Opinion Mining Techniques on Social Media Data

Ritu Mewari
Department of Computer Science
B.C.T.K.I.T, Dwarahat

Ajit Singh, Ph.D
Associate professor
B.C.T.K.I.T, Dwarahat

Akash Srivastava
Assistant professor
CSE Dept. DIT University

ABSTRACT
In the current scenario, at the crossroad of computational linguistics and data retrieval opinions and emotions are more valuable than the subject of the document. Linguistic resources are used to retrieve sentiments and also to classify it. Over the internet, not only the large volume of unstructured data is available but also the large amount of text is also generating day by day in the form of blogs, emails, tweets and feedbacks e.t.c. Text analysis is much more mature than unstructured data. Mining is tough for these types of data because of its noisiness and this is the chief bottleneck for designing text mining system.They suffer from spelling mistakes, grammatical errors and improper punctuations because they are informally written.Opinion mining provides a clear platform to catch public's mood by filtering the noise data. It also provides computational techniques used to extract and consolidate individual's opinion from unstructured and noisy text data.This paper tries to cover some techniques and approaches of opinion mining process and also highlight comparative study of some techniques.

Keywords
Opinion mining, Sentiment analysis, Social Network Sentiment classification , Classification ,machine learning

1. Introduction
Social Network is the chaining of organizations or persons in the real world. We can consider each person in the social network as nodes and can exist in the real world with both implicit(friendship and common interest) and explicit(kinship and classmate). Using traditional data processing applications Big data is very difficult to process because it comprises very large and complex data. With advent of web 2.0, not only the large amount of unstructured data is available but also the large amount of text is also generating day by day on popular micro-blogging sites such as Facebook , Twitter ,Tumblr. Now study of the user data or content of social networks is one of the current trends of the times.

Opinion mining is described as the processing of unstructured data and text data to categorize it into some results like positive, negative and neutral or good , bad and average so that we can predict the product..We can also describe the opinion as the private state of any person by which he/she can express their personal emotions, ideas, assessments, judgment and evaluation about a specific topic. It is the domain of natural language processing and text analytics. It produce the subjective qualities from texual sources. There are mainly 3 tasks of opinion mining process. The first task is to find out the polarity of the text on the basis of sentiment analysis.Second task is related to opinion extraction. The third task is to discover and summarize the opinions.

Components of opinion-
- The opinion holder –source of the opinion
- The object-for which we are expressing the opinions
- View or appraisal-that is the opinion

2. WEB MINING
WWW is large source for data mining. There exist a lot of online sites which contains large volume of information mining. The Web mining research is at the crossroads of research from several research communities such as database, information retrieval, and Artificial Intelligence. [1] As web knowledge is scattered and due to lack of any uniform format, web mining is a difficult task and involves many issues.

Definition (Web Mining)
In web mining process we discover the different patterns by applying data mining techniques.
Web mining is divided into three different types.
1-Web usage mining
2-Web content mining
3- Web structure mining.
These categories are shown in figure 3.

2.1 Web Usage Mining
Web usage mining depend upon the user's requirement like some users are interested in multimedia data others are interested in textual data .This type of mining discover the user's need on the internet. This process use user's log

2.2 Web Structure Mining
Web structure mining is the process of extracting knowledge from web pages by focusing the structure. According to the type of web structural data, web structure mining can be divided into two kinds: 1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.

