Opinion mining for reputation evaluation on unstructured Big Data

Mrs. Uma Gurav, Prof. Dr. Nandini sidnal

Abstract— **Big Data analysis is a current research trend in computer science field. It is used for reputation evaluation based on customer reviews of any kind of product and applications. Opinion mining also known as sentiment analysis is one of the most important part of this research area. Big Data is a new term used to identify the datasets that due to their large size, we cannot manage them with the typical data mining software tools. This data is in the order of magnitude of petabytes. It can be found easily on web, Social media, remote sensing data and medical records in the form of customer reviews etc., it may be structured, semi-structured or unstructured data and we can use this big data for opinion mining. This paper describes the methods used for reputation evaluation on big unstructured data, it also focuses on combination of different classifiers techniques to overcome the challenges and incrementally enhance the granularity of opinion capturing.**

Index Terms **- Big Data, opinion mining, sentiment analysis, Data mining, Machine learning**

I. INTRODUCTION

The Big Data opinion mining is becoming an important tool to improve efficiency and quality in organizations, and its importance is going to increase in the coming years. It is the important aspect for capturing public opinion about product preferences, marketing campaigns, political movements, social events and company strategies. In recent times, research activities in the areas of Opinion, Sentiments and/or Emotions in natural language texts and other social media are gaining momentum based on subjectivity or objectivity analysis. The reason may be the huge amount of available text data in the social Web in the forms of news, reviews, blogs, chats and even twitter . Though , opinion mining from natural language text is a multifaceted and multidisciplinary problem, in general, the term "sentiment" is used in reference to the automatic analysis of natural language text. Research efforts are being carried out for identification of positive or negative or neutral polarity of evaluative text and for development of opinion mining tools and human sentiment recognition devices. Artificial Intelligence (AI) techniques play important role in these tasks. The main four aspects of the opinion mining problem are Object identification, Feature extraction, Orientation classification and Integration. The important issues that need attention include how various psychological phenomena can be explained in computational terms and which AI concepts and computer modelling methodologies will prove most useful from the human sentiment's point of view. In the following sections, analysis of various methods is done in a more descriptive way:

Manuscript received April, 2015.

Mrs. Uma Gurav, Assistant professor Information Technology Department, K.I.T's College of engineering, Kolhapur India, Mobile No-9867277931

Prof . Dr. Nandini Sidnal, Head ,Computer Science *Department, Associate Professor, K.L.E.s College of engineering, Visvesvaraya Technological University ,Belgaum Karnataka, ,India.*

2. A Study and Comparison of various Opinion Extraction Methods for Reputation Evaluation:

2.1 The machine learning methods: These perform the supervised or semi- supervised learning by extracting the features from the text and learn the model. It's a part of artificial intelligence techniques which uses several learning algorithms to determine the sentiment by training on a known dataset. The aim of Machine Learning is to develop an algorithm so as to optimize the performance of the system using example data or past experience. The Machine Learning provides a solution to the classification problem that involves two steps:

1) Learning the model from a corpus of training data
2) Classifying the unknown data based on the trained model.

In general, classification tasks are often divided into several sub-tasks:

1) Data preprocessing
2) Feature selection and/or feature reduction
3) Representation
4) Classification
5) Post processing

Feature selection and feature reduction attempt to reduce the dimensionality (i.e the number of features) for the remaining steps of the task. The classification phase of the process finds the actual mapping between patterns and labels (or targets). Active learning, a kind of machine learning is a promising way for sentiment classification to reduce the annotation cost. The following are some of the Machine Learning approaches commonly used for Sentiment, categorized document or sentences into positive, negative or neutral categories. Machine learning techniques classified into two basic techniques as defined below [1-4].

2.1.1 Supervised Machine Learning Techniques

Supervised machine learning techniques are used for classified document or sentences into finite set of class i.e. into positive, negative and neutral. Training data set is available for all kind of classes. Support Vector Machine

(SVM), Naive-Bayes, K-nearest neighbor (KNN), Logistic regression for classification purpose can be used. Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization, generally outperforming Naive Bayes [40]. They are large-margin, rather than probabilistic, classifiers, in contrast to Naive Bayes and Maximum Entropy.SVM efficiently classify news articles, Blogs into positive, negative or neutral category. Naive-Bayes efficiently classifies tweets or small piece of sentences called "Crunches". KNN also give good result for sentence level sentiment analysis. It is an approach to text classification that assigns the class c* = argmaxc P(c | d), to a given document d.

A naive Bayes classifier is a simple probabilistic classifier based on Bayes theorem and is particularly suited when the dimensionality of the inputs are high. Its underlying probability model can be described as an "independent feature model".

2.1.2 Unsupervised Machine Learning Techniques

Unsupervised machine learning techniques don't use training data set for classification. Clustering algorithms like K-means clustering, Hierarchical clustering used to classify data into categories. Semantic Orientation also provides to generate accurate result for classification. Neural network can be also used for defining threshold values to the words and classify them based on the defined values. Point wise mutual information (PMI) is also one of the unsupervised classification methods for sentiment analysis.

