International Journal of communication and computer Technologies, ISSN: 2278-9723

Available at http://www.ijccts.org

Techniques of Opinion Mining and Sentiment Analysis: A Survey

D.Vidhya[1], G. Sivaselvan[2], V.Vennila[3]

[1] PG Scholar, [2, 3] Asst. Prof, Department of Computer Science and Engineering,
K.S.R. College of Engineering, Tiruchencode.

Received: 02-06-2015, **Revised:** 16-08-2015, **Accepted:** 18-10-2015, **Published online:** 29-11-2015

ABSTRACT

Sentiment analysis is the most emerging field, that extract the customer's opinion about particular product or service. In order to enhance the sales of a product and to improve the customer satisfaction, most of the on-line shopping sites provide the opportunity for customers to write reviews about products. For a popular product, the number of reviews can be in hundreds. Due to immense amount of customer's opinions, views and feedback available, it is very much significant to analyse, explore and organize their views for better decision making. Opinion Mining or Sentiment Analysis is the mining of opinions and sentiments automatically from text, speech, and database sources through Natural Language Processing (NLP). This survey paper gives an outline of the techniques that classify the opinions as positive or negative.

KEYWORDS

Clustering, Feature Extraction Techniques, Naïve Bayes Classifiers, Opinion Mining, Support Vector Machines.

1. INTRODUCTION

Language is a powerful tool for communication. It is also a means to express emotion and sentiment. In current search engine people to search for other's opinions from the Internet before purchasing a product, when we are not aware with a specific product, we ask the trusted sources to recommend one. The web is a huge warehouse of structured and unstructured data. The analysis of this data to extract suppressed public opinion and sentiment is a challenging task. In data mining research field, machine learning techniques have been applied to automatically identify the information content in text.

2. RELATED WORKS

Bakhtawar Seerat et al [3] proposed the work on how opinions are being extracted from online reviews and challenges of opinion mining. [4] present an insight into task of opinion mining. Vijay B.Raut et al [5] have compared the methods and produced the summary of different approaches used for opinion mining and the results obtained. G.Vinodhini et al [2] presented an overview of different opinion mining techniques with approaches used. Pravesh Kumar Singh, Mohd Shahid Husain [6] investigated movie review mining using machine learning and semantic orientation.Grouping feature expressions, which are domain synonyms, is critical for effective opinion summary [1].

3. DATA SOURCE

User opinion are considered as major criterion for the improvement of the quality of services. Blogs, review sites, data and micro blogs provide empathetic for the deliverable level of the products and services provided to customers.

Blogs
The name related to universe of all the blog sites is called blogosphere. People who wish to share with others can write the topic on a blog. Blog pages have become the popular means to express ones personal opinions about any topic or product.

Review sites

The decision of any user in purchasing, the opinions of others is being an important factor. A large number of user reviews are available on the web. The reviewers data used in most of the sentiment classification studies are collected from websites like www.amazon.com (product reviews).

Data Set

The collected dataset contains different types of product reviews extracted from the website Amazon.com including Books, DVDs, Electronic items and Kitchen appliances.

4. OPINION MINING

Opinion Mining or sentiment analysis can be defined as a sub-discipline of computational linguistics that emphases on extracting opinion of user. It is a Natural Language Processing (NLP) and Information

Extraction (IE) task that goals is to obtain feelings of writer expressed in positive or negative comments by analyzing a large number of documents.

The evaluation of opinion can be done in two ways:

- Direct opinion, gives positive or negative opinion directly about the object. For example, “The picture quality of this camera is poor” expresses a direct opinion.
- Comparison means to compare the object with other similar objects. For example, “The picture quality of camera-y is better than that of Camera-x.” expresses a comparison.

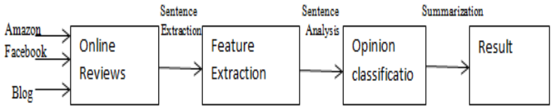

Fig.1 Opinion Mining Process

Opinion Mining contain three main components, they are:

1. Opinion Holder: The person who expresses their opinion.
2. Opinion Object: Opinion about the specific feature of an object.
3. Opinion Orientation: Customer view on an object.

