Comparative Study and Analysis on various Opinion Mining Techniques

Deepak Kumar Yadav
Department of Computer Applications, SSTC, SSGI, FET Bhilai, India

Sampada Vishwas Massey
Department of Computer Science & Engineering, SSTC, SSGI, FET Bhilai, India

***Abstract*— Sentiment analysis is a kind of text classification that classifies texts based on the sentiment orientation of opinions they contain. It is also known as opinion mining refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis seeks to identify the viewpoint(s) underlying a text span; an example application is classifying a movie review. To determine this sentiment polarity. Sentiment Analysis can be used to determine sentiment on a variety of levels. It will score the entire document as positive or negative, and it will also score the sentiment of individual words or phrases in the document. The previews and review of this paper has significant research on the subject of sentiment analysis, expounding its basic terminology, tasks and granularity levels. Sentiment analysis aims to define the attitude of a speaker or a writer regarding some topic or the overall contextual polarity of a document. Among all varieties of social media, Twitter is a valuable resource or main data source for collecting data such as views, reviews etc. for data mining. In this paper we have collected some tweets from social networking sites. So that analysis will be done on those tweets to provide some prediction of business intelligence. Results of trend analysis will be display as tweets with different segments presenting positive, negative and neutral.**

***Index Terms*—computational linguistics, Data Mining, Natural language processing, analysis, twitter social media.**

I. INTRODUCTION

Opinion is a person's perspective about an issue or an object. Mining refers to extraction of knowledge from raw data or facts [1]. Thus opinion mining is the technique used to extract intelligent information based on a person's opinion from raw data available on internet. The answer lies within the increasing use of internet by people for searching about various products, news, latest information etc. Today people are also placing their comments & opinions on social media so that they can be seen by other people too. Survey has shown that such opinion also affects the people reading those opinions [2], [3]. Sentiment analysis can be done at three different levels: document level, sentence level and feature (or aspect) level [4]. At the document level, the goal is to find the opinion direction of the whole document. Hence, it can be seen as a classification task that classifies each document to one of the positive or negative classes. At the sentence level, the goal is to find the opinion orientation of the opinionated sentences. A common approach is to first identify the subjective sentences, and then determine the sentiment of each of the subjective sentences. In aspect level sentiment analysis, the aspects of the object that the user has commented on is first identified, and then the sentiment of the sentence about that aspect is discovered [5].

A. Basic Terminology of sentiment analysis

Sentiment analysis analyses the polarity of opinion (positive or negative). One might think the need of opinion mining and sentiment analysis as why would one need to know about someone's opinion. The words opinion, sentiment, view and belief are used interchangeably but there are subtle differences between them [6].

- *Opinion:* A conclusion thought out yet open to dispute ("each expert seemed to have a different opinion").
- *View:* subjective opinion ("very assertive in stating his views").
- *Belief:* deliberate acceptance and intellectual assent ("a firm belief in her party's platform").
- *Sentiment:* a settled opinion reflective of one's feelings ("her feminist sentiments are well-known").

B. Learning Methods

There are different types of learning types:

- Supervised learning**:** Learning classifier from training data and assign class labels to test data.
- Unsupervised learning: Learning without training data.
- Semi-supervised learning: Amalgamate both labeled and unlabeled training data. The Sentiment

learning uses Machine Learning or Lexicon based learning.

C. Classification of textual review

Classification is a supervised procedure that learns to classify new instances based on learning from a training set of instances that have been properly labeled with the correct classes. An algorithm that implements classification, especially in a concrete execution is classifier. The piece of input data is formally termed an instance, and the categories are termed classes. Text Classification (TC) is one of the prime techniques to deal with the textual data. TC systems are used in a number of applications such as, filtering email messages, classifying customer reviews for large e-commerce sites, web page classification for an internet directory, evaluating exams paper answers and organizing document databases in semantic categories.

