IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-3, Issue-3 E-ISSN: 2347-2693

Different Approaches of Sentiment Analysis

Supriya B. Moralwar[1], Sachin N. Deshmukh[2]

[1,2]Department of computer science and IT, Dr. B.A.M. University
Aurangabad-431004, India
sbmoralwar@gmail.com, sndeshmukh@hotmail.com

www.ijcseonline.org

Received: Feb/23/2015 Revised: Mar/01/2015 Accepted: Mar/16/2015 Published: Mar/31/2015

***Abstract*—** Sentiment analysis is a machine learning approach in which machines analyzes and classifies the sentiments, emotions, opinions about any particular topics or entity which are expressed in the form of text or speech. Due to large volume of textual data increasing on the web so much of the current research is focusing on the area of sentiment analysis. People are interested to develop and design a system that can identify and classify the sentiments as represented in textual form. Sentiment analysis is used to extract the subjective information in source material by applying various techniques such as Natural language Processing (NLP), Computational Linguistics and text analysis and classify the polarity of the opinion. In this paper, we are going to discuss different levels of sentiment analysis, approaches for sentiment classification, Data Source for sentiment analysis and comparative study of approaches for sentiment classification.

***Keywords*—** Sentiment Analysis, Opinion Extraction, Text Mining, Natural Language Processing, Subjective Analysis, Machine Learning Algorithm.

I. INTRODUCTION

OPINION mining (often referred as Sentiment Analysis) refers to identification and classification of opinion expressed in the text span; using information retrieval and computational linguistics [1].

Sentiment analysis is an emerging discipline whose aim is to get real voice or opinion of people towards specific product, persons, services, organizations, news, events, issues and their attributes. It is used to indentify positive, negative or neutral opinions, emotions and evaluations. Sentiment analysis is performed by using techniques like Natural Language Processing (NLP), Machine Learning, Text Mining and Information Theory and Coding, Semantic Approach. By using this approaches, methods, techniques and models, we can categorized our data which is unstructured data may be in form of news articles, blogs, tweets, product reviews etc. into positive, negative or neutral sentiment according to the sentiment is expressed in them. If data is present in huge amount, then there is need for preprocessing.

Now a day's, very large amounts of information are available in on-line documents. As part of the effort to better organize this information for users, researchers have been actively investigating the problem of automatic text categorization. The bulk of such work has focused on topical categorization, attempting to sort documents according to their subject. Now, recent years have shown rapid growth in online discussion groups and review sites where a crucial characteristic of the posted articles is their sentiment.

For example, whether a product review is positive or negative labeling these articles with their sentiment would provide succinct summaries to readers; indeed, these labels are part of the appeal which both labels movie reviews that do not contain explicit rating indicators and normalizes the different rating schemes that individual reviewers use. Sentiment classification would also be helpful in business intelligence applications and recommender systems where user input and feedback could be quickly summarized; in-deed, in general, free-form survey responses given in natural language format could be processed using sentiment categorization.

II. LEVELS OF SENTIMENT ANALYSIS

A. *Document Level Sentiment Analysis*

The Document Level Sentiment analysis is performed for whole document [2]. The basic unit of information is a single document of opinionated text. In this type of document level classification a single review about a single topic is considered. But in case of forums or blogs, comparative sentences may appear and customers may compare one product with the other that has similar characteristics and hence document level analysis is not desirable in forums and blogs. While doing document level classification, irrelevant sentences must be eliminated at preprocessing phase. For document level classification both supervised and unsupervised machine learning classification methods are used. Supervised machine learning algorithm such as Support Vector Machine (SVM), Naïve Baye's, KNN and Maximum Entropy can be used to train the system. For training and testing dataset, the reviewer rating (in the form of 1-5 stars) and review text can be used. The features that can be used for the machine learning are term frequency, document frequency, tf-idf measure, Part of speech tagging, Opinion words, opinion phrases, negations and dependencies. Manually labeling the polarities of the document is time consuming task and hence the user rating available can be made use of. The unsupervised machine learning can be done by extracting the opinion words inside a document. The point-wise mutual

information [3] can be made use of to find the semantics of the extracted words.