5. ARCHITECTURE OF OPINION MINING

Opinion Mining is a process of finding user’s opinion towards a topic or a product. Opinion mining determines whether customer view is positive, negative, or neutral about product, topic, event etc. The process of opinion mining involve three main steps,

1) Opinion Retrieval
2) Opinion Classification
3) Opinion Summarization

Review Text is retrieved from the websites. Opinion text in blog, reviews, comments etc. contains subjective information about topic. Reviews classified as positive or negative review. Opinion summary is made based on features opinion sentences by considering frequent features about a topic.

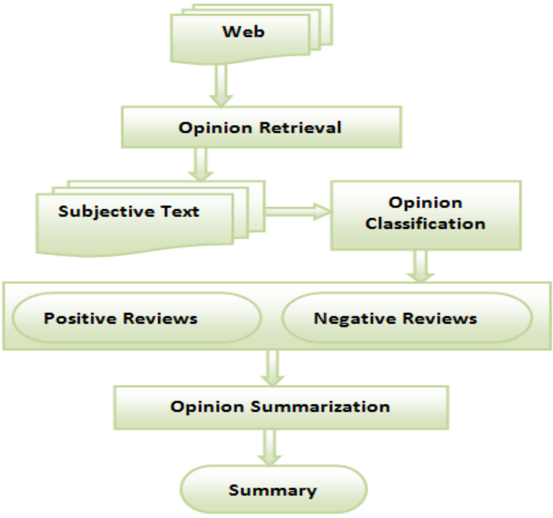

Fig.2 Architecture of Opinion Mining

5.1 Opinion Retrieval

It is the process of assembling review text from review websites. Different review websites contain reviews for products, movies, hotels. Information retrieval techniques such as web crawler can be applied to collect the review from many sources and store them in database. This step involves retrieval of reviews and comments of user.

5.2 Opinion Classification

Key step in sentiment analysis is classification of review text. Given a review document D = {d1…..d1} and a predefined categories set C = {positive, negative}, sentiment classification is to classify each di in D, with a label expressed in C. The approach involves classifying review text into two forms namely positive and negative.

5.3 Opinion Summarization

Summarization of opinion is a major part in opinion mining process. Summary of reviews provided should be based on features or subtopics that are mentioned in reviews. Many work have been done on summarization of product reviews. The opinion summarization process mainly involves the following two approaches. Feature based summarization is a type of summarization involved in finding of frequent terms or features that are appearing in many reviews. Sentences that contain particular feature information is presented as summary. Features present in review text can be identified using Latent Semantic Analysis (LSA) method. Term frequency is count of term occurrences in a document. If a term has higher frequency it means that term is more import for summary presentation. In many product reviews certain product features are frequently appeared and associated with user opinions about it.

6.OPINION MINING TASK AT DIFFERENT LEVELS

TABLE.1 Presents insight into opinion mining at different levels

Classification of Opinion mining at different levels	Assumptions made at different levels	Tasks associated with different levels
1. Opinion Mining at Sentence level.	1. A sentence contains only one opinion posted by single opinion holder; this could not be true in many cases e.g. there could be multiple opinions in compound and complex sentences. 2. Secondly the sentence boundary is defined in the given document	Task 1: identifying the given sentence as subjective or opinionated Classes: objective and subjective (opinionated) Task 2: opinion classification of the given sentence. Classes: positive, negative and neutral.
2. Opinion Mining at Document level.	1.Each document focuses on a single object and contains opinion posted by a single opinion holder. 2.Not applicable for blog and forum post as there could be multiple opinions on multiple objects in such sources.	Task 1: opinion classification of reviews Classes: positive, negative, and neutral
3. Opinion Mining at Feature level.	1. The data source focuses on features of a single object	Task 1: Identify and extract object features that have been

	posted by single opinion holder. 2. Not applicable for blog and forum post as there could be multiple opinions on multiple objects in such sources.	commented on by an opinion holder (e.g., a reviewer). Task 2: Determine whether the opinions on the features are positive, negative or neutral. Task 3: Group feature synonyms. Produce a feature-based opinion summary of multiple reviews.

7. TECHNIQUE

Major data mining techniques that used to extract the knowledge and information are: generalization, classification, clustering, association rule mining, data visualization, neural networks, fuzzy logic, Bayesian networks, and genetic algorithm, decision tree.