2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

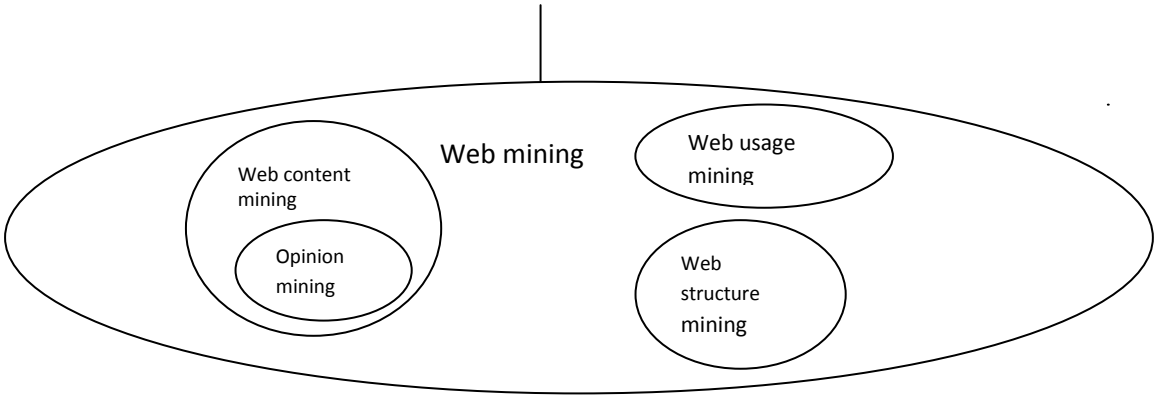

Fig:1

2.3 Web Content Mining

Web content mining aims to extract useful information from contents of the web page. It involves scanning of all the contents on web page to find its relevance with the search query.

3. Levels of Sentiment Analysis-

3.1 Document level sentiment analysis

In this we consider only a single review about a single topic. Supervised and unsupervised learning methods can be used for it. In case of forums, blogs comparative sentence may appear. It means comparison of two similar type of product may possible.

Advantage-from one document we can find out the overall polarity.

Disadvantage- different emotions about different features of an entity could not be extracted separately.

For this purpose most suitable techniques directly refers Artificial Intelligence. Some popular Data Mining techniques are- Classification , clustering generalization ,association, rule mining, data visualization ,neural network, fuzzy logic , genetic algorithm ,Bayesian network, churn prediction, multi agent system, decision tree and many more.

3.2 Sentence level sentiment analysis

In this we find the polarity of each sentence. After that we classify it into classes as positive, negative, neutral.

Advantage- Lies in subjectivity/objectivity classification.

3.3 Phrase level sentiment analysis

We extract the phrase which contains opinion words and a phrase level classification is done. This can be advantageous or disadvantageous.

4.OPINION MINING TECHNIQUES

The Rapid growth of social media is directly affecting the database complexity of database is increasing day by day. We need to develop such technology which will be able to extract nuggets of knowledge.

4.1 SUPERVISED LEARNING

Classification is a supervised learning used to find the relationship among attributes. In supervised learning, we have our training data set contains two things .1-Independent variables (dataset related properties) 2-Dependent attribute (predicted attributes.

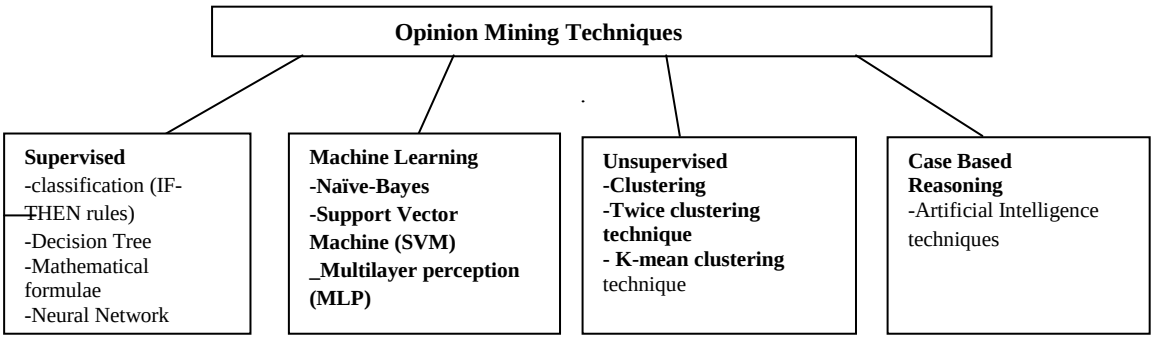

Fig:2 Classification of Opinion mining techniques

On the basis of these two data, we predict the all possible outcomes from given data set. So we can say supervised technique find a function or a model to differentiate data classes or concepts, for the purpose of finding the class of the object whose class label is unknown .This derived model is dependent upon the set of training data (i.e. Data objects whose class label is known).