B. Sentence Level Sentiment Analysis

The Sentence level sentiment analysis is related to find sentiment form different sentences whether the sentence expressed is positive, negative or neutral sentiment. The Sentence level sentiment analysis is closely related to subjectivity classification. Here, the polarity of each sentence is calculated and then same document level classification methods are used for the sentence level classification problem. Then the objective and subjective sentences must be found out. The subjective sentences must contain opinion words which help in determining the sentiment about entity. After that the polarity classification is done into positive, negative and neutral classes [3].

C. Entity or Aspect Level Sentiment Analysis

The Entity or Aspect Level sentiment analysis performs finer-grained analysis. The goal is to find out the sentiment on entities or aspect of those entities. For example consider a statement "My Nokia Lumina 510 cell phone has good picture quality but it has less battery backup." So the opinion on Nokia's camera and display quality is positive but the opinion on its cell phone battery backup is negative. We can generate summery of opinions about entities. Comparative statements are part of the entity or aspect level sentiment analysis but deal with techniques of comparative sentiment analysis.

D. Phrase Level Sentiment Analysis

In phrase level sentiment classification, the phrases that contain opinion words are found out and a phrase level classification is done. This is advantageous or may be disadvantageous. It is advantageous where the exact opinion about an entity can be correctly extracted. But in other cases, where contextual polarity matters, so result may not be accurate. So the negation of words can occur locally. In such cases, this type of sentiment analysis suffices [3].

E. Feature Level Sentiment Analysis

Product features are considered as product attributes. Analysis of these features for identifying sentiment of the document is called as feature based sentiment analysis. In this approach positive, negative or neutral opinion is identified from the extracted features. It is the fine grained analysis model among all other model [4, 5].

III. DATA SOURCE

A. Review Sites

A review site is a website where users can post reviews, which give an opinion about people, businesses, products, services and particular entity. Most of the sentiment analysis work has been done on movie and product review sites [6]. The review data used in most of the sentiment classification studies are collected from the e-commerce websites like www.amazon.com (product reviews), www.yelp.com (restaurant reviews), www.CNET download.com (product reviews) and www.reviewcentre.com, which have millions of product reviews by customer. Other than these reviews the available are professional review sites such as www.dpreview.com, www.zdnet.com and customer opinion sites on broad topics and products such as www .consumerreview.com,www.epinions.com, www.bizrate.com [7].

B. Blogs

With an increasing usage of the internet, blogging and blog posts are growing rapidly [7]. The term blog refers to a webpage consisting of brief paragraphs of opinion, information, personal diary entries, or links, called posts, which are arranged chronologically with the most recent first, in the style of an online journal [8]. Sentiment analysis on blogs [9] has been used to predict the product sales, movie sales, political mood and in many of the studies related to sentiment analysis.

C. Forums

Forums or message boards allow its members to hold conversations by posting it on the site. These are dedicated to a topic and thus using forums as a database allows us to do sentiment analysis in a single domain.

D. Datasets

Most of the work in the field uses movie reviews and product reviews data for classification. The movie review dataset which is available online is (http:// www.cs.cornell.edu/People/pabo/movie-review-data) [7]. Other dataset which is available online is multi-domain sentiment (MDS) dataset (Blitzer et. al., 2007) (http://www.cs.jhu.edu/mdredze/datasets/sentiment) [10]. Is ; Zhu Jian ,2010 ; Pang and Lee ,2004; Bai et al. ,2005; Kennedy and Inkpen ,2006; Zhou and Chaovalit ,2008; Yulan He 2010; Rudy Prabowo ,2009; Rui Xia ,2011) [7].

E. Micro-blogging

Twitter is a popular micro-blogging service where users create or write status messages called "tweets". These tweets express opinions about different topics. Tweets are also used as data source for classifying sentiment.

F. Google Play Android Application Store

Google Play Android Application Store has a large and variety of collections of Android Applications with

rankings and user reviews. It extracts textual reviews having rich content from the App Store site [11].