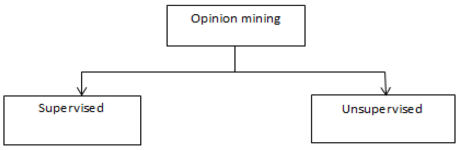

Fig.3 Techniques of Opinion Mining

7.1 SUPERVISED MACHINE LEARNING

Classification is most widely used supervised machine learning technique. Classification used to predict the possible outcome from given data set based on defined set of attributes and a given predictive attributes. The given dataset known as training dataset contains independent variables (dataset related properties) and a dependent attribute (predicted attribute). A training dataset created model test on test corpus contains the same attributes but no predicted attribute. Majorly used supervised algorithm are naïve bayes classifier, Support Vector Machine (SVM).

7.1.1 Naive Bayes Classification

A Naive Bayes Classifier calculate the probability based on Bayes' theorem and is particularly suited for high dimensional inputs. Naïve Bayes classification is an approach to text classification that assigns the class c to a given document d.

$$c* = \text{argmaxc } P(c|d) \qquad (1)$$

The Naive Bayes (NB) classifier uses the Bayes rule given in eq(2)

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \qquad (2)$$

Where P(c|d) is the probability of instance d being in class c , P(d|c) is the probability of generating instance d given class c , P(c) is the probability of occurrence of class c and P(d) is the probability of instance d occurring. To estimate the term P(d |c),

This theorem determines a conditional probability having the probability of dissimilar event and independent probabilities of events. Thus, we can evaluate the probability of an event based on the examples of its occurrence. In this case, the document probability is estimated as positive or negative. This is facilitated by the collection of positive and negative examples chosen. The process is termed as naive Bayesian because of how we calculate the probability of occurrence of an event - is the product of probability of occurrence of each word in the

document. This postulates that there is no connection between the words. This assumption of independence is introduced to facilitate the construction of classifier, it is not true in all cases, and there are words that appear together more frequently than individual.
The probability of a word with positive or negative meaning is estimated by analyzing a series of positive and negative examples and estimate the frequency of each of the classes. This learning process is supervised, requiring the existence of pre-classification examples for training.

$$P(sentiment|sentence) = \frac{P(sentiment)P(sentence|sentiment)}{P(sentence)}$$

we assume that P(sentence|sentiment) is a sentence. We estimate P(word|sentiment) product of P(word|sentiment) for all words in as:

$$P(word|sentiment) = \frac{\text{number of word occurences in class} + 1}{\text{number of words to a class} + \text{total number of words}}$$

Accuracy

For about 1000 sentences, train the Naïve Gauss Algorithm and got 79% accuracy, Where number of groups (n) is 2.

Advantages

1. Model is easy to understand.
2. Efficient computation.

Disadvantage

It is not necessary that assumptions of attributes being independent.

7.1.2 Support Vector Machines

Support vector machines (SVMs) is considered to be highly effective at traditional text categorization. They are large-margin, rather than probabilistic, classifiers.
.

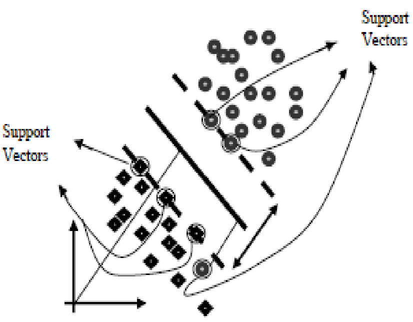

Fig.4 Principle of SVM

The basic idea behind the training procedure is to find a maximum margin hyper plane, represented by vector →ω, that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible. This corresponds to a constrained optimization problem; letting cj {1, −1} (corresponding to positive and negative) be the correct class of document dj, the solution can be written as,

$$\vec{\omega} = \sum_j \alpha_j c_j \vec{d_j}\ , \alpha_j \geq 0 \qquad (3)$$

Where c_j is a class and d_j is a document. Those d_j for which α_j are greater than zero are called support vectors.

Accuracy

SVM frequency model produce better output with an accuracy of 83%.