II. SENTIMENT ANALYSIS TASKS

Sentiment analysis is a challenging interdisciplinary task which includes natural language processing, web mining and machine learning. It is a complex task and encompasses several separate tasks, viz:

- Subjectivity Classification
- Sentiment Classification
- Complimentary Tasks
  - Object Holder Extraction
  - Object/ Feature Extraction

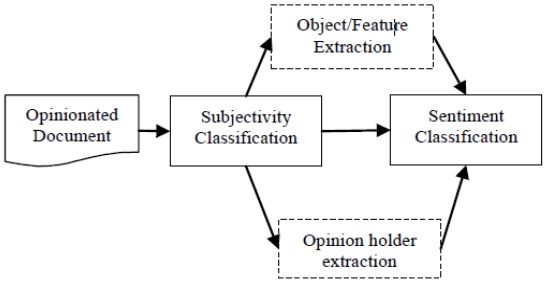

Fig. 1 Tasks of Sentiment Analysis

A. Subjectivity classification

Typically, any given document will contain sentences that express opinion and some that do not. That is, a document is a collection of objective sentences, sentences that state a fact, and subjective sentences, sentences that represents the author's opinion, point of view or emotion. Subjectivity classification is the task of classifying sentences as opinionated or not opinionated [7], [8].Tang et al.[9], stated subjectivity classification as follows: Let $S = \{s_1, \ldots, s_n\}$ be a set of sentences in document D. The problem of subjectivity classification is to distinguish sentences used to present opinions and other forms of subjectivity (subjective sentences set S_s) from sentences used to objectively present factual information (objective sentences set S_o), where $S_s \cup S_o = S$.

B. Sentiment Classification

Once the task of finding whether a piece of text is opinionated is over we have to find the polarity of the text i.e., whether it expresses a positive or negative opinion. Sentiment classification can be a binary classification (positive or negative) [10], multi-class classification (extremely negative, negative, neutral, positive or extremely positive), regression or ranking [11].

Depending upon the application of the sentiment analysis, sub -tasks of opinion holder extraction and object feature extraction are optional. (They have been represented by dashed boxes in Fig.3).

C. Opinion Holder Extraction

Sentiment Analysis also involves elective tasks like opinion holder extraction, i.e. the discovery of opinion holders or sources[12],[13]. Detection of opinion holder is to recognize direct or indirect sources of opinion. They are vital in news articles and other formal documents because multiple opinions can be expressed in the same article corresponding to different opinion holders. In documents like these, the multiple opinion holders may explicitly be mentioned by name. In social networks, review sites and blogs the opinion holder is usually the author who may be identified by the login credentials.

D. Object /Feature Extraction

An additional task is the discovery of the target entity. In contrast with review sites, blogs and social media sites tend not have a set intention or predefined topic and are thus, inclined to discuss assorted topics. In such platforms it becomes necessary to know the target entity.

Also as mentioned before target entities can have features or components that are being reviewed. A reviewer can have differing opinions about the different features or components of the target entity. As a result, feature based sentiment analysis, i.e. extraction of object feature and the related opinion, is an optional task of sentiment analysis [14], [15], [16].

Table 1: Summary of Sentiment Analysis Tasks

At Document Level 1. Task: Sentiment Classification of whole document 2. Classes: Positive, negative and neutral 3. Assumption : Each Document focuses on a single object (not true in discussion posts, blogs ,etc.) and contain opinion from a single
At Sentence Level 1. Task 1: Identifying Subjective/ Objective Sentences • Classes: Objective and Subjective 2. Task 2: Sentiment Classification of Sentences • Classes: positive and negative • Assumption: A sentence contains only one opinion which may not always be true *Prior polarities of words determined at*

word level sentiment analysis is used here
At Feature Level 1.Task 1: Identify and extract object features that have been commented on by an opinion holder (eg. A reviewer) 2. Task 2: Determining whether the opinions on features are negative, positive or neutral 3. Task 3: Find feature synonyms

III. METHOD

A. Architecture

One can consider document-level polarity classification to be just a special (more difficult) case of text categorization with sentiment- rather than topic-based categories. Hence, standard machine learning classification techniques, such as support vector machines (SVMs), can be applied to the entire documents themselves. We therefore propose, *as depicted in Figure* 1, to first employ a subjectivity detector that determines whether each sentence is subjective or not: discarding the objective ones creates an extract that should better represent a review's subjective content to a default polarity classifier.

