ISSN: 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Sentiment Analysis for Web-based Big Data: A Survey

Sufal Das
Department of Information Technology
North-Eastern Hill University
Shillong India

Hemanta Kumar Kalita
Department of Information Technology
North-Eastern Hill University
Shillong India

Abstract: These days, customers rely on the Web for the opinions on various products and services. Due to rapid increase of data availability in web, it is very difficult to manage these for an application. Again, these data are in heterogeneous formats as well as rapid changing in nature. Thus there is a need for an effective system to classify analyse web reviews as a big data analysis. This task of classifying and analyzing such collective web data together is known as opinion mining, and it is also known as sentiment analysis. Sentiment analysis is a very challenging and promising discipline which uses both intersection of information retrieval and computational linguistic techniques to deal with the reviews expressed in a source material [1]. This work talks about the sentiment analysis process and focus on some machine learning techniques for sentiment classification and future challenges in opinion mining for big data.s

Keywords: Web Mining, Opinion Mining, Sentiment Classification, Text Classification, Big Data Analysis

I. Introduction

In today’s world users can post their comments on any internet forums, review sites, blogs and discussion group which are commonly known as user generated content which contains the important information. This online word-of-mouth behavior represents new and considerable sources of information and their applications. These online comments are expanded on a global or web scale. Users can view reviews on a particular product, but the large scale reviews make it almost to impossible to read them all. So Opinion mining helps classify and analyze these reviews and opinionated texts and produce a summary that the user can look into efficiently [2, 3].

Sentiment analysis is also known as opinion mining and refers to the use of natural language processing (NLP), text analysis and computational linguistics to identify and extract subjective information in source materials [4, 5]. Textual information can be facts or opinions. The facts are the objective expressions which describe the entities, events and properties whereas the opinion is the subjective expression which describes people's opinions, emotions and sentiments towards entities and their properties. The current search engine searches for facts and can be expressed with keywords unlike opinions that cannot be expressed with keywords and ranking cannot be done. The art Opinion Mining is to recognize the subjectivity and objectivity of a text and further classify the opinion orientation of subjective text. It used Natural Language Processing and Machine Learning ethics to determine opinion in the text. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level positive, negative, or neutral. Spam filtering is also required and refers to detection and removal of fake opinions that mislead the users by giving unworthy positive or negative opinions to some objects in order to sponsor or spoil the objects reputations respectively. It is also a research issue. There are lots of Free and Open Source tools available for performing Natural Language Processing and Machine Learning tasks [3]. Major applications of sentiment analysis are in financial markets. For investors is important information the analysts and other investors’ opinions about the stocks of a company, to identify price trends. It can be useful where a company is interested in customers' perceptions about its products. Information may be used to improve products and identifying new marketing strategies. Sentiment analysis is also used by tourists to know the best places to visit or famous restaurants. Applying sentiment analysis can be obtained relevant information for planning a trip. In politics sentiment analysis can be used on elections. Using sentiment analysis we can identify the voter opinions about a certain candidate. Movies or software programs reviews can detect users' sentiments.

II. Background Study

Big Data Charateristics

Every day, we create 2.5 quintillion bytes of data so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, post to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This is big data. The data sets so large or complex that traditional data processing applications are inadequate.

The Big data [6] can be described in 4 V’s: Volume, Velocity, Veracity and Variety.

Volume: It relates to the quantity of data that is generated everyday in large scale and its size is increasing continuously.

Velocity: This term refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and challenges which lie ahead in the path of growth and development.

Veracity: It refers to the quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

Variety: There are different types of data and data sources available and all these refer to this term. The availability of information to analyze is of different types, such as mainly coming from social media and communication devices. The term ‘variety’ includes structured data like tabular data (databases), transactions etc. and unstructured and semi-structured data like hierarchical data, documents, e-mail, video, images, audio etc.

Sentiment Analysis/ Opinion Mining

Sentiment analysis is a classification method based on the words included in each social post. This classification has mainly three class labels as positive, negative and neutral sentiment. It is also known as opinion mining [7].

There are numerous application areas of sentiment analysis including

a. Product/services research for marketing purposes,
b. Better search engines,
c. Trend monitoring (e.g. social, cultural, political etc.)
d. Recommendation systems.