2.2 Natural Language Processing

Natural language processing techniques plays important role to get accurate sentiment analysis. NLP techniques like Bag of words, Hidden markov model, part of speech (POS), N-gram algorithms, large sentiment lexicon acquisition and parsing techniques are used to express opinion for document level, sentences level and aspect level [1,2,12]. Large sentiment lexicon acquisition is used sentiment word dictionary which contains lot of sentiment words with their numeric threshold value for particular domain [1].The lexicon-based approach involves calculating sentiment polarity for a review using the semantic orientation of words or sentences in the review. The "semantic orientation" is a measure of subjectivity and opinion in text. It deals with the actual text element. It transforms it into a format that the machine can use.

2.2.1Artificial intelligence

It uses the information given by the NLP and uses a lot of maths to determine whether something is negative or positive: it is used for clustering.

2.2.2 Maximum Entropy

Maximum Entropy (ME) classification is yet another technique, which has proven effective in a number of natural language processing applications [26]. Sometimes, it outperforms Naive Bayes at standard text classification [27].

2.2.3 SentiWordNet dictionary

SentiWordNet dictionary is used for subjective sentiment analysis [21]. The method defines distance d(t1, t2) between terms t1 and t2 as the length of the shortest path between t1 and t2 in WordNet. The orientation of t is defined as SO(t) =(d(t, Like) − d(t, Hate))/d(Like,Hate). |SO(t)| is the strength of the sentiment of t, SO(t) > 0 entails t is positive, and t is negative otherwise[1],[5],[12],[21].

For objective sentiment classification we have to expand the vocabulary of SentiWordNet or WordNet by adding more words with proper threshold value. Noun phrase (NP), verb oriented, adjective oriented sentiment analysis concentrate on NP, verb and adjective respectively to classify the sentence or entity into positive, negative or neutral.[2], [5], [13] Word based techniques, Emotional based techniques are part of the NLP domain for sentiment analysis classification particularly for twitter message analysis [6], [7] .

2.3 Text Mining Techniques

Text mining techniques are also useful for efficient automatic sentiment analysis for twitter messages. Text mining process divides into four stages. In this approach supervised machine learning algorithms are used for classification purpose.

Text Mining Process is explained as Text collection --> pre-processing --> analysis -->validation

Text mining classifier architecture stages are Tweets --> validation -->selection --> classification (positive, negative , neutral)

2.4 Techniques of Information Theory and Coding [14], [18]

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

2.5 Semantic Approach [22]

For sentence level and entity or aspect level SA the semantic approach is really useful and gives efficient result. We can use Ontology learning techniques or description logic (DL) for defining semantic rules and put them together in the knowledge base. The rule-based approach looks for opinion words in a text and then classified it is based on the number of positive and negative words. It considers different rules for classification such as dictionary polarity, negation words, booster words, idioms, emoticons, mixed opinions etc. Using

ISSN: 2278 – 1323

the rules of ontology and/or DL we can attach semantic orientation to the sentences or to the entities for proper opinion capturing.

2.6. Sampling

Sampling is based on the fact that if the dataset is too large and we cannot use all the examples, we can obtain an approximate solution using a subset of the examples. A good sampling method will try to select the best instances, to have a good performance using a small quantity of memory and time. An alternative to sampling is the use of probabilistic techniques. [7], "Big data does not need big machines, it needs big intelligence".

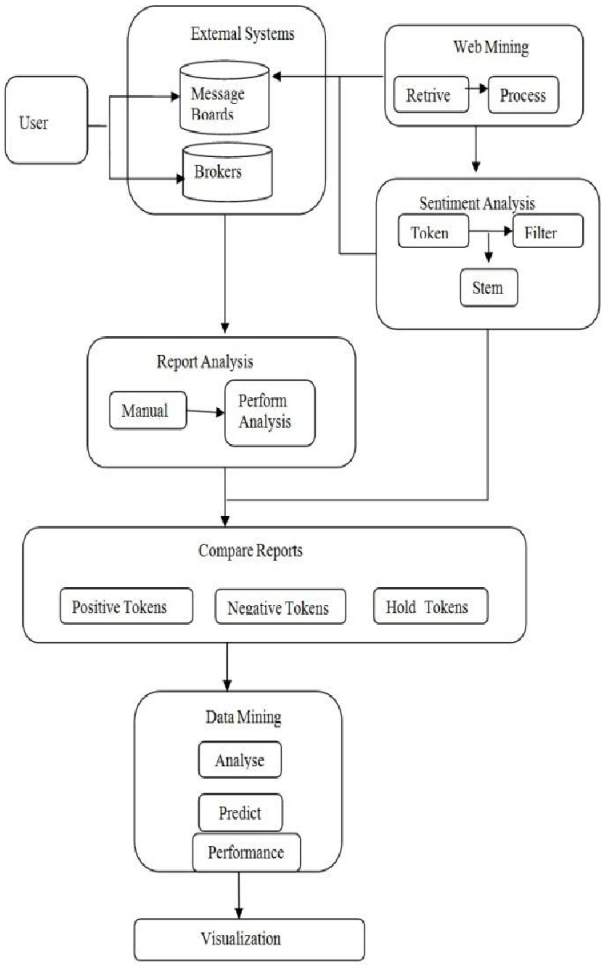

Figure 1 : Explains the process of data mining with sentiment analysis[11].

