Opinion Mining Techniques for Social Network Analysis: A Survey

Poornima Singh[1] Gayatri S Pandi (Jain)[2]
[1,2]Department of Computer Engineering
[1,2]L. J. Institute of Engineering and Technology, Gujarat Technological University, Ahmedabad, Gujarat, India

Abstract— Opinion Mining is a process of automatic extraction of knowledge from the opinion of others about some particular topic or problem. With the growing popularity of the Web 2.0, people are more and more provided with documents expressing opinions on different topics. In recent times, new research approaches were defined in order to automatically extract such opinions on the Internet. This review article focuses on the basics of Opinion Mining, analysis of data sources required for Opinion Mining. Further, machine learning techniques on the basis of their usage and importance for the analysis, tools, evaluation of Sentiment classifications and its various applications have also been overviewed.

Key words: Opinion Mining, Social Network Analysis, Comparative Opinions

I. INTRODUCTION

Opinion mining is the field of study that analyzes people's opinions, sentiments, attitudes, appraisals, evaluations, and emotions towards entities such as products, services, events, organizations, individuals, issues, topics and their attributes[1]. Microblogging has become a very popular communication tool among Internet users in recent years. As there are a lot of raw data of people posting real time messages about their opinions on a variety of topics in daily life, it is a worthwhile research endeavour to collect and analyze these data. Opinions of people's always play an important role in decision making process because whenever one need to make a decision one wants to hear other's opinions. This is true for individuals as well as for organizations. The analysis of sentiments of people, and accessing their emotions, attitudes and opinions is the main task of opinion mining. The advent of Web 2.0 and social media content has stirred much excitement and created abundant opportunities for understanding the opinions of the general public and consumers toward social events, company strategies, political movements, product preferences and marketing campaigns[2].

With the growing popularity of the Web 2.0, peoples are more and more provided with documents expressing opinions on different topics. Micro-blogging has become a very popular communication tool among Internet users in recent years. Recently, new research approaches were defined in order to automatically extract such opinions on the Internet.

Opinion mining aims to extract attributes and components of the object that have been commented on documents. It also aims to determine whether the comments are positive, negative or neutral as shown in fig. 1. Opinions of people's always play an important role in decision making process because whenever one need to make a decision one wants to hear other's opinions. This is true for individuals as well as for organizations. The analysis of sentiments of people, and accessing their emotions, attitudes and opinions is the main task of opinion mining.

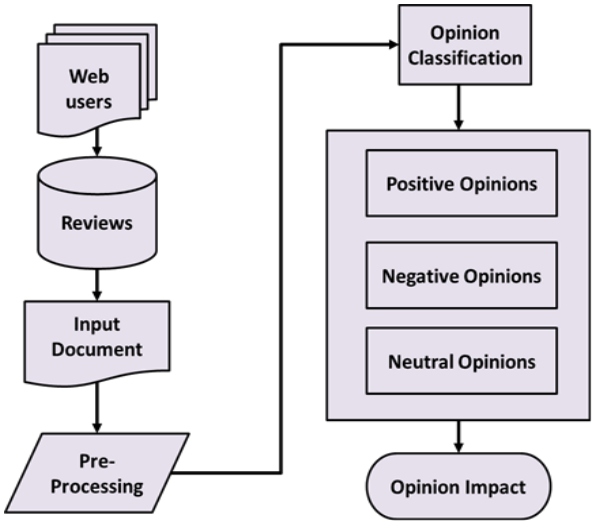

Fig. 1: Workflow of opinion mining

II. TYPES OF OPINIONS

A. Regular and Comparative Opinions:

1) Regular Opinion:

A regular opinion is often referred to simply as an opinion in the literature. It has two main sub-types:

2) Direct Opinion:

A direct opinion refers to an opinion expressed directly on an entity or an entity aspect. For example, the sentence, "The product feature is good", indicates a direct opinion about product. Much of the current research focuses on direct opinions because they are simpler to handle.