4.1.1 Classification

In this category ,rules generate in the form of IF-THEN as result from training data. We apply the condition on the data and according to the fulfillment of condition we generate the output.

Ex.-A classified model executed on medical check database and extracted many rules, one rule could be

"IF (Age=25 AND Height>172 AND Weight >90)

OR (Age=65 AND Height>172AND Weight<45)

THEN Medical-Unfit=Yes"

Accuracy is measured in terms of prediction hit rate .It is vary between 80% to 100%.But 100% is impossible and minimum is 80%.

4.1.2Decision Tree

It is a flow chart of nodes .It look like a tree structure. Every node denotes or asks a question/test on an attribute value and every branch represents an output of this test and class distribution is showed by leave nodes. In Fig:3 there is a decision tree with result of 4 classes A,B,X and Y according to different categories of decision tree.

4.1.3Neural Network

Neural network is a connection of neurons. Each neuron in the network work like processing which further will generate the desired output. Every neuron in the network connected with our neurons by weights and frequency distribution to show the polarity of online customer reviews. So to show the rating graph of any customer this schema is very useful to customer .In Fig:4 there is a hospital report neural network chart to determine the fitness of the patient according to his /her age ,height and weight.

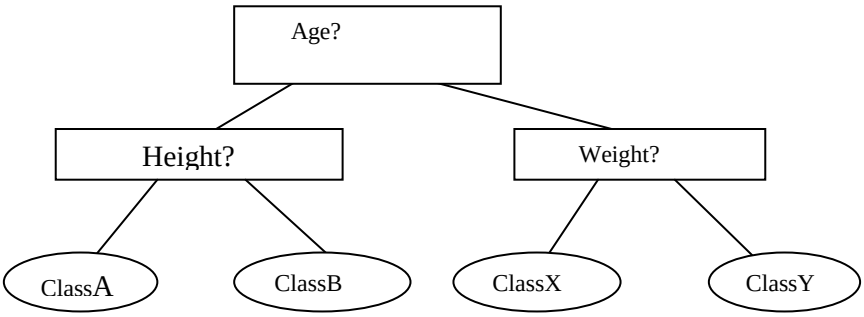

Fig3 : Decision Tree

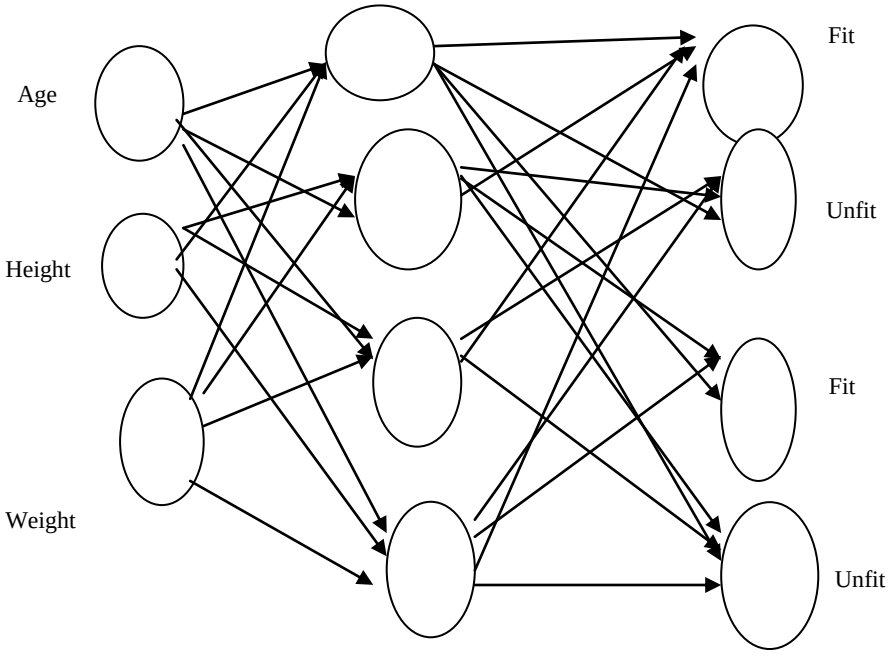

Fig 4:Neural Network

There exist mainly 4 machine learning techniques [table-1] used for sentiment classification [10]. Word based unigram, Bigram, Chins character based bigram and Trigram methods are used to find out the features.