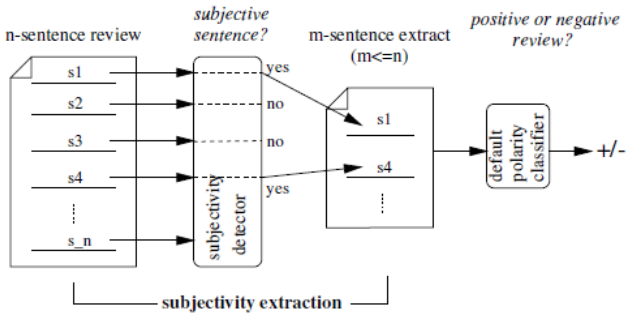

Fig 2: Polarity classification via subjectivity detection.

B. Contextual and subjectivity detention

As with document-level polarity classification, we could perform subjectivity detection on individual sentences by applying a standard classification algorithm on each sentence in isolation. However, modeling proximity relationships between sentences would enable us to leverage coherence: text spans occurring near each other (within discourse boundaries) may share the same subjectivity status,

We would therefore like to supply our algorithms with pair-wise interaction information, e.g., to specify that two particular sentences should ideally receive the same subjectivity label but not state which label this should be. Incorporating such information is somewhat unnatural for classifiers whose input consists simply of individual feature vectors, such as Naive Bayes or SVMs, precisely because such classifiers label each test item in isolation. Using an efficient and intuitive graph-based formulation relying on finding minimum cuts. Our approach is inspired by Blum and Chawla (2001), [17] although they focused on similarity between items (the motivation being to combine labeled and unlabeled data), whereas we are concerned with physical proximity between the items to be classified; indeed, in computer vision, modeling proximity information via graph cuts has led to very effective classification (Boykov, Veksler, and Zabih, 1999)[18].

C. Cut-Based Classification

Figure 2 shows a worked example of the concepts in this section. Suppose we have *n* items $x_1 \ldots\ldots x_n$ to divide into two classes $\mathbf{C_1}$ and C_2, and we have access to two types of information:

- Individual scores $ind_j(xi)$: non-negative estimates of each xi's preference for being in C_j based on just the features of x_i alone; and
- Association scores $assoc(x_i , x_k)$:non-negative estimates of how important it is that x_i and x_k be in the same class[1].

We would like to maximize each item's "net happiness": its individual score for the class it is assigned to, minus its individual score for the other class. But, we also want to penalize putting tightly associated items into different classes. Thus, after some algebra, we arrive at the following optimization problem: assign the x_i s to C_1 and C_2 so as to minimize the partition cost.

$$\sum_{x \in C_1} ind_2(x) + \sum_{x \in C_2} ind_1(x) + \sum_{\substack{x_i \in C_1, \\ x_k \in C_2}} assoc(x_i, x_k).$$

The problem appears intractable, since there are 2^n possible binary partitions of the x_i's. However, suppose we represent the situation in the following manner. Build an undirected graph G with vertices $\{v_1, \ldots.. v_n, s, t\}$; the last two are, respectively, the source and sink. Add n edges (s, v_i) each with weight $ind_1(xi)$ and n edges (v_i , t) each with weight $ind_2(xi)$ Finally, (n/2)edges (v_i , v_k) each with weight $assoc(x_i , x_k)$.

Then, cuts in G are defined as follows:

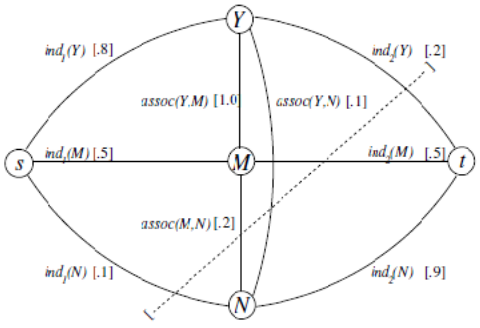

Fig.3(a)

C_1	Individual Penalties	Association Penalties	cost
{Y,M}	.2+.5+.1	.1+.2	1.1
(none)	.8+.5+.1	0	1.4
{Y,M,N}	.2+.5+.9	0	1.6
{Y}	.2+.5+.1	1.0+.1	1.9
{N}	.8+.5+.9	.1+.2	2.5
{M}	.8+.5+.1	1.0+.2	2.6
{Y,N}	.2+.5+.9	1.0+.2	2.8
{M.N}	.8+.5+.9	1.0+.1	3.3

Fig.3(b)

Fig 3(a) and 3(b) Graph for classifying three items. Brackets enclose example values; here, the individual scores happen to be probabilities.