2.7 Distributed systems

The most popular distributed systems used nowadays are based in the map-reduce framework. The map-reduce methodology started in Google, as a way to perform crawling of the web in a faster way. Hadoop is an open-source implementation of map-reduce started in Yahoo and is being used in many non-streaming big data analysis. A way to speed up the mining of streaming learners is to distribute the training process onto several machines. Hadoop Map Reduce is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes. A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job. The map-reduce model divides algorithms in two main steps: map and reduce, inspired in ideas in functional programming. The input data is split into several datasets and each split is send to a mapper that will transform the data. The output of the mappers will be combined in reducers that will produce the final output of the algorithm.

3. Sentiment analysis is done on three levels:

1. Document Level
2. Sentence Level
3. Entity or Aspect Level.

3.1. Document Level Sentiment analysis

Is performed for the whole document and then decide whether the document express positive or negative sentiment. The basic information unit is a single document of opinionated text. In this document level classification, a single review about a single topic is considered. But in the case of forums or blogs, comparative sentences may appear. Customers may compare one product with another that has similar characteristics and hence document level analysis is not desirable in forums and blogs.
The challenge in the document level classification is that the entire sentence in a document may not be relevant in expressing the opinion about an entity. Therefore subjectivity/objectivity classification is very important in this type of classification. The irrelevant sentences must be eliminated from the processing works. Both supervised and unsupervised learning methods can be used for the document level classification. Any supervised learning algorithm like naive Bayesian, Support Vector Machine, can be used to train the system. For training and testing data, the reviewer rating (in the form of 1-5 stars), can be used. The features that can be used for the machine learning are term frequency, adjectives from Part of speech tagging, Opinion words and phrases, negations, dependencies etc. Labeling the polarities of the document manually is time consuming and hence the user rating available can be made use of. The unsupervised learning can be done by extracting the opinion words inside a document. The point-wise mutual information can be made use of to find the semantics of the extracted words. Thus the document level sentiment classification has its own advantages and disadvantages. Advantage is that we get an overall polarity of opinion text about a particular entity from a document. Disadvantage is that the different emotions about different features of an entity could not be extracted separately.

3.2 Sentence level sentiment analysis

In the sentence level sentiment analysis, the polarity (positive /negative /neutral) of each sentence is calculated. The same

document level classification methods can be applied to the sentence level classification problem. Objective and subjective sentences must be found out. The subjective sentences contain opinion words which help in determining the sentiment about the entity. After which the polarity classification is done into positive and negative classes. Sentence level sentiment classification is not desirable in case complex sentences, it is useful in case of single and simple sentence. Knowing that a sentence is positive or negative is of lesser use than knowing the polarity of a particular feature of a product. The advantage of sentence level analysis lies in the subjectivity/ objectivity classification. The traditional algorithms can be used for the training processes.
Many of the statements about entities are factual in nature and yet they still carry sentiment. Current opinion mining methods express the sentiment of subjective statements and neglect such objective statements that carry sentiment [1]. For Example, "I bought a Samsung grand phone two weeks ago. Everything was good initially. The voice was clear and the battery life was long, although it is a bit slim and lightweight model. Then, it stopped working yesterday. [1]" The first sentence expresses no opinion as it simply states a fact. All other sentences express either explicit or implicit sentiments. The last sentence "Then, it stopped working yesterday" is objective sentences but current techniques can't express sentiment for the above specified sentence even though it carry negative sentiment or undesirable sentiment.

3.3. Entity or Aspect/phrase Level sentiment analysis

The phrase level sentiment classification is a much more pinpointed approach to opinion mining. The phrases that contain opinion words are found out and a phrase level classification is done. This can be advantageous or disadvantageous. In some cases, the exact opinion about an entity can be correctly extracted. But in some other cases, where contextual polarity also matters, the result may not be fully accurate. Negation of words can occur locally. In such cases, this level of sentiment analysis suffices. But if there are sentences with negating words which are far apart from the opinion words, phrase level analysis is not desirable.The words that appear very near to each other are considered to be in a phrase. For example consider a statement "My Samsung Galaxy S3 phone has good picture quality but it has low phone memory storage." so sentiment on Samsung Galaxy's camera and display quality is positive but the sentiment on its phone memory storage is negative. Hence , we can generate summery of opinions about entities. Comparative statements are also part of the entity or aspect level sentiment analysis.

4. Related work

Often based on a combination of machine learning methods with dedicated background information, such as dictionaries, opinion mining techniques with a good accuracy can be developed relatively quickly by using labeled examples and sentiment words as features. After an initial training phase based on a supervised classification of regression technique, the polarity of the opinion expressed in free texts can be automatically estimated, enabling large scale analyses of opinions [23]. Following are a study analysis on a few of them.