WBB and CBB give good result in classification of text. All of these Naïve Bayes Classifier show best result.

Kernel based –based machine learning approach – It integrated multiple features from syntactical and lexical levels to fetch the opinions at sentence level based on SVM SMO technique. [11]

4.2 Unsupervised Learning

Unsupervised learning is about to learn by observation instead of learn by example [7] because in this no explicit targeted output attached with input and class label for any unknown step. We collect the object of similar qualities in one group and object of another quality in another group. Opinion words were extracted in different concept ANGER, DISGUST, FEAR, JOY, SADNESS .Divide it into clusters using fuzzy c clustering technique and the assign (WNA)word net score to extended opinions .K-mean and bayes classifier We can use to cluster the attribute words along opinions from sentence group in [8]

An unsupervised twice clustering technique proposed in [9].K-mean cluster technique used to cluster the videos according to create signature.

4.3 Case Based Reasoning

It is a popular and recent problem solving technique .The problem which are related to real time scenario can easily solved by CBR tool. Case based reasoning is an emerging Artificial Intelligence supervised technique mostly used to solve those problems which are similar to past problems.

CBR contains the past problem's solution in CBR repository .This repository is called knowledgebase or case base. CBR use this knowledgebase to solve a new problem by reuse the solution not use 'first principal' reasoning. We can also modify a little bit if it will produce the desired output.

Table1.Different features of machine learning techniques

Model	Concept	Extensions	Accuracy	Advantage	Disadvantage
SVM(support vector machine)	Based on decision plane that defines the boundaries of decision.	1-Soft Margin 2-Non Linear 3-Multi Class Sum	With unigram-82.9%	1-low dependency on data set 2-good for experimental result	1-Reads preprocessing for missing values 2-Interpretation is difficult
MLP(Multilayer perception)	In this 1 or N layer exist for input or output .Also called feed forward neural network.	2 phase 1-forward phase input layer to output layer 2-Change the weight and bias value error.	84.25 % - 89.50%	1-Act as a universal function 2-Can learn each relationship	1-Needs more time for execution 2-Considerd as a complex black box.
Naïve Bayes Classifier	For 2 event conditional prob.- P(e1/e2)= P(e2/e1)P(e1)/e2	Used Accuracy Precision Recall Relevance	Got 0.79939 accuracy.	1-Easy to implement. 2-Efficient computation.	1-Assumption of attributes being independent which may not be necessarily valid.

Table 2. Comparison between different techniques used previously

STUDIES	YEAR	SENTIMENT	LEARNING ALGORITHM	PREDICTION ACCURACY
Pravesh Kumar Singh	Feb 2014	Positive Negative	1-Naïve Bayes 2-SVM 3-Multilayer - Perception 4-Clustring	Naïve-.79939 SVM-82.9% MLP-84.25to 89.50% Clustering-65.33to99.57%
Gayathri Deepthi K.Sashi Reha	April 2014	Positive Negative Neutral	Naïve Bayessian Classifier	.846

Poongodi s Radha oj	Sep 2013	Positive Negative	1-Naïve Bayes 2-Support Vectoer – Machine(SVM) 3-Multi Layer Perception(MLP)	Using NLP NB-87.98% SVM-92.84% MLP-93.70% NLP+emoticons NB-90.41% SVM-97.28% MLP-99.57%
Article Dr.M.S.Vijya V Pream Sudha	Dec 2013	Positive Negative Neutral	1-SVM 2-K-NN with Rapidoniner	SVM-85.5% K-NN-70%
Myunsook	2013	Positive Negative	1-libSVM 2-MultilayerNN 3-J48 4-Random forest	SVM-87.90% MLNN-87.72% J48-85.98% RF-88.39%
Akash Srivastava Bhaskar pant	April 2012	Positive Negative Average	Lib-SVM	74.8268%
Alexander Pak Patrick Paroubek	2012	Positive Negative Neutral	1-Unigram 2-Bigram 3-Trigram 4-Ngram	Best performance is recorded using Bigram And with classifier very accurate