An opinion/sentiment is quintuple (o, f, so, h, t)

a. o- target object.
b. f-feature of object o.
c. so-sentiment value of o, on feature f, by holder h, at time t.
d. h-opinion holder.
e. t-time at which opinion is expressed.

Objective: Given an opinionated document, discover all quintuples (o, f, so, h, t), unstructured text can be converted to structured data and will enable opinion mining.

Opinions may be of two types:

a. Regular opinion: eg. Coke tastes very good.
b. Comparative opinion: eg. Coke tastes better than pepsi.

Generally, Opinion Mining can be applied in three different levels for text mining [7, 8].

Document Level: At this level, the task is to classify whether a whole opinion document expresses a positive or negative sentiment. It assumes that each document expresses opinions on a single entity or product. Thus, it is only applicable to documents describing a single entity or product.

Sentence Level: At this level, the task goes to the sentences that determine whether each sentence expressed a positive, negative, or neutral opinion. At this level analysis of the sentence to classify it as objective or subjective is done. Many objective sentences can imply opinions but it is very difficult to determine some complex subjective sentences and determine their sentiment. Researchers have also analyzed clauses, but the clause level is still not enough.

Feature-based Level: There are three tasks involved. Firstly, identification of features is done. Secondly, the sentences involving these features are taken and classified as positive, negative or neutral. Thirdly, a summary of the opinions is produced.

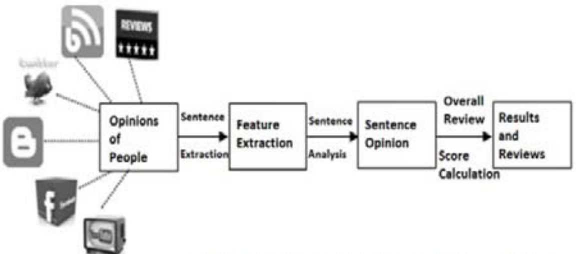

Figure 1. Process of Opinion Mining & Sentiment Analysis

Opinion Extraction: Opinions are extracted by crawlers. A crawler is a small program that visits websites to get their contents and information [9].

Feature Extraction: Features are usually nouns. Hence, Part-of-Speech tagging is the most common technique for feature extraction. Frequent features can be identified where features on which many people have expressed their opinions. Association mining is used. Infrequent features can also be identified. If a sentence has no frequent features then sentences that have opinion words are taken, and the nearest noun/noun-phrases are chosen as features. Frequently occurring features are given more weightage.

Sentence Opinion: Opinion words are extracted from the sentences that features have been discovered. These opinion words or phrases are then applied to classification techniques. Techniques can be lexicon based or machine learning. Some existing and popular machine learning techniques discussed and compared here.

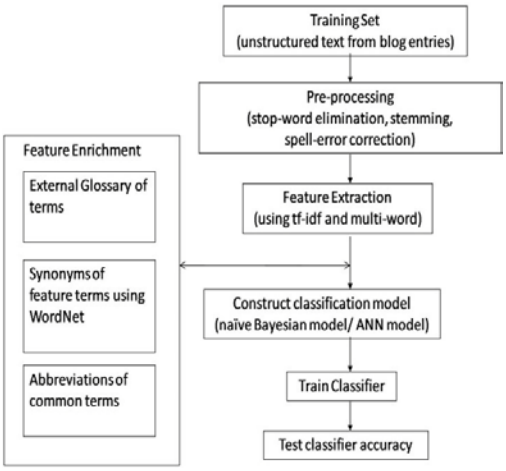
Training Set
(unstructured text from blog entries)
Pre-processing
(stop-word elimination, stemming, spell-error correction)
Feature Enrichment
External Glossary of terms
Synonyms of feature terms using WordNet
Abbreviations of common terms
Feature Extraction
(using tf-idf and multi-word)
Construct classification model
(naïve Bayesian model/ ANN model)
Train Classifier
Test classifier accuracy

Figure 2: Block Diagram for Sentiment Analysis as Text Classifier [9]

Challenges:

a. Very large dataset should be handled.
b. Multilingual- reviews may not always be in English.
c. Sarcasm- sarcasm cannot be detected by opinion mining.
d. Group Synonym words- different words that depict same meaning or feature for a product may not be detected.
e. Detecting emoticons is difficult.
f. Different writing style- abbreviations, poor spelling, poor grammar, poor punctuation.
g. Word ambiguity, i.e. the same word has more than one meaning.
h. Fake reviews like undeserving positive reviews or unfair malicious reviews.