IV. APPROACHES FOR SENTIMENT CLASSIFICATION

A. *Natural Language Processing*

It is the branch of computer science and technology which focused on developing systems that allow computers to communicate with people using natural language. Natural language processing technique plays important role to get accurate sentiment analysis. NLP techniques like Bag of words, Hidden markov model (HMM), part of speech (POS), N-gram algorithms, large sentiment lexicon acquisition and parsing techniques are used to express opinion for document level, phrase level, sentences level and aspect level [12,13].

Large sentiment lexicon acquisition is used sentiment word dictionary which contains lot of sentiment words with their numeric threshold value for particular domain [14]. SentiWordNet dictionary is used for subjective sentiment analysis. Part-of speech (POS) tagging is often the most time consuming and challenging task before doing sentiment analysis of any documents. As online textual reviews are short, non-grammar sentences and contain slangs, abbreviations, and symbols which make the POS tagging even more difficult. For example, consider the statement. “The camera is good. I love its picture quality.” Here, “camera” is referred as a product and “picture quality” is referred as a feature. We know, Products and features are tagged as nouns. We can define the synonym list of products and features. This feature can be because of uncertain and non-grammar online reviews. For example, consider the following comment. “I like the high res”. Here “res” refers to resolution, and resolution is similar to graphics. Sometimes textual reviews may contain mixture sentiment. For example, “I like the graphics, but it takes battery a lot”. Now we are doing feature based sentiment analysis, so it is easy to handle such reviews. In this case, the sentiment is positive for “graphics” and negative for “battery”. For this CLASSIFIER, CONCEPT, CONCEPT_RULE, and PREDICATE_RULE rules can be used [6].

B. *Machine Learning Techniques*

Machine learning techniques are most useful techniques for the sentiment classification for categorized text into positive, negative or neutral categories. in machine learning technique, training and testing datasets are required. A training dataset is used to learn the documents and test dataset is used to validate the performance. There are number of machine learning algorithms used to classify reviews.

There are two types of machine learning techniques such as supervised machine learning algorithm like maximum entropy, SVM, Naïve bayes, KNN, etc and unsupervised machine learning algorithm such as HMM, Neural network, PCA, ICA, SVD, etc.

i. Naïve Bayes

Naïve bayes is a simple and easy but effective classification algorithm. It is mostly used for document level classification. The basic idea is to calculate the probabilities of categories given a test document by using the joint probabilities of words and categories. Naive Bayes is optimal for certain problem classes with highly dependent features.

Naive Bayes classifiers are computationally fast when taking decisions. It does not require large amounts of data before learning can begin [15].

ii. Support Vector Machine

SVM is a discriminative classifier considered as the best text classification method. It is a statistical classification method proposed by Vapnik. SVM maps input (real-valued) feature vectors into a higher-dimensional feature space through some nonlinear mapping. SVMs are developed on the principle of structural risk minimization. The structural risk minimization seeks to find a hypothesis (h) for which one can find lowest probability of error whereas the traditional learning techniques for pattern recognition are based on the minimization of the empirical risk, which are attempt to optimize the performance of the learning set. Computing the hyper plane to separate the data points i.e. training an SVM leads to a quadratic optimization problem. SVMs can learn a larger set of patterns and able to scale better, because of classification complexity it does not depend on the dimensionality of the feature space. SVM have the ability to update the training patterns dynamically whenever there is a new pattern during classification [16].

iii. k-Nearest Neighbor

KNN is a classifier that relies on the category labels attached to the training documents similar to the test document. It is a method to classify an object based on the majority class amongst its k-nearest neighbors. It is a type of lazy learning where the function is only approximated locally and all computation is deferred until classification [17].

KNN algorithm usually uses the Euclidean or the Manhattan distance. However, any other distance such as the Chebyshev norm or the Mahalanob is distance can also be used [18].

iv. Winnow

It is a well-known online mistaken-driven technique. It works by updating its weights in a sequence of trials and on each trial; it first makes a prediction for one document and then receives feedback. If a mistake is made, then it updates its weight vector using the document. During the training phase, with a collection of training data we can

process it repeatedly several times by iterating on the data [6].

v. Maximum Entropy

Maximum Entropy (MaxEnt) classification is a technique which has proven effective in a number of natural language processing applications. It doesn't make any assumptions about the relationships between features, so might potentially perform better when conditional independence assumptions are not met.