4.1 Naive Bayesian classifier

A notable approach in [3] uses a sentence level sentiment analysis. The word level feature extraction is done using Naive Bayesian Classifier. The semantic orientation of the individual sentences is retrieved from the contextual information. This machine learning approach on average claims an accuracy rate of 83%. For classifying and analyzing of the sentiment from the reviews, machine learning and lexical contextual information are used. The above paper [3] focuses on sentence level to check whether the sentences are objective or subjective and to classify the polarity of the sentences to positive or negative opinion. The naive bayes approach is used to annotate (online dictionary) each sentence as positive and negative on the bases of useful word level feature.

4.2 SVM classifier

SVM classifier is trained on the annotated sentences for the positive and negative classification. Contextual information is used to calculate the polarity of sentence and mark it as either negative or positive. The paper [4] presents experiments for sentiment analysis to automatically distinguish prior and contextual polarity. Beginning with a large stable of clues marked with prior polarity, method identifies the contextual polarity of the phrases that contain instances of those clues in the corpus. A two-step process is used in [4] that employ machine learning and a variety of features. Firstly the method classifies each phrase containing a clue as neutral or polar. Secondly it takes all phrases marked in previous step as polar and disambiguates their contextual polarity (positive, negative, both, or neutral) of sentiment expressions, achieving reliable results.

4.3 Natural language processing

Another significant work is the implementation of both Natural Language understanding and Generation in Sentiment analysis [5]. A method describes a system that automatically identifies the contextual polarity based on algorithms to search and predict the orientation of opinions is specified. In this research work, a review database that stores the opinionated texts. The method then finds frequent features that many people have expressed their opinions on. After that, the opinion words are extracted using the resulting frequent features, and semantic orientations of the opinion words are identified with the help of WordNet. The system then finds those infrequent features. The orientation of each opinion sentence is identified and a final text summary is generated in this work. The part of speech tagging from natural language processing is used to find opinion features. Thus, text summary of opinions is generated. Summarization of text is also done as a subsystem. But this summarization work is

ISSN: 2278 – 1323

truly dependent on the features and hence is far from the automatic summarization work in the field of NLP. The paper proposes a method by utilizing the adjective synonym set and antonym set in WordNet to predict the semantic orientations of adjectives.

A method of sentiment analysis which does not use conventional natural language rules is specified in [6]. The work uses a machine learning approach (Naive Bayesian) for classification. The class association rules are used to extract the associations between term features appearing in consumer review opinions and product features for a particular consumer product. A set of pre-classified opinion sentences is utilized as training data to develop class association rules. The f-measure is used as metric for evaluation, and claims efficiency up to 70%. In the above paper[6], the review sentences are divided into various classes according to the association rules. The classification of the opinionated text is done using both class association rules and naive Bayesian classifier. After which the experiments done proves that Class association rules perform better than the traditional naive Bayesian classifiers.

In [7], the authors present an approach for opinion mining which relies on natural language processing techniques. The work is accomplished by the sentiment lexicon and a pattern database. The two feature selection algorithms discussed in this work are based on mixture model and the likelihood ratio. They propose a sentiment pattern based analysis. In [8], an in-depth study of short range and long range dependency relations among the words of a sentence is discussed. They use a clustering approach after the parsing is done. In the paper [9] a combined model of sentiment analysis is done. Considering every levels of analysis like phrase level, sentence level and document level have their own advantages**.**

4.4 Combination model

Combination model including all the three may achieve better performance. A combined model based on phrase and sentence level analyses and a description on the implementation of different levels of analyses are presented. For the phrase level sentiment analysis, a template is used. The newly defined template is Left-Middle-Right template. The Conditional Random Fields are used to extract the sentiment words. The Maximum Entropy model is used in the sentence-level sentiment analysis. The combination model with specific combination of features performs slightly better than the traditional single level models. Another paper which studies the mining of on-line reviews in the movie domain is [10]. In the paper they come up with a proposal of a model called S-PLSA (Sentiment Probabilistic Latent Semantic Analysis). This is a generative model for sentiment analysis that does a deeper comprehension of the sentiments in blogs.

4.5 Combination of different classifiers like Naive Bayesian classifier and genetic algorithms:

An important advantage for combining redundant and complementary classifiers is to increase robustness, accuracy, and better overall generalization. The base classifiers such as Naïve Bayes (NB), and Genetic Algorithm (GA) are combined instructed to predict classification scores. The reason for that choice is that they are representative classification methods and heterogeneous techniques in terms of their strengths. Well known heterogeneous techniques are performed with base classifiers to obtain a very good generalization performance. , which can generate better results on sentiment analysis. This is a multi–step process which includes, Pre-processing phase, Document Indexing phase, feature reduction phase, classification phase and combining phase to aggregate the best classification results. These combination methods can prove more accurate, since GA has better performance rate than NB in the important respects of accuracy.