III. FEW MACHINE LEARNING TECHNOQUES FOR SENTIMENT ANALYSIS

Naive Bayes Classifier (NBC)

Naive Bayes classifier is a simple technique for constructing classifier models that assigns class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. The following equation based on conditional probability is user for naive bayes classifier [10], where every feature f_i is independent of every other feature f_j , for $j \neq I$ and P is the probability.

$$P(C \mid f_1, \ldots, f_d) = \frac{P(C)\ P(f_1,\ldots,f_d)|C}{p(f_1,\ldots,f_d)}$$

Support Vector Machine (SVM)

In machine learning, support vector machines [10] are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. It constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

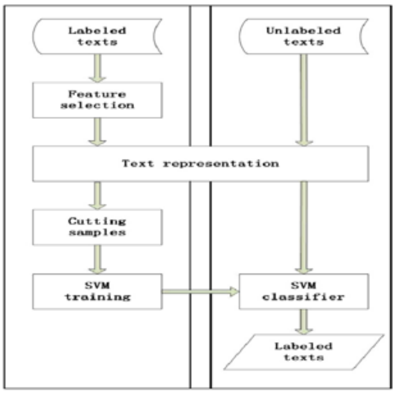

Figure 3: Block Diagram for Sentiment Analysis using SVM

Multilayer Perceptron (MLP)

A multilayer perceptron [11] is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network.MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.

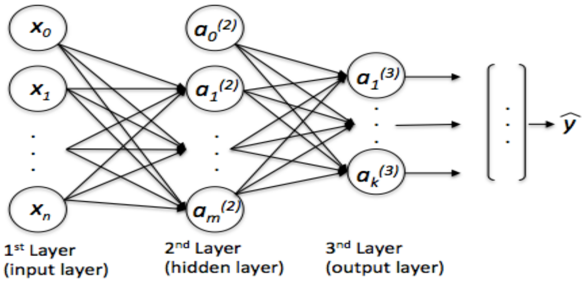

Figure 3: Block Diagram for a Multi-Layer Perceptron

Comparision Amnong NBC, MLP and SVM

Techniques	Advantages	Disadvantages
Naive Bayes Classifier (NBC)	Model is easy to interpret. Eefficient computation.	Assumptions of attributes being independent, which may not be necessarily valid.
Support Vector Machine (SVM)	Very good performance on experimental results. Low dependency on data set dimensionality.	In case of categorical or missing value it needs preprocessed. Difficultly in interpretation of resulting model.
Multilayer Perceptron (MLP)	It acts as a universal function approximation. MLP can learn each and every relationship among input and output variables.	MLP needs more time for execution compare to other technique because flexibility lies in the need to have enough training data. It is considered as complex black box.

IV. Related Works

There are many techniques for sentiment analysis. They may be lexicon-based or machine learning. Lexicon-based uses words from dictionary of a particular language whereas machine learning uses the concept of pattern recognition and computational learning theory in artificial intelligence [12]. Here we have discussed some sentiment analysis techniques.

Bing Liu et al. [13] developed a feature-based opinion summarization system that mines and summarizes all the customer reviews of a product. This system differs from other text summarization systems in number of ways. It mines only the features of the product that customers have expressed their opinions on and the summary generated is structured rather than a free text document as generated by other text summarization systems. The summarization is performed by the system in three steps:

(i) Finding the features on which the customers have expressed their opinions.

(ii) Identifying opinion sentences in each review and analyzing the polarity of the sentence.

(iii) Producing a summary.

N Mishra et al. [14] used techniques that decompose the reviews into segments that evaluate the individual characteristics of a product and then adapt methods from the econometrics literature, specifically the hedonic regression concept, to estimate:

(i) The weight that customers place on each individual product feature.