Table 3. Comparison between different techniques used previously[4]

Studies	**Year**	**Mining Techniques**	**Feature Selection**	**Data Source**	**performance**
Rui	2011	Naïve Bayes	Unigram Bigram Trigram	Movie Review	NB-85.8% ME-85.4%
Kaiquan	2011	Multiclass SVM	Lingustic Feature	Multiclass SVM	61%
Gamgam	2010	Maximum Entropy	Dependency Relation	Amazon Review	72.6%
Gang li	2010	K means Clustering	TF-IDF	Movie Review	78%
Songhoton	2008	Centroid Classifier	MI JG CHI	Chnsentico rp	90%SVM
Melville	2000	Bayesian Classifier	Ngram	Blogs	91.21%
Rudy	2000	SVM Hybrid	Document Frequency	Movie Review	89%

5. CONCLUSION

Opinion mining is a burning field of web mining. There exist a lot of benefits of opinion mining at customer and business level. Opinion study about a particular product provides us a clear picture of future.

So company can modify their product according to customer's need and customer can aware about that particular product before going to purchase it. A bulk of data is daily posted on web sites like face book ,twitter e.t.c. User post their sentiments in the form of comments, reviews and feedback daily .An opinion mining process gives us the way to extract

pearl knowledge from it. Now people are moving towards the latest social media like twitter. And business organizations are also moving toward these social sites. They fetch feedback from there to decide the future direction. So by knowing the key features of supervised, unsupervised and case based reasoning techniques we will be able to better result in future.

6. FUTURE WORK

Yet we find out the accuracy of mining process at a measurable level but still there is a great space of improvement in this area. The comprehensive and comparative literature review
Of different mining models give us a better understanding for future work .We can work on models to improve their accuracy. There is demand of a technique which will able to address all the challenges simultaneously.

7.REFERENCES

[1] South Morgan Street, Bing Liu. 2011. Identifying Noun Product Features that Imply Opinions.

[2] Alexandra Balahur and Andrés Montoy.2010.OpAL. Applying Opinion Mining Techniques for the Disambiguation of Sentiment Ambiguous Adjectives in SemEval-2 Task 18.

[3] Andrea Esuli. 2008. Automatic Generation of Lexical Resources for Opinion Mining. Models, Algorithms and Applications.

[4] Nidhi Mishra,C.K.Jha , oct2012.Classification of Opinion Mining Techniques ,IJCA(0975-8887) vol-56-no-13.

[5] Ayesha Rashid, Naveed Anwer2, Dr. Muddaser Iqbal3, Dr. Muhammad Sher vol.10 Issue 6, No 2,November 2013 (A Survey Paper: Areas, Techniques and Challenges of Opinion Mining

[6] Fan, G WU "Opinion Summarization of Customer comments"International conference on Applied Physics and IndustrialEngineering in 2012.

[7] X. Su, G. Gao, Y. Tian, "A Framework to Answer Questions of Opinion Type" Seventh Web Information Systems and Applications Conference in 2010.

[8] X. Ding, B. Liu and P. S. Yu "A Holistic Lexicon-Based Approach to Opinion Mining" Proceedings of the International Conference on Web Search and Data Mining Pages 231-240 in 2008.

[9] Timothy L. Acorn, Sherry H. Walden "Sm

[10] art: Support: Management Automated Reasoning Technology for Compaq Customer Service" In proceeding of: Proceedings of the Fourth Conference on Innovative Applications of Artificial Intelligence in 1992.

[11] S. Wang ; Taiyuan X. Yin ; J. Zhang and Ru Li "Sentiment clustering of product object based on feature reduction" Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on Date of Conference: 29-31 May 2012.

[12] D. Martens "Predicting going concern opinion with data mining"Decision Support Systems 45 765–77 international journal in 2008