Based on individual scores alone, we would put Y ("yes") in C_1, N ("no") in C_2, and be undecided about M ("maybe"). But the association scores favor cuts that put Y and M in the same class, as shown in the table.
Thus, the minimum cut, indicated by the dashed line, places M together with Y in C_1.

IV. DATA SOURCE

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

A. 3.1. Blogs

With an increasing usage of the internet, blogging and blog pages are growing rapidly. Blog pages have become the most popular means to express one's personal opinions. Bloggers record the daily events in their lives and express their opinions, feelings, and emotions in a blog (Chau & Xu, 2007). Many of these blogs contain reviews on many products, issues, etc. Blogs are used as a source of opinion in many of the studies related to sentiment [19].

B. 3.2. Review sites

For any user in making a purchasing decision, the opinions of others can be an important factor. A large and growing body of user-generated reviews is available on the Internet. The reviews for products or services are usually based on opinions expressed in much unstructured format. The reviewer's data used in most of the sentiment classification studies are collected from the e-commerce websites like www.amazon.com (product reviews), www.yelp.com (restaurant reviews), www.CNET download.com (product reviews) and www.reviewcentre.com, which hosts millions of product reviews by consumers. Other than these the available are professional review sites such as www.dpreview.com , www.zdnet.com and consumer opinion sites on broad topics and products such as www .consumerreview.com, www.epinions.com, www.bizrate[20].

C. 3.3. Data Set

Most of the work in the field uses movie reviews data for classification. Movie review data's are available as dataset (http://www.cs.cornell.edu/People/pabo/movie-review-data). Other dataset which is available online is multi-domain sentiment(MDS)dataset.(http://www.cs.jhu.edu/mdredze/datasets/sentiment). The MDS dataset contains four different types of product reviews extracted from Amazon.com including Books, DVDs, Electronics and Kitchen appliances, with 1000 positive and 1000 negative reviews for each domain. Another review dataset available is http://www.cs.uic.edu/liub/FBS/CustomerReviewData.zi p. This dataset consists of reviews of five electronics products downloaded from Amazon and Cnet

D. 3.4. Micro-blogging

Twitter is a popular microblogging service where users create status messages called "tweets". These tweets sometimes express opinions about different topics. Twitter messages are also used as data source for classifying sentiment.

V. RELATED WORK

A. 4.1. Social Network Analysis

Social network analysis is a methodology mainly developed by sociologists and researchers in social psychology. Social network analysis views social relationships in terms of network theory, while individual actor being seen as a node and relationship between each node are presented as an edge. Social network analysis has been defining in [21] as an assumption of the importance of relationships among interacting units, and the relations defined by linkages among units are a fundamental component of network theories. Social network analysis has emerged as a key technique in modern sociology. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, information science, organizational studies, social psychology, and sociolinguistics 1. In 1954, Barnes [22] started to use the term systematically to denote patterns of ties, encompassing concepts traditionally. Afterwards, there are many scholars expanded the use of systematic social network analysis. Due to the growth of online social networking site, online social networking analysis becomes a hot research topic recently.

B. 4.2. Twitter

Twitter is an online social network used by millions of people around the world to be connected with their friends, family and colleagues through their computers and mobile phones [23]. The interface allows users to post short messages (up to 140 characters) that can be read by any other Twitter user. Users declare the people they are interested in following, in which case they get notified when that person has posted a

new message. A user who is being followed by another user need not necessarily reciprocate by following them back, which renders the links of the network as directed. Twitter is categorized as a micro-blogging service. Micro-blogging is a form of blogging that allows users to send brief text updates or other media such as photographs or audio clips. Among variety of microblogging include Twitter, Plurk, Tumblr, Emote.in, Squeelr, Jaiku, identi.ca, and others, Twitter contains an enormous number of text posts and grows quickly every day. Also, audience on Twitter varies from regular users to celebrities, company representatives, politicians [24], and even country presidents therefore provide a huge base for data mining. We choose Twitter as the source for trend analysis simply because of its popularity and data volume.