4.6 Topic modeling and Sentiment Analysis

Even if opinions are correctly extracted from texts, they need to be aggregated and summarized to be properly analyzed. Creating single-document summaries of reviews is recognized to be a difficult task [22]. The general approach consists in first clustering the texts by topics and then organizing the texts by the type of sentiment for each topic [26]. The most popular topic identification technique is Latent Dirichlet allocation [27] and its application to sentiment summarization [29]. Some researchers applied text categorization methods to extract sentiment at the document level. Pang and Lee have classified movie reviews with bag-of-words features and an SVM classifier. Later they adopted a hierarchical approach, using a classifier to first find subjective sentences and a second one to determine their polarity [23]. Other approaches focus on identifying opinions inside sentences at the sub-sentence expression level. Wilson et al. introduced Opinion Finder, which employs several different classifiers to identify subjective sentences, speech acts and direct subjective expressions, opinion sources and opinion polarities [33].Godbole et al. use a simple rule-based approach, utilizing custom sentiment lexicons, identifying negations and using named-entity and co reference resolution [47]. Breck et al. use conditional random field classifiers to identify direct speech events [27], while Ding et al. use lexicons to extract specific product features from customer reviews to automatically generate opinion summaries [33]. Lun-Wei et al. also generate opinion summaries from news and blog articles in addition to extracting opinion polarity, degree and correlated events [42].

The approach presented in [21] is based on

1) An information retrieval system identifying relevant tweets,

2) An opinion detection algorithm based on counting positive and negative words,

3) A predictive model based on a moving average time series model.

Previous opinion mining techniques are mostly based on word count and rarely use advanced NLP techniques, such as a syntax analyzer. However, many opinions are expressed in an ambiguous way (a survey is given in [22]). Building on [28] lexical semantic information can be used together with a data-driven approach based on natural language processing as input to a Bayesian machine learning method.

4.6 Sentiment Analysis for Subjective Sentences

Bo Pang and Lillian Lee, research paper explains degree of positivity polarity, subjectivity detection and Opinion identification using SVM and N-gram algorithms [8]. Pang and Lee, a mincut-based algorithm was proposed to classify each sentence as being subjective or objective [9]. The algorithm works on a sentence graph of an opinion document. They also express supervised, unsupervised approaches for classification for sentiment analysis. Ana C.E.S Lima and Leandro N.de Castro presents hybrid approach of emotional-based and word-based for automatic sentimental analysis of twitter messages (i.e. Tweets) and they also use basic text mining techniques and naive-Byes classification algorithm which provide good efficiency.[6] Generally sentimental word dictionaries will be used for labeling of Small piece of data called "crunches". These kinds of dictionaries contain certain threshold value for sentiment word and the defined value is used to decide sentiment of word is positive or negative for subjective sentences. SentiWordNet V3.0 or WordNet are the online available sentiment word dictionaries [21].

1.) Positive Sentiment in subjective sentence: "I like my new Dell Laptop" Defined sentence is expressed positive sentiment about the laptop brand Dell and we can decide that from the sentiment threshold value of word "like". Threshold value of word "like" has positive numerical threshold value. Use this threshold value in the classification algorithm like naive-Bayes.

2.) Negative sentiment in subjective sentences: "Phata poster nikala hero is the flop movie" defined sentence is expressed negative sentiment about the movie named "Phata poster nikla hero"and we can decide that from the sentiment threshold value of word "flop". Threshold value of word "flop" has negative numerical threshold value. Use this threshold value in the classification algorithm like naive-Bayes.

3.) Neutral sentiment in subjective sentences: "I'm going for a long drive" defined sentence is expressed fact. It doesn't carry any sentiment so we put this kind of statement in the neutral category. We can decide that the defined sentence is neutral because there is absence of words that express sentiment. Polarity, subjective detection and opinion identification all are important in sentiment analysis. .

4.7 Sentiment Analysis for Objective Sentences

Sentiment Analysis for objective sentences is a research topic now-a-days because there are so many data sources which have objective sentences that carry sentiment but because of lack of proper algorithms and contexts we can't get the good result from the objective sentences. According to recent article published by Ronen Feldman express that objective sentences that carry sentiment should be analyzed for getting efficient sentiment analysis and this is one of the challenging task in sentiment analysis. [1], [5] Source of objective sentences are including news articles, blogs, social media etc. where we get good amount of objective sentences. [5] We consider following examples which are objective sentences but still carry sentiment. [1], [5], [12].

"Firefox keeps crashing." defined sentences carry negative sentiment about Firefox web browser.

"The earphone broke in two days." defined sentence carry negative sentiment about the earphones.

"I get relaxed time after today's session." define positive sentiment about person's routine.

Available sentiment dictionaries don't have enough vocabulary to get analyzed objective sentences and categorized them efficiently into positive, negative or neutral. Provide proper context or semantic orientation is also very important part of sentiment analysis of objective sentences. [31] Discusses a survey of sentiment analysis, and [24] for opinion mining techniques. To build classifiers for sentiment analysis, we need to collect training data so that we can apply appropriate learning algorithms. Labeling tweets manually as positive or negative is a laborious and expensive, if not impossible, task. However, a significant advantage of Twitter data is that many tweets have author provided sentiment indicators, changing sentiment is implicit in the use of various types of emoticons. Smiley's or emoticons are visual cues that are associated with emotional states. They are constructed using the characters available on a standard keyboard, representing a facial expression of emotion. Hence this can be used these to label our training data. When the author of a tweet uses an emotion, they are annotating their own text with an emotional state. Such annotated tweets can be used to train a sentiment classifier [8, 10].