The parameter values are set so as to maximize the entropy of the induced distribution subject to the constraint that the expected values of the feature/class functions with respect to the model are equal to their expected values with respect to the training data: the underlying philosophy is that we should choose the model making the fewest assumptions about the data while still remaining consistent with it, which makes intuitive sense [19].

vi. *Association Rule Learning*

In association rule learning the association is find out between different variables. Variables are having different values so its changes never be constant. This type of relationship can be used while using huge database where organized collection of data is present [20].

C. Decision Tree Learning

Decision Tree Learning is a tree based approach, where collection of child and root node which focus on the target value. It is a flow chart like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent child node or class distributions [21]. The popular Decision Tree algorithms are ID3, C4.5 and CART. The ID3 algorithm is considered as a very simple decision tree algorithm. It uses information gain as splitting criteria. C4.5 is an evolution of ID3. It uses gain ratio as splitting criteria [22]. The CART algorithm uses Gini coefficient as the test attribute for selection criteria, and each time selects an attribute with the smallest Gini coefficient as the test attribute for a given set [23].

D. Techniques of Information theory and Coding [24]

The concept of mutual information (MI), Residual Inverse Document Frequency (RIDF), TF-IDF and random process are also used for sentiment analysis and its classification.

E. Semantic Orientation Approach

The Semantic orientation approach to Sentiment analysis is "unsupervised learning" because it does not require prior training in order to extract the data. Instead, it measures how far a word is inclined towards positive and negative.

F. Hybrid Approaches

In Hybrid approach, we can combine any of the above approaches or techniques as and when needed for efficient sentiment analysis it is a Hybrid Approach.

V. COMPARATIVE STUDY OF SENTIMENT ANALYSIS TECHNIQUES

An unsupervised classification method for extracting aspects and determining sentiment in review text is designed. This method is simple and flexible for any domain and language. He introduces a local topic model, which works at the sentence level and employs a small number of topics that automatically infer the aspects [25].

An approach to extract product features and to classify the sentiment associated with these product features from the reviews through syntactic information [26].

He uses an automated consumer review agent for collecting and creating review models [27]. Classification, clustering, summarization etc are used as machine learning technique for performing analysis.

A probabilistic topic model is utilized to capture the mixture of aspects and sentiments simultaneously [28].

A mutual reinforcement strategy is designed to cluster product aspects and opinion words by iteratively fusing both content and sentiment link information [29].

VI. CONCLUSION

Sentiment analysis is also known as opinion mining or opinion extraction. Sentiment analysis is helpful in different field for calculating, identifying and expressing sentiment. This paper illustrates the research area of Sentiment Analysis on reviews on product like amazons, android apps and its latest advances. It affirms the levels of sentiment classification, data source for review collection, and Approaches for sentiment classification. Most work has been done on product reviews downloaded from Amazon.

ACKNOWLEDGMENT

The author will like to thank the university authorities and department of computer science and information technology Dr. B.A.M.U Aurangabad for providing the infrastructure to carry out the work. This work is supported by university commission.

REFERENCES

[1] A. Nisha, Jebaseeli, E. Kirubakaran, PhD., "A Survey on Sentiment Analysis of (Product) Reviews", International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012

[2] Jalaj S. Modha, Prof & Head Gayatri S. Pandi Sandip J. Modha, "Automatic Sentiment Analysis

for Unstructured Data", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 12, ISSN: 2277 128X, December 2013.

[3] Raisa Varghese1, Jayasree M2, "A SURVEY ON SENTIMENT ANALYSIS AND OPINION MINING", IJRET:International Journal of Research in Engineering and Technology ISSN: 2319-1163 | ISSN: 2321-7308.

[4] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, "OPINION MINING AND ANALYSIS: A SURVEY", International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June 2013.