C. 4.3. Social Network Analysis on Twitter

A social networking service is an online service that focuses on building social network among people who are willing to share interests, activities, information, or real-life connections. As the fast-growing popularity on the Internet, social network service platform therefore provide adequate information for social network analysis[25].

Among all kinds of social networking service, Twitter, as a micro-blogging service is the second popular social networking site [26]. With its special limitation that only 140 characters can be entered in each tweet, Twitter therefore provide a good position for social network analysis. Many researches has focus on social network analysis on Twitter. Longueville, Smith, and Luraschi [27] focus on how Twitter can be used as a source of spatiotemporal information; Sakaki, Okazaki, and Matsuo [28] present an investigation of the real-time nature of Twitter and proposes an event notification system that monitors tweets and delivers notification promptly; Pak and Paroubek [29] used Twitter as a source of opinion mining and sentiment analysis tasks.

VI. PROPOSED WORK

An overview of steps and techniques commonly used in sentiment classification approaches, as shown in Figure 3. Part of speech model in which a document is represented as a vector, whose entries correspond to individual terms of a vocabulary. Part-of-speech information is supposed to be a significant indicator of sentiment expression. The work on subjectivity detection [30] reveals a high correlation between the presence of adjectives and sentence subjectivity.

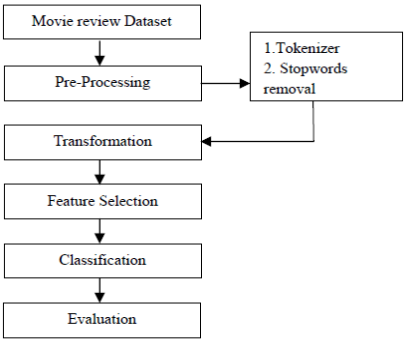

Fig.4: Steps and techniques used in sentiment classification

Indeed, in the study [31], the experimental results show that using only adjectives as features actually results in much worse performance than using the same number of most frequent words.

A. 5.1 Text Preprocessing

Text pre-processing techniques are divided into two subcategories.

1. Tokenization: Textual data comprises block of characters called tokens. The documents are separated as tokens and used for further processing.

2. Removal of Stop Words: A stop-list is the name commonly given to a set or list of stop words. It is typically language specific, although it may contain words. A search engine or other natural language processing system may contain a variety of stop-lists, one per language, or it may contain a single stop-list that is multilingual. Some of the more frequently used stop words for English include "a", "of", "the", "I", "it", "you", and ”and” these are generally regarded as 'functional words' which do not carry meaning. When assessing the contents of natural language, the meaning can be conveyed more clearly by ignoring the functional words. Hence it is practical to remove those words which appear too often that support no information for the task. If the stop word removal is applied, all the stop words in the particular text file will not be loaded. If the stop word removal is not applied, the stop word removal algorithm will be disabled when the dataset is loaded.

B. 5.2 Text Transformation

The score of each sentence in the source document is calculated by sum of weight of each term in the corresponding sentences. The weight of each term is calculated by multiplication of TF and IDF of that word based on adjective word extracted from Parts of speech tags. The TF and IDF are defined as

TF (t) = Number of times the adjective term occurs in document(d)/Total Number of adjective in document(d)…….(3)

IDF(t)=log{ND/DF(t)} ……(4)

Here ND is total number of document in the document collection and DF (t) is number of documents in which adjective term (t) occurs in the document collection.

C. 5.3 Feature Selection

Many statistical feature selection methods for document level classification can also be used for sentiment analysis. The simplest statistical approach for feature selection is to use the most frequently occurring words in the corpus as polarity indicators. The majority of the approaches for sentiment analysis involve a two-step process:

- Identify the parts of the document to contribute the positive or negative sentiments.
- Join these parts of the document in ways that increase the odds of the document falling into one of these two polar categories.

D. 5.4 Sentiment Fuzzy Classification

Sentiment polarity is vague with regard to its conceptual extension. There is not a clear boundary between the concepts of "positive", "neutral" and "negative". To better handle such intrinsic fuzziness in sentiment polarity, we apply the fuzzy set theory to sentiment classification. To do so, we first redefine sentiment classes as three fuzzy sets, and then apply existing fuzzy distributions to construct membership functionsfor the three sentiment fuzzy sets. A fuzzy set is defined by a membership function. These functions can be any arbitrary shape but are typically triangular or trapezoidal. In our formulation, the entire opinionated documents under discussion are represented as a sorted set, denoted by *X*, in terms of their opinion weight (calculated by TF-IDF).