Advantages

1. Good performance on experimental results.
2. Low dependency on data set dimensionality.

Disadvantages

1. Pre-processing is needed in case of missing values.
2. Difficult interpretation of resulting model.

7.2 UNSUPERVISED LEARNING

In divergence of supervised learning, unsupervised learning has no explicit targeted output associated with input. Class label for any unknown instance is unsupervised learning considered to learn by observation. Clustering is major used technique for unsupervised learning. The process of assembling objects of similar characteristics into a group is called clustering. Objects in one cluster are dissimilar to the objects in other clusters.

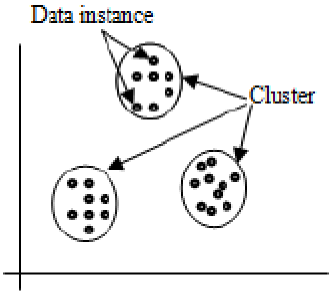

Fig.5 Clustering

7.2.1 Clustering Algorithms

A number of popular clustering algorithms are available. The basic reason of a number of clustering methods is that "cluster" is not accurately defined. As a result many clustering methods have been developed, using a different induction principle.

1. Exclusive Clustering :
In this clustering algorithm, clustering of data are done in an exclusive way, so that a data fits to only one certain cluster. K-means clustering is the example for exclusive clustering.

2. Overlapping Clustering :
This clustering algorithm uses ambiguous sets to grouped data, so each point may fit with two or more groups or various degree of membership are clustered.

3. Hierarchical Clustering :
Hierarchical clustering has two variations:

Agglomerative clustering is based on the union between the two nearest groups. The start state is recognized by setting every data as a group or cluster. After some iteration, the final needed clusters is obtained. It is a bottom-up version.

Divisive clustering begins from one group or cluster which containing all data items. At every step, clusters are consecutively fragmented into smaller groups or clusters according to some variance. It is a top down version.

4. Probabilistic Clustering
It is a mix of Gaussian, and uses totally a probabilistic approach.

7.2.2 Evaluation Criteria Measures for Clustering Technique

It is divided into internal quality criteria and external quality criteria.

1. Internal Quality Criteria
Using similarity measure it measures the density for clusters. It generally takes into consideration intra-cluster similarity, the inter cluster separability or both. It doesn't use any exterior information other than the data.

2. External Quality Criteria
External quality criteria are important for perceiving the cluster structure match to some classification of the instance or objects that are previously defined .

Accuracy

Depending on the data accuracy of the clustering techniques varied from 65% to 99%.

Advantages

The main advantage is that it offers the classes or groups that fulfil (approximately) an optimality measure.

Disadvantages

1. There is no learning set .
2. Number of groups is usually unknown.

8. CONCLUSION

Opinion Mining plays major role in making a decision about product or services. Opinion Mining has large application areas like education, shopping. For example websites like amazon.com allow customers to express their opinions on their websites. In this survey several machine learning techniques have been discussed and the related work has been done by using these techniques which automatically mines and ranking of products. Researchers have been carried out in other areas like android apps where trusted applications are downloaded based on reviews.

REFERENCES

[1] Zhai Z, Liu B, Xu H, and Jia P, Grouping Product Features Using Semi-supervised Learning with Soft-Constraints, in Proceedings of COLING. 2010.

[2] G.Vinodhini et al, “Sentiment Analysis and Opinion Mining: A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol 2, Issue 6, June 2012.

[3] Bakhtawar Seerat, FarouqueAzam, “Opinion Mining: Issues and Challenges (A Survey),” International Journal of Computer Applications, Vol49 No 9 July 2012, Pg No 42-51.

[4] N.Mishra and C.K.Jha, “An insight into task of opinion mining,” Second International Joint Conference on Advances in Signal Processing and Information Technology – SPIT 2012.

[5] Vijay .B.Raut et al, “Survey on Opinion Mining and Summarization of User Reviews on Web”, International Journal of Computer Science and Information Technologies (IJCSIT),Vol 5(2), 2014.

[6]Pravesh Kumar Singh, Mohd Shahid Husain , “Methodological Study Of Opinion Mining And Sentiment Analysis Techniques,” International Journal on Soft Computing (IJSC) Vol. 5, No. 1, February 2014.