3) Indirect Opinion:

An indirect opinion is an opinion that is expressed indirectly on an entity or aspect of an entity based on its effects on some other entities. For example, the sentence "After injection of the drug, my joints felt worse" describes an undesirable effect of the drug on "my joints", which indirectly gives a negative opinion. Indirect opinions are often harder to deal with.

4) Comparative Opinion:

A comparative opinion expresses a relation of similarities or differences between two or more entities and/or a preference of the opinion holder based on some shared aspects of the entities[3]. For example, the sentences, "Thums Up tastes better than Coke" and "Thums Up tastes the best" express two comparative opinions.

B. Explicit and Implicit Opinions:

1) Explicit Opinion:

An explicit opinion is a subjective statement that gives a regular or comparative opinion. For example, "Thums Up tastes great," and "Thums Up tastes better than Coke."

2) Implicit (Or Implied) Opinion:

An implicit opinion is an objective statement that implies a regular or comparative opinion. Such an objective statement usually expresses a desirable or undesirable fact. For

All rights reserved by www.ijsrd.com

example, "The screen size of iPhone 6 is smaller than iPhone 6 plus."

III. LEVELS OF OPINION MINING

Opinion mining works mainly at three levels:

A. Document Level:

The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment. Document level analysis assumes that each document expresses opinions on a single entity. Thus, it is not applicable to documents which evaluate or compare multiple entities.

For example, given a news article, the system determines whether the article expresses an overall positive or negative opinion about news.

B. Sentence Level:

The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion where neutral usually means no opinion. The same document level classification methods can be applied to the sentence level classification problem also but Objective and subjective sentences must be found out. In sentence level Opinion Mining, the polarity of each sentence is calculated. This level of analysis is closely related to subjectivity classification, which distinguishes sentences (called objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions.

C. Entity and Aspect level:

Both the document level and the sentence level analyses do not discover what exactly people liked and did not like. The Aspect level performs finer-grained analysis. Aspect level directly looks at the opinion itself Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases). Aspect level was earlier called feature level (feature-based opinion mining and summarization). It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion).

IV. DATA SOURCES

User's opinion is a major criterion for the improvement of the quality of services rendered and enhancement of the deliverables. Review sites, Blogs, datasets and micro blogs provide a good understanding of the reception level of the products and services.

A. Blogs:

With an increasing usage of the internet, blog pages and blogging are growing rapidly; blog pages have become the most popular means to express one's personal opinions. The name associated to universe of all the blog sites called blogosphere. Bloggers record the daily events in their lives and express their opinions, feelings, and emotions in a blog. Blogs are used as a source of opinion in many of the studies related to sentiment analysis. Many of these blogs contain reviews on many products, issues, etc. Blogs provide a type of website that contains information and personal opinions of the individual authors[4].

B. Review sites:

The opinions of others can be an important factor for any user while making a purchasing decision. A large and growing body of user-generated reviews is available on the Internet. The reviewer's data used in most of the sentiment classification studies are collected from the e-commerce websites like www.amazon.com (product reviews), www.CNET download.com (product reviews), www.yelp.com (restaurant reviews). Other than these the available are professional review sites such as www.zdnet.com, www.dpreview.com and consumer opinion sites on broad topics and products such as www.consumerreview.com, www.bizrate.com and www.epinions.com.

C. Data Set:

Most of the work in the field uses movie reviews data for classification.

The datasethttp://www.cs.cornell.edu/People/pabo/movie-review-data contains Movie review data. Other dataset which is available online is multi-domain sentiment (MDS) dataset (http: // www.cs.jhu.edu/mdredze/ datasets / sentiment). This dataset contains four different types of product reviews extracted from Amazon.com including Books, DVDs, Kitchen appliances and Electronics, with 1000 positive and 1000 negative reviews for each domain.

D. Micro-Blogging:

The most prominent example of microblogging platforms, which allows for real-time analysis, is Twitter. Twitter is a popular microblogging service where users create status messages called "tweets". Twitter messages are also used as data source for classifying sentiment. These tweets sometimes express opinions about different topics. Its vast user community, all-around presence and informal style of conversation make twitter a rich source of up-to-date information on different events and a good indicator of users' moods.