X = [Min (Opinion weight (Si)), …, Max(Opinion weight(Si))]

Where, i= {1,…,n}, Min(Opinion weight(Si)) and Max(Opinion weight(Si)) denotes the respective minimum and maximum opinion weight.

1. *Positive sentiment fuzzy set:* If X is a collection of sentiment opinions denoted by *x*, then a positive sentiment fuzzy set P in X can be defined as a set of ordered pairs are $P = \{(x, \mu P(x)) \mid x\ X\} \in$ .....(5) where $\mu_P(x)$denotes the membership function of *x* in P that maps *X* to the membership space *M*. We select the rise semi-trapezoid distribution as the membership function of the positive sentiment fuzzy set, namely.

E. Parameters for evaluation

In the context of classification, True Positives (TP), True Negatives (TN), False Negatives (FN) and False Positives (FP) are used to compare the class labels assigned to documents by a classifier with the classes the items actually belongs to. True positive means, which are truly classified as the positive terms. True positives (TP) are examples that the classifier correctly labeled as belonging to the positive class. False positive (FP) are examples which were not labeled by the classifier as belonging to the positive class but should have been. True Negative (TN) is examples that the classifier correctly labeled as belonging to the negative class. True Negative means, which are truly classified as the Negative terms. At last there is False Negative (FN), which is an example which was not labeled by the classifier as belonging to the negative class but should have been. Other evaluation measures like precision, recall, F-measure, specificity and accuracy can easily be calculated from these four variables.

Table: 2 Contegency table

		Correct labels	
		Positive	Negative
Classified Labels	Positive	TP(True Positive)	FN(False Negative)
	Negative	FN(False Negative)	TP(True Positive)

VI. SYSTEM FRAMEWORK

We present a model which collects tweets from social networking sites and thus provide a view of business intelligence. In our framework, there are two layers in the sentiment analysis tool, the data processing layer and sentiment analysis layer. Data processing layer deals with data collection and data mining, while sentiment analysis layer use a application to present the result of data mining. More details will be introduced in the following sections.

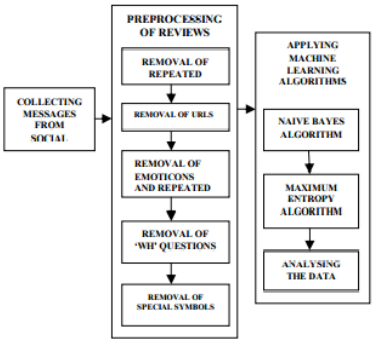

Fig. 5: Architecture for SAT using Machine learning algorithm

A. 5.1. Data Collection and Preprocessing

As now, we have set up the list of tweets or comments on different products manually. We then go through the website of social network sites to collect tweets. All data collected will be stored in a database for further analysis. During the analysis process, words and their polarities are taken into considerations. Combining with social semantic analysis and natural language processing, tweets about daily gossips or unrelated contents will be discarded, and thus relative contents are accurately extracted.

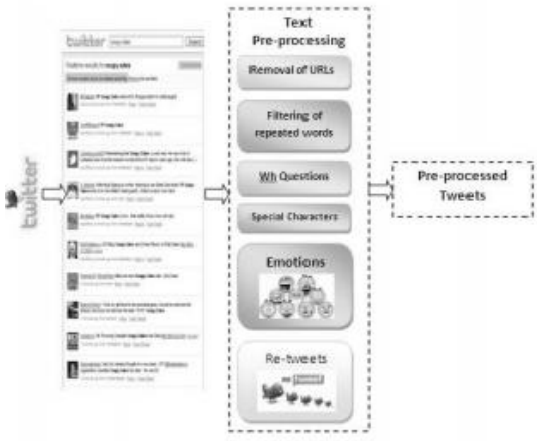

Fig. 6: Overview of Data Collection and Preprocessing Process

The system architecture consists of three parts i.e. collecting messages from social networking sites, preprocessing and applying algorithms. First we have to select the messages from text file or excel file to preprocess the messages. In preprocessing, they remove the unnecessary data like repeated messages (tweets), repeated letters, urls, emotion icons, WHQuestions, special symbols. We have to select an

algorithm after preprocess the messages by classifying to get the reviews on any product like cinemas, phones, iPods, electronic media etc…so this project contains two modules.