[5] Zhongwu Zhai, Bing Liu, Hua Xu and Hua Xu, "Clustering Product Features for Opinion Mining", WSDM'11, February 9–12, 2011, Hong Kong, China. Copyright 2011 ACM 978-1-4503-0493-1/11/02...$10.00

[6] Siddhi Patni, Avinash Wadhe, "Review Paper on Sentiment Analysis is – Big Challenge", International Journal of Advance Research in Computer Science and Management Studies Volume 2, Issue 2, ISSN: 2321-7782 (Online), February 2014 .

[7] G.Vinodhini, RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 6, ISSN: 2277 128X, June 2012 .

[8] Anderson, P., "What is Web 2.0? Ideas, technologies and implications for education", Technical report, JISC, 2007.

[9] Mishne G. and Glance N., "Predicting movie sales from blogger sentiment", In AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW), 2006: 155–158.

[10] Maria Tchalakova, Dale Gerdemann, Detmar Meurers, "Automatic Sentiment Classification Of Product Reviwes Using Maximal Phrases Based Analysis", Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 111-117, Portland, Oregon, USA 2011 Association for Computational Linguistics, 24 June 2011,.

[11] Jiawen Liu, Mantosh Kumar Sarkar and GoutamChakraborty, "Feature-based Sentiment Analysis on Android App Reviews Using SAS® Text Miner and SAS® Sentiment Analysis Studio", SAS Global Forum 2013.

[12] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan and Claypool Publishers, p.18-19, 27-28, 44-45, 47, 90-101, May 2012.

[13] Nitin Indurkhya, Fred J. Damerau, "Handbook of Natural Language Processing", Second Edition, CRC Press, 2010.

[14] Ronen Feldman, "Techniques and Application of Sentiment Analysis", Communication of ACM, vol. 56.No.4, April 2013.

[15] Ahmad Ashari, Iman Paryudi, A Min Tjoa, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11, 2013.

[16] Ajayi Adebowale, Idowu S.A, Anyaehie Amarachi A., "Comparative Study of Selected Data Mining Algorithms Used For Intrusion Detection", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-3, July 2013.

[17] Wikipedia, "k-Nearest Neighbor Algorithm," Availableat:http://en.wikipedia.org/wiki/K_nearest_neighbor_algorithm.

[18] V. Garcia, C. Debreuve, "Fast k Nearest Neighbor Search using GPU", IEEE, 2008.

[19] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proceedings of EMNLP 2002, pp. 79-86, 2002.

[20] Abdullah Dar*, Anurag Jain, "Survey paper on Sentiment Analysis: In General Terms", International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-3, Issue-11).

[21] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan-Kaufmann Publishers, San Francisco, 2001.

[22] O. Maimon and L. Rokach, "Data Mining and Knowledge Discovery", Springer Science and Business Media, 2005.

[23] X. Niuniu and L. Yuxun, "Review of Decision Trees", IEEE, 2010.

[24] Jintao Mao and Jian Zhu, "Sentiment Classification based on Random Process", IEEE Computer Society, International Conference on Computer Science and Electronics Engineering, p.473-476, 2012.

[25] S. Brody, N. Elhadad, "An unsupervised aspect-sentiment model for online reviews", in: Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics. Publishing, Association for Computational Linguistics, pp. 804-812, 2010.

[26] Pimwadee Chaovalit, Lina Zhou, "Extracting Product Features and Opinions from Product Reviews Using Dependency Analysis", Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, Shandong, pp. 2358-2362, 2010.

[27] Pollach, I., "Automating user reviews using ontologies: an agent-based approach", Springer Journal on World Wide Web, Vol. 15, No.3, pp. 285-323, 2012.

[28] Q. Mei, X. Ling, M. Wondra, H. Su, and C. X. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs", in *Proc. 16th Int. Conf. WWW*, Banff, AB, Canada, pp. 171–180, 2007.

[29] Q. Su *et al*., "Hidden sentiment association in Chinese web opinion mining", in *Proc. 17th Int. Conf. WWW*, Beijing, China, pp. 959–968, 2008.

AUTHORS PROFILE

Supriya B. Moralwar is pursing Masters in Computer Science and Engineering from Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad -431001, Maharashtra India.
sbmoralwar@gmail.com

Assistant Prof. Sachin N. Deshmukh
Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad -431001, Maharashtra, India.
sndeshmukh@hotmail.com