V. MACHINE LEARNING TECHNIQUES

The Major data mining techniques used to extract the knowledge and information are: Classification, Clustering, Generalization, Data visualization, Association rule mining, neural networks, Bayesian networks, Fuzzy logic, and Genetic algorithm, decision tree. Machine learning techniques are most powerful techniques used in opinion mining for categorizing documents or sentences into positive, negative or neutral categories. Machine learning techniques are classified into two basic categories.

1) Supervised Machine Learning:

Classification is most frequently used popular data mining technique [19]. Classification used to predict the possible outcome from given data set on the basis of defined set of attributes and a given predictive attributes. The given dataset is found to be the training dataset consist on independent variables (dataset related properties) and a dependent attribute. A training dataset created model test on test corpus contains the same attributes but no predicted attribute.

2) Unsupervised Learning:

Unsupervised learning has no explicit targeted output associated with input in contrast of supervised learning. Class label for any instance is unknown so unsupervised

All rights reserved by *www.ijsrd.com*

learning is about to learn by observation. Unsupervised machine learning techniques do not use training data set for classification. Various Clustering algorithms like K-means clustering, Hierarchical clustering used to classify data into categories. The process of gathering objects of similar characteristics into a group is called clustering. Semantic Orientation also provides to generate accurate result for classification.

The machine learning approach applicable to sentiment analysis mostly belongs to supervised classification in general and text classification techniques in particular. Machine learning techniques like Naive Bayes (NB), support vector machines (SVM) and maximum entropy (ME) have achieved great success in text categorization. Some other most well-known machine learning methods in the natural language processing area are K-Nearest neighbourhood, C5, ID3, centroid classifier, the N-gram model and winnow classifier.

A. Naive Bayes:

The Naive Bayes classifier is based on Bayesian probability and assumed that feature probabilities are independent of one another. One approach to text classification is to assign to a given document d the class c^*. Where,

$$c^* = arg\ max_c\ P(c \mid d) \qquad (1.1)$$

Naive Bayes is optimal for certain problem classes with highly dependent features [5]. A list comprising the existing work done using Naive Bayes classifier is shown in Table 1.

S. No.	Data Source	Performance (Accuracy)	References
1	Tweets	71.88%	[6]
2	Microblogs	70.39%	[7]
3	Cantonese reviews	93%	[8]
4	Movie reviews	85.8%	[9]
5	Microblogs	82.5%	[10]
6	Microblogs	82.7%	[11]
7	Reviews	77.5%	[12]
8	Chinese sentiment corpus	85.58%	[13]
9	Reviews	77.5%	[14]
10	Car reviews	86%	[15]
11	Reviews	81.9-87.0%	[16]
12	Reviews	81.5%	[17]

Table 1: List of existing work done using Naive Bayes classification approach

B. Maximum Entropy:

Maximum entropy classification (MaxEnt, or ME, for short) is an alternative technique which has proven effective in a number of natural language processing applications. Maximum Entropy Classifier makes no conditional independence assumption between features, as the Naive Bayes classifier does. Maximum Entropy Classifier provides the least biased estimate possible based on the given information. Its estimate of $P(c \mid d)$ takes the following exponential form:

$$P_{ME}(c \mid d) = \frac{1}{Z(d)} exp(\sum_i \lambda i, cFi, c(d, c)) \qquad (1.2)$$

Where $Z(d)$ is a normalization function and $F_{i,c}$ is a feature/class function for feature f_i.

The classification system is well described as: "Maximum Entropy models offer a way to combine diverse pieces of contextual evidence in order to estimate the probability of a certain linguistic class occurring with a certain linguistic context….in which task is to estimate the probability of class 'a' occurring with context 'b' ".

A list comprising the existing work done using Maximum Entropy classifier is shown in Table 2.