B. 5.2. Machine Learning Algorithms

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. A learner can take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. Data can be seen as examples that illustrate relations between observed variables. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies in the fact that the set of all possible behaviors given all possible inputs is too large to be covered by the set of observed examples (training data). Hence the learner must generalize from the given examples, so as to be able to produce a useful output in new cases.

C. 5.3 Naive Bayes algorithm

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

Proposed Naive bayes classifier Input: messages m= {m_1, m_2, m3.....m_n}, Database: Naive Table NT Output: Positive messages p= {p_1, p_2...}, Negative messages n= {n_1, n_2, n_3....}, Neutral messages nu= {nu_1, nu_2, nu_3...}

- Proposed Naive bayes classifier

Input: messages m= {m_1, m_2, m_3.....m_n},
Database: Naive Table N_T
Output: Positive messages p= {p_1, p_2...},
Negative messages n= {n_1, n_2, n_3....},
Neutral messages nu= {nu_1, nu_2, nu_3...}

M= {m_1, m_2, m_3.....}
Step: 1 Divide a message into words m_i = {w_1, w_2, w_3.....}, i=1, 2... n
Step 2: if $w_i \in N_T$ Return +ve polarity and -ve polarity Step 3: Calculate overall polarity of a word=log (+ve polarity)-log(-ve polarity)
Step 4: Repeat step 2 until end of words
Step 5: add the polarities of all words of a message i.e. total polarity of a message. Step 6: Based on that polarity, message can be positive or negative or neutral. Step 7:
repeat step 1 until M ∈ NULL

D. 5.3 Maximum Entropy algorithm

Maximum entropy (ME) models, variously known as log linear, Gibbs, exponential and multinomial logic models, provide a general purpose machine learning technique for classification and prediction which has been successfully applied to fields as diverse as computer vision and econometrics. In natural language processing, recent years have seen ME techniques used for sentence boundary detection, part of speech tagging, parse selection and ambiguity resolution, and stochastic attribute-value grammars, to name just a few applications. A leading advantage of ME models is their flexibility: they allow stochastic rule systems to be augmented with additional syntactic, semantic, and pragmatic features. However, the richness of their presentations is not without cost. Even modest ME models can require considerable computational resources and very large quantities of annotated training data in order to accurately estimate the model's parameters. While parameter estimation for ME models is conceptually straightforward, in practice ME models for typical natural language tasks are usually quite large, and frequently contain hundreds of thousands of free parameters. Estimation of such large models is not only expensive, but also, due to sparsely distributed features, sensitive to round-off errors. Thus, highly efficient, accurate, scalable methods are required for estimating the parameters of practical models. In this paper, we consider a number of algorithms for estimating the parameters of ME models, including Generalized Iterative Scaling and Improved Iterative Scaling, as well as general purpose optimization techniques such as gradient ascent, conjugate gradient, and variable metric methods. Surprisingly, the widely used iterative scaling algorithms perform quite poorly, and for all of the test problems, a limited memory variable metric algorithm outperformed the other choices.

Theoretically, MaxEnt performs better than Naive Bayes because it handles feature overlap better. However, in practice, Naive Bayes can still perform well on a variety of problems.

Proposed Maximum Entropy classifier

Input: messages m= {m_1, m_2, m_3....., m_n},
Database: Naive Table N_T
Output: Positive messages p= {p_1, p_2...},
Negative messages n= {n_1, n_2, n_3....}
Neutral messages nu= {nu_1, nu_2, nu_3...}

M= {m_1, m2, m3.....}
Step: 1 Divide a message into words
m_i = {w_1, w_2, w_3.....},i=1,2,....,n
Step 2: if w_i N_T return +ve polarity and -ve polarity
Step 3: Calculate overall polarity of a word= ((+ve polarity)*log (1/+ve polarity)) - ((-ve polarity)*log (1/-ve polarity)) Step 4: Repeat step 2