4.8 Domain Dependency

A sentiment classifier that is trained to classify opinion polarities in a domain may produce miserable results when the same classifier is used in another domain. Sentiment is expressed differently in different domains. For instance, consider two domains, digital camera and car. The way in

ISSN: 2278 – 1323

which customers express their thoughts, views and prospective about digital camera will be different from those of cars. But some similarities may also be present. So Sentiment analysis is a problem which has high domain dependency. Therefore cross domain sentiment analysis is a challenging problem that has to be unfolded.

4.9 Opinion Spam Detection

A key feature of social media is that it enables anyone from anywhere in the world to freely express his/her views and opinions without disclosing his/her true identify and without the fear of undesirable consequences. These opinions are thus highly valuable. However, this anonymity also comes with a price. It allows people with hidden agendas or malicious intentions to easily game the system to give people the impression that they are independent members of the public and post fake opinions to promote or to discredit target products, services, organizations, or individuals without disclosing their true intentions, or the person or organization that they are secretly working for. Such individuals are called opinion spammers and their activities are called opinion spamming (Jindal and Liu, 2008; Jindal and Liu,2007). Opinion spamming has become a major issue. Apart from individuals who give fake opinions in reviews and forum discussions, there are also commercial companies that are in the business of writing fake reviews and bogus blogs for their clients. Several high profile cases of fake reviews have been reported in the news. It is important to detect such spamming activities to ensure that the opinions on the Web are a trusted source of valuable information. Unlike extraction of positive and negative opinions, opinion spam detection is not just a NLP problem as it involves the analysis of people's posting behaviors. It is thus also a data mining problem.

5. Research Objective:

Specification and investigation of methods to incrementally enhance the granularity of opinion capturing has to be done to overcome the challenges stated in above sections.

Following points need to be focused:

1. Most of the solutions focusing on global review classification consider only the polarity of the reviews (positive/negative) and rely on machine learning techniques. We aim for solutions that aim a more detailed classification of reviews (e.g., three or five star ratings) use more linguistic features including negation, modality and discourse structure.

2. We can use hybrid approaches like SVM, Naive-Bayes, BOW, POS, Large sentiment lexicon acquisition,(Subjective lexicon is a list of words where each word is assigned a score that indicates nature of word in terms of positive, negative or objective.) SentiWordNet or WordNet, N-gram, statistical modeling and rule-based natural language processing techniques ,grammar rules and text mining techniques and methods to do classification and efficient SA on subjective and objective sentences.

3. Although based on more advanced NLP techniques, our focus is on a similar method based on

1) Identification of the documents related to the topic of interest

2) Opinion detection algorithm

3) Construction of the dynamic model of the opinion diffusion, which can be further used for simulation and prediction on big data.

4) To detect opinion spamming.

5) Can be one of the useful parameter for trust evaluation in cloud about cloud services, it's providers, user satisfaction, and cloud based applications.

7. CONCLUSION

Opinion mining on unstructured big data is a machine learning problem that has been a research interest for recent years. Although several notable works have come in this field, a fully automated and highly efficient system has not been introduced till now. This is because of the unstructured nature of natural language, Big Data. The vocabulary of natural language is very large that things become even hard. Various methods and hybrid approaches discussed above can be used for fully automated and efficient sentiment analysis on big data. By performing an extensive research in the related area, we identified many research challenges in sentiment analysis that are yet to be addressed. Several challenges still exist in the field of machine learning and some of them are co-reference Resolution, domain dependency etc. These problems have to be tackled separately and those solutions can be used to improve the methods to do effective sentiment analysis and opinion extraction from big data.

8. REFERENCES

1. Bing Liu, Sentiment Analysis and Opinion Mining,Morgan and Claypool Publishers, May 2012.p.18-19,27-28,44-45,47,90-101.
2. Nitin Indurkhya, Fred J. Damerau , Handbook of Natural Language Processing, Second Edition, CRC Press, 2010.
3. B. B. Khairullah Khan, Aurangzeb Khan, "Sentence based sentiment classification from online customer reviews," ACM, 2010.
4. P. H. Theresa Wilson, Janyce Wiebe, "Proceedings of human language technology conference and conference on empirical methods in natural language processing," Association for Computational Linguistics, p 347354, 2005.
5. M. Hu and B. Liu, "Mining and summarizing customer review," KDD04, ACM, 2004.
6. C.W. C.C.Yang, Y.C. Wong, "Classifying web review opinions for consumer product analysis," ICEC09,ACM, 2009.
7. R. B. W. N. Jeonghee Yi, T Nasukawa, "Sentiment analyzer: Extracting sentiments about a given topic