S. No.	Data Source	Performance (Accuracy)	References
1	Tweets	71.88%	[6]
2	Movie review	85.4%	[9]
3	Microblogs	83%	[11]
4	Reviews	77.1%	[18]
5	Reviews	81.0%	[17]

Table 2: List of existing work done using Maximum Entropy classification approach

C. Support Vector Machines:

In contrast to Naive Bayes and Maximum Entropy Classifier, Support Vector Machines are large-margin, rather than probabilistic classifiers. Support Vector Machines (SVMs) operate by separating points in a d-dimensional space using a (d-1)-dimensional hyperplane. In a nutshell, a support vector machine (or SVM) is an algorithm that works as follows. Firstly, it uses a nonlinear mapping to transform the original training data into a higher dimension. Inside this new dimension, it searches for the linear optimal separating hyperplane (that is, a "decision boundary" separating the tuples of one class from another). Data from two classes can always be separated by a hyperplane with an appropriate nonlinear mapping to a sufficiently high dimension. The SVM finds this hyperplane using support vectors ("essential" training tuples) and margins (defined by the support vectors)[19]. By the practical point of view, it is usually not possible to find a hyperplane that separates the classes perfectly, so points are permitted to be inside the margin or on the wrong side of the hyperplane. Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization.

Classification of test instances consists simply of determining which side of ω's hyperplane they fall on [18]. A list comprising the existing work done using Support Vector Machine classification method is shown in Table 3.

S. No.	Data Source	Performance (Accuracy)	References
1	Tweets	82.2%	[6]
2	Movie review	86.4%	[9]
3	Microblogs	71.56%	[11]
4	Reviews	77.4%	[12]
5	Chinese sentiment corpus	86.64%	[13]
6	Reviews	84.6%	[14]
7	Movie reviews	86.2%	[20]
8	Reviews	82.9%	[17]

Table 3: List of existing work done using Support Vector Machine classification approach

D. K-Nearest Neighbor:

K-nearest neighbor (KNN) is a simple machine learning algorithm. Nearest-neighbor classifiers are based on learning by analogy, i.e., by comparing a given test tuple with

All rights reserved by *www.ijsrd.com*

training tuples that are similar to it [22]. In this algorithm, the objects are classified based on the majority of its neighbor. The training tuples are described by n attributes. Where, each tuple represents a point in an n-dimensional space. In this manner, all of the training tuples are stored in an n-dimensional pattern space. While given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. All these k training tuples are the k "nearest neighbors" of the unknown tuple. In KNN, selection is based on majority voting or distance weighted voting [18]. "Closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, $X_1 = (x_{11}, x_{12},...., x_{1n})$ and $X_2 = (x_{21}, x_{22},...., x_{2n})$, is

$$dist(X_1, X_2) = \sqrt{\Sigma (x_{1i}-x_{2i})^2} \quad (1.3)$$

A list comprising the work done using K-nearest neighbor classification approach is shown in Table 4.

S. No.	Data Source	Performance (Accuracy)	References
1	Chinese sentiment corpus	86.64%	[13]
2	Tweets	66.0-87.0	[21]

Table 4: Enlist of existing work done using K-nearest neighbor classification approach

VI. Opinion Mining Tasks

Usually people think that opinion mining is all about to predict the polarity of the text as positive or negative. But it's not true. It is much broader task. The Tasks included are:

A. Subjectivity Detection:

This task includes whether the text contains opinions or not (subjective or objective).

B. Sentiment Prediction:

This task includes the prediction of polarity of the text i.e. whether the text is positive or negative. It can be at document level, sentence level and at aspect level.

C. Aspect Based Sentiment Summarization:

This task includes the discovery of features of products and then discovering the sentiments for each feature. Such as, for Samsung mobile you may have features like design, sound, screen etc.

D. Contrastive Viewpoint Summarization:

This task includes the contradicting opinions .For Example, some peoples like the policies of some particular company and some may not. So this creates contrastive viewpoints of people.

E. Text Summarization for Opinions:

This task includes the useful summary format to generate textual summaries .Example: a few sentences summarizing the reviews of a product.

