, classification, clustering, genetic algorithm, association rule mining, data visualization, neural networks, fuzzy logic, Bayesian networks, and, decision tree. Number 5 has the techniques of Opinion Mining. Figure 5. Techniques of Opinion Mining

- Supervised Machine Learning: Classification is most oftenly used and very popular data mining technique [11]. Classification used to divide the possible results from a given data set is based on the basis of a defined set of attributes and a given predictive attributes. The given dataset is used as the training dataset consist of independent variables (properties of the dataset) and a dependent attributes (predicted attribute). A training dataset created model test on text corpus holds the same attributes but no predicted attribute. Accuracy of model checks on how faultless it is making a prediction. Double Propagation Algorithm is used to extract Product features and sentenced words.

- Unsupervised Learning: It differs to supervised learning, because unsupervised learning does not have definite targeted output connected with the input. Class label for any instance is not known so this technique of learning is about to learn by observation. Clustering is a technique which is also used in unsupervised learning. Clustering is an approach of bunching objects with similar properties into a group. Objects in a cluster are always dissimilar to the objects in other clusters.

- Case Based Reasoning: Case based reasoning is one of the emerging Artificial Intelligence supervised techniques. CBR is a fierce tool of computer reasoning and crack the problems (cases) in the closest way to real time scenario. This is a problem solving technique in which knowledge is personified as past cases in the library and it is not dependent on classical rules. The solutions of all the cases are stored in CBR warehouse known as Knowledge base or Case base.

8. Semantic Orientation

Problem of Opinion mining can be divided into two parts which are sentiment classification [13] and feature based opinion mining. The trouble of taking out the semantic orientation (SO) of a text (i.e., whether the text is positive or negative towards a peculiar subject matter) often takes as a
starting point the problem of determining semantic orientation for individual speech. The hypothesis is that, if the SO of relevant words in a text is given, SO for the entire text can be determined. The SO approach to Sentiment analysis is an unsupervised learning because it does not need advance training in order to mine the data. Figure 6 shows the details of the classification of approaches of semantic orientation.

- Corpus Based Approach- Emotional affinity of words is determined by Popular corpus-driven method. Emotional affinity is meant to learn their probabilistic affective scores from large corpora. The method to assign a happiness factor to words depending on the frequency of their occurrences in happy-labelled blog posts compared to their total frequency in a corpus containing blog posts labelled with “happy” and “sad” mood annotations. They also compare the happiness factor scores of words with the scores in the list.

- Dictionary Based Approach- Dictionary based approach contains used lexical resources (e.g-Word Net) which work as an asset to automatically acquire emotion-related words for emotion classification experiments. They start from a set of primary emotion adjectives, and then retrieve alike words from Word Net by utilizing all senses of all words in the synsets that contain the emotion adjectives. The process takes advantage of the synonym and hyponym relations in Word Net to manually find alike words to nominal emotion words. The affective weights are automatically acquired from a very large text corpus in an unsupervised fashion.

9. Tools Used In Opinion Mining

The tools used in the process of tracking the opinion or polarity from the user’s generated contents are:

- Review Seer tool – Work done by aggregation sites is automated by this tool. To collect positive and negative opinions for assigning a score to the extracted feature terms, the Naive Bayes classifier approach is used. The results are displayed as a simple opinion sentence [10].

- Web Fountain - Beginning definite Base Noun Phrase (BNP) heuristic approach is used here for extracting the product features. Development of a simple web interface is also possible.

- Red Opal –This tool allows the users to determine the features based opinion orientations of products. It assigns the scores to each and every product based on features extracted from the customer reviews. The results are displayed by a web based interface [1].

- Opinion observer-This is an opinion mining system which is used to analyze and compare different opinions [5] on the cyber space by using user generated contents. This system illustrates the results in a graph format clearly showing opinion of the product feature by feature. It uses a WordNet Exploring method to assign prior polarity.

10. Conclusion

Opinion mining is an emerging sphere of data mining used to receive the knowledge of the huge mass of data (data may be customer comments, feedback and reviews on whatever product or topic etc). Much research has been carried on to mine the opinions in the contour of a document, sentence and feature level sentiment analysis. It has been examined that now the opinion mining trend is proceeding to the sentimental reviews of twitter data, comments used in Facebook on pictures, videos or Facebook status. Therefore, this paper discusses about an overview of the sentimental analysis approach of Opinion Mining in detail with the techniques and tools.

References


[16] Rudy Prabowo, Mike Thelwell, "Sentiment Analysis: A Combined Approach".

[17] Bo Pang, Lillian Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts".

[18] Alec Go, Richa Bhayani, Lei Huang, "Twitter Sentiment Classification Using Distant Supervision".

[19] Archana Shukla, "Sentiment Analysis of Document Based on Annotation".

[20] Meena Rambocas, Joao Gama, "Marketing Research: The Role of Sentiment Analysis".

Author Profiles

Abhishek Kaushik is currently working in Siemens, Germany as a Master thesis student. He is in the final phase of completing his Masters degree in Information Technology from Kiel University of Applied Sciences. Before starting his Masters he received his Bachelor’s of Technology in Computer Science Engineering from Kurukshetra University in 2012.

Anchal Kaushik is pursuing his MBA in Competitive Intelligence and Strategy Management from Amity University Noida. Before this she received her Bachelor's of Commerce Degree in 2014 from CCS Meerut.

Sudhanshu Naithani has received his Bachelor's of Technology in Computer Science Engineering from Kurukshetra University in 2015. He is currently working as a research assistant under Assistant Professor Ravinder Madan at Manav Bharti University, Solan.