(b)The implicit evaluation score that customers as sign to each feature, and

(c) How these evaluations affect the revenue for a given product.

Finally, they developed a novel hybrid technique combining text mining and econometrics that models consumer product reviews as elements in a tensor product of feature and evaluation spaces and then impute the quantitative impact of consumer reviews on product demand as a linear functional from this tensor product space. They evaluate our technique using a data set from Amazon.com. They could extract actionable business intelligence from the data and better understand the customer preferences and actions and showed that the textual portion of the reviews can improve product sales prediction compared to a baseline technique that simply relies on numeric data.

M Eirinaki et al. [15] evaluated the effectiveness of different classifiers, and shows that the use of multiple classifiers in a hybrid manner can improve the effectiveness of sentiment analysis in terms of micro and macro-averaged than any individual classifier. This paper combines rule-based classification, supervised learning and machine learning into a new combined method. This method reviews number of automatic document classification techniques. This also includes semi-automatic, complementary approach in which each classifier can contribute to other classifiers to achieve a good level of effectiveness.

I Smeureanu et al. [16] proposed an algorithm for detecting sentiments on movie user reviews, based on naive Bayes classifier. Analysis of the opinion mining domain, techniques used in sentiment analysis and its applicability are done To improve classification they removed insignificant words and introduced in classification groups of words (n-grams).

M. Hu et al. [17] presented a method to summarize all the customer reviews of a product. This summarization task is different from traditional text summarization as there are not only interested in the specific features of the product that customers have opinions on and also whether the opinions are positive or negative. This method summarize the reviews by selecting or rewriting a subset of the original sentences from the reviews to capture their main points as in the classic text summarization.

V. CONCLUSION

Sentiment analysis is technically very challenging but more promising techniques are available, and it will become increasingly important as more people are buying and expressing their opinions on the web. Summarizing the reviews is not only useful to common shoppers, but also crucial to product manufacturers and has wide applications. Since people are interacting through internet, a huge data is being generated every second. Thus, a distributed parallel computing environment is very much needed to perform sentiment analysis efficiently.

VI. REFERENCES

[1] P. K., & Husain, M. S. (2014). Methodological study of opinion mining and sentiment analysis techniques. International Journal on Soft Computing, 5(1), 11.

[2] Sharma, N. R., & Chitre, V. D. (2014). Opinion mining, analysis and its challenges. International Journal of Innovations & Advancement in Computer Science, 3(1), 59-65.

[3] Bhattacharyya, D., Biswas, S., & Kim, T. H. (2010). A review on natural language processing in opinion mining. International Journal of Smart Home, 4(2), 31-38.

[4] Dale, R., Moisl, H., & Somers, H. (Eds.). (2000). Handbook of natural language processing. CRC Press.

[5] Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), 2032-2033.

[6] Agrawal, D., Das, S., & El Abbadi, A. (2011, March). Big data and cloud computing: current state and future opportunities. In Proceedings of the 14th International Conference on Extending Database Technology (pp. 530-533). ACM.

[7] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86).

[8] Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003, November). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 427-434). IEEE.

[9] Remus, R. (2011, May). Improving sentence-level subjectivity classification through readability measurement. In NODALIDA-2011 Conference Proceedings (pp. 168-174).

[10] Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14-46.

[11] Li, S., Wang, Z., Zhou, G., & Lee, S. Y. M. (2011, June). Semi-supervised learning for imbalanced sentiment classification. In IJCAI proceedings-international joint conference on artificial intelligence (Vol. 22, No. 3, p. 1826).

[12] Khairnar, J., & Kinikar, M. (2013). Machine learning algorithms for opinion mining and sentiment classification. International Journal of Scientific and Research Publications, 3(6), 1-6.

[13] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[14] Mishra, N., & Jha, C. K. (2012). Classification of opinion mining techniques. International Journal of Computer Applications, 56(13).

[15] Eirinaki, M., Pisal, S., & Singh, J. (2012). Feature-based opinion mining and ranking. Journal of Computer and System Sciences, 78(4), 1175-1184.

[16] Smeureanu, I., & Bucur, C. (2012). Applying supervised opinion mining techniques on online user reviews. Informatica economica, 16(2), 81.

[17] Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).