F. Product Feature Extraction:

This task includes the extraction of features of products. It includes the following steps:

- Identify product features.
- Identify opinions regarding product features.
- Determine the polarity of opinions.
- Rank opinions based on their strength.

G. Opinion Retrieval:

This task includes the retrieval of documents and ranking according to their opinions about a query topic.

H. Opinion-Based Entity Ranking:

All the candidate entities are ranked that are based on how well opinions on these entities match the user's preferences.

VII. Commonly Used NLP Tools for Opinion Mining

A variety of open-source text-analytics tools like natural-language processing for information extraction and classification can be applied for sentiment analysis.

- OpenNLP
- LingPipe
- Part-of-Speech (POS) Tagger
- NTLK
- Tawlk/osae
- Opinion Finder
- GATE
- NLP Tool suite
- Textir

VIII. Conclusion

This review article focuses on the tasks, resources and techniques involved in opinion mining and sentiment analysis along with some machine models being used to implement it. Although a lot of research has been performed in past few years on its wide area but due to lack of availability of corpus we mainly concentrated on its lexical resource and machine learning techniques. From the literature survey, it is also found that different types of features and classification algorithms are combined in an efficient way in order to overcome their individual drawbacks and finally enhance the sentiment classification performance.

References

[1] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, pp. 1-167, May 2012.
[2] C. Hsinchun and D. Zimbra, "AI and Opinion Mining," Intelligent Systems, IEEE, vol. 25, pp. 74-80, 2010.
[3] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," presented at the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA, 2006.
[4] D. E. O'Leary, "Blog mining-review and extensions: "From each according to his opinion"," Decision Support Systems, vol. 51, pp. 821-830, 2011.
[5] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," Machine Learning, vol. 29, pp. 103-130, 1997.
[6] T. Trilla and F. Alias, "Sentence-Based Sentiment Analysis for Expressive Text-to-Speech," Audio,

All rights reserved by *www.ijsrd.com*

Speech, and Language Processing, IEEE Transactions on, vol. 21, pp. 223-233, 2013.
[7] L. Po-Wei and D. Bi-Ru, "Opinion Mining on Social Media Data," in Mobile Data Management (MDM), 2013 IEEE 14th International Conference on, 2013, pp. 91-96.
[8] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of Internet restaurant reviews written in Cantonese," Expert Systems with Applications, vol. 38, pp. 7674-7682, 2011.
[9] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences, vol. 181, pp. 1138-1152, 2011.
[10] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," presented at the Proceedings of the 13th international conference on Discovery science, Canberra, Australia, 2010.
[11] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, pp. 1-12, 2009.
[12] M. Annett and G. Kondrak, "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs," in Advances in Artificial Intelligence. vol. 5032, S. Bergler, Ed., ed: Springer Berlin Heidelberg, 2008, pp. 25-35.
[13] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents," Expert Systems with Applications, vol. 34, pp. 2622-2629, 2008.
[14] C. Chen, F. Ibekwe-SanJuan, E. SanJuan, and C. Weaver, "Visual Analysis of Conflicting Opinions," in Visual Analytics Science And Technology, 2006 IEEE Symposium On, 2006, pp. 59-66.
[15] M. Gamon and A. Aue, "Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms," presented at the Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing, Ann Arbor, Michigan, 2005.
[16] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," presented at the Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, 2003.
[17] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, 2002.
[18] K. Shimada and T. Endo, "Seeing Several Stars: A Rating Inference Task for a Document Containing Several Evaluation Criteria," in Advances in Knowledge Discovery and Data Mining. vol. 5012, T. Washio, E. Suzuki, K. Ting, and A. Inokuchi, Eds., ed: Springer Berlin Heidelberg, 2008, pp. 1006-1014.
[19] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques: Morgan Kaufmann Publishers Inc., 2011.
[20] A. Kennedy and D. Inkpen, "Sentiment Classification Of Movie Reviews Using Contextual Valence Shifters," Computational Intelligence, vol. 22, pp. 110-125, 2006.
[21] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using Twitter hashtags and smileys," presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China, 2010.

All rights reserved by *www.ijsrd.com*