using natural language processing techniques," ICDM03, IEEE, 2003.
8. P. B. Subhabrata Mukherjee, "Feature specific sentiment analysis for product reviews."
9. W. X. G. C. Si Li, Hao Zhang and J. Guo, "Exploiting combined multi-level model for document sentiment analysis," International Conference on Pattern Recognition IEEE, 2010.
10. Ronen Feldman, James Sanger, The Text Mining Handbook-Advance Approaches in Analyzing Unstructured Data, Cambridge University Press,2007.
11. Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publications, 2006.
12. Ronen Feldman, "Techniques and Application of Sentiment Analysis", Communication of ACM, April 2013, vol. 56.No.4.
13. Ana C.E.S Lima and Leandro N.de Castro, "Automatic Sentiment Analysis of Twitter Messages", *I*EEE Fourth International Conference on Computational Aspect .of Social Networks (CASoN), p.52-57, 2012.
14. Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P.Sheth, "Harnessing Twitter Big Data for Automatic Emotion Identification ",ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on privacy, Security, Risk and Trust,p.589-592, 2012.
15. Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, vol.2, No1-2(2008)1-135.
16. Bo Pang and Lillian Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Proceedings of ACL, 2004.
17. Huising Xia, Min Tao and Yi Wang, "Sentiment Text classification of customers Reviews on the Web Based on SVM", IEEE Circuts and System Society, Sixth International Conference on Natural Computation (ICNC), p.3633-3937, 2010.
18. Bruno Ohana,Brendan Tierney and Sarah-Jane Delany, "Domain Independent Sentiment Classification with Many Lexicons ", IEEE Computer Society, Workshops of International conference on Advanced Information Networking and Application, p.632-637, 2011.
19. Chihil Hung and Hao-kai Lin, "Using Objective Word in SentiWordNet to Improve Word-of-Mouth Sentiment Classification", IEEE Computer Society, P.47- 54, March-April 2013.
20. Mostafa Karamibekr and Ali A.Ghorbani, "Verb Oriented Sentiment Classification", IEEE Computer Society, IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent technology,p.327-331, 2012.
21. 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.
22. Mostafa Karamibekr and Ali A.Ghorbani, "Sentiment Analysis of Social Issues", IEEE Computer Society, International Conference on Social Informatics, p. 215-221, 2012.
23. Shichang Sun, Hongbo Liu,Hongfei Lin, Ajith Abraham, "Twitter Part of Speech Tagging Using Pre- Classification Hidden markov Model", IEEE International Conference on Systems, Man and Cybernetics, October 14-17,p.1118-1123, 2012.
24. Keisuke Mizumoto, Hidekazu Yanagimoto and Michifumi Yoshioka, "Sentiment Analysis of Stock Market News with Semi-supervised Learning", IEEE Computer Society,IEEE/ACIS 11th International Conference on Computer and Information Science, p.325-328,2012.
25. Sang-Hyun Cho and Hang-Bong Kang, "Text Sentiment Classification for SNS-based Marketing Using Domain Sentiment Dictionary", IEEE International Conference on Conference on consumer Electronics (ICCE), p.717-718, 2012.
26. Aurangzeb Khan and Baharum Baharudin, "Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms form Blogs", 2011.
27. Ms.K.Mouthami, Ms. K .Nirmala Devi, Dr.V.Murali Bhaskaran, "Sentiment Analysis and Classification Based on Textual Review".
28. Online SentiWordNet dictionary source http://sentiwordnet.isti.cnr.it/.
29. Gautam Shroff, Lipika Dey and Puneet Agrawal, "Social Business Intelligence Using Big Data",CSI Communications, April 2013,p.11-16.
30. Wikipedia article on supervised machine learning http://en.m.wikipedia.org/wiki/Supevised_learning.
31. J. Aasman. Unication of geospatial reasoning, temporal logic, & social network analysis in event-based systems. Proc. of the 2nd Intl. Conf. on Distributed event-based systems, pages 139{145, New York, NY, USA, 2008.
32. D. Anicic, P. Fodor, S. Rudolph, R. Stühmer, N. Stojanovic, R.Studer. "A Rule-Based Language for Complex Event Processing and Reasoning." Proc. of the 4th Intl. Conf. on Web reasoning and rule systems, 2010.
33. Crina Costea ,Damien Joyeux ,Omar Hasan,Lionel Brunie ,"A Study and Comparison of Sentiment Analysis ",2008.
34. David M. Blei, et al. "Latent Dirichlet allocation" Journal of Machine Learning Research , 2003.
35. E. Breck, Y. Choi, C. Cardie. "Identifying expressions of opinion in context." Intl Conf. on Artifical intelligence, 2007.
36. Yongzheng (Tiger), Zhang Dan Shen, Catherine Baudin ,"Sentiment Analysis in Practice ",December 12, 2011@ICDM'11.
37. Y. L. Chang and J. T. Chien. Latent Dirichlet learning for document summarization. Proc. of the IEEE Intl. Conf. on Acoustics, Speech,and Signal Processing, 2009.
38. Shradha Tulankar1, Dr Rahul Athale2, Sandeep Bhujbal3, Sentiment Analysis of Equities using Data Mining Techniques and Visualizing the Trends, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No 2, July 2013.
39. Y. Choi and C.Cardie. "Learning with compositional semantics as structural inference for subsentential

ISSN: 2278 – 1323

sentiment analysis." Proc. of the Conf. on Empirical Methods in Natural Language Processing: , 2008.
40. X. Ding, B. Liu, P. S. Yu." A holistic lexicon-based approach to opinion mining. ",Proc. of the 1st ACM Intl. Conf. on Web Search and Data Mining, Feb 11-12, 2008, Stanford University, Stanford, California, USA.
41. O. Etzion. "Semantic approach to event processing" Proc. of the Inaugural Intl. Conf. on Distributed event-based systems, pages 139, New York, NY, USA, 2007. ACM (DEBS 07).
42. Gartner Inc. ," Gartner Identifies the Top 10 Strategic Technologies for sentiment analysis in cloud", sema2010, Oct 2010.
43. A. Go, R. Bhayani, and L. Huang." Twitter sentiment classification using distant supervision" Processing: 1-6, 2008.
44. N. Godbole , M. Srinivasaiah , S. Skiena." Large-scale sentiment analysis for news and blogs" Proc. of ICWSM, Boulder, Colorado, USA, 2007.
45. P. Goyal, R. Mikkilineni." Policy-based event-driven services-oriented architecture for cloud services operation & management". IEEE 2009,Intl. Conf. on Cloud Computing. Bangalore, India, September 2009.
46. Kaschesky, M. and R. Riedl. "Tracing opinion-formation on political issues on the internet" Proc. of the Hawaii Intl. Conf. on System Sciences, 2011.
47. Stella Gatziu Grivas,Michael Kaschesky,Marc Schaaf, "Feature based Opinion mining - towards Performance Measure" ,Published in Proc. of the IEEE International International Journal of Advanced Computer Research September-2013
48. Diego Reforgiato Recupero, Sergio Consoli, Aldo Gangemi "A Semantic Web Based Core Engine to Efficiently Perform Sentiment Analysis" Andrea Giovanni, Nuzzolese, and Daria Spampinato
49. K. Lun-Wei, L. Yu-Ting and C. Hsin-Hsi" Opinion extraction, summarization and tracking in news and blog corporation" Proc. of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006.
50. T. Mullen and R. Malouf. "A preliminary investigation into sentiment analysis of informal political discourse." AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW): 159–162, 2006.
51. B. O'Connor, R. Balasubramanyan, B. R. Routledge, N. A. Smith." From tweets to polls: Linking text sentiment to public opinion time series." Intl. AAAI Conf. on Weblogs and Social Media, Washington,DC, May 2010.
52. B. Pang and L. Lee." Opinion mining and sentiment analysis Found Trends" , 2008.
53. B. Pang and L. Lee. "Thumbs up? Sentiment classification using machine learning". Proc. of EMNL, 2002.
54. A. Paschke. "A Semantic Design Pattern Language for Complex Event Processing." Proc. of AAAI, 2009.
55. D. R. Radev, E. Hovy, and K. McKeown. "Introduction to the special issue on summarization Computational Linguistics",2002.
56. M. Schaaf, A. Koschel, S. Gatziu Grivas, I. Astrova." An Active DBMS Style Activity Service for the Cloud Environments", Proc. of the 1st Intl Conf. on Cloud Computing, GRIDs, and Virtualization November 2010, Lisbon (IARIA 2010).
57. H. Saggion and A. Funk." Extracting Opinions and Facts for Business Intelligence". RNTI, 2009.
58. L. Specia, m. Turchi, N. Cancedda, M. Dymetman and N. Cristianini. "Estimating the sentence-level quality of machine translation systems" Proc. of the 4th Intl. Workshop on Statistical Machine Translation,Athens, Greece, 30-31 March, 2009.
59. V. Stoyanov and C. Cardie."Partially supervised coreference resolution for opinion summarization through structured rule learning",Proc. of Conf. on Empirical Methods in Natural Language Processing, 2006.
60. D. Suthers. Interaction, Mediation, and Ties, "An Analytic Hierarchy for Socio-Technical Systems" Proc. of the 44th Hawaii Intl. Conf. on System Sciences, 2011.
61. K. Teymourian, A. Paschke." Towards Semantic Event Processing." In DEBS'09, July 6-9, Nashville, TN, USA.
62. T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan." OpinionFinder: A system for subjectivity analysis,"Proceedings of HLT/EMNLP, 2005.
63. G. Wishnie and H. Saiedian,"A complex event routing infrastructure for distributed systems." IEEE Computer Society, Vol. 2, pp. 92–95.
64. S. ChandraKala and C. Sindhu," Opinion mining and sentiment classification: A survey ",2009
65. Ravikiran Kalava , G.Anil Kumar , Ch.Vasavi, Workshop on Management in Cloud Computing, ,"Cloud-based Event-processing Architecture for Opinion Mining" (MCC 2011), July 2011, Washington.

Mrs. Uma Gurav, (B.E, M.Tech,) - Currently pursuing P.H.D at Visvesvarya Technological University, Belgaum, Karnataka, in computer science and engineering. She is Working as a Assistant professor, K.I.T's college of engineering, Kolhapur, having industrial and academic experience around 12 years. Her research areas includes Analysis and design of algorithms, Data structures , and distributed systems, cloud computing and Big data Analytics.

Prof. Dr. Nandini Sidnal (B.E, M.Tech, P.H.D)- is working as a Head of Department, Computer science, at K.LE's college of engineering , Belgaum, Karnataka, She has 21 years of experience .Her research areas includes networking, M-commerce, cloud computing and Big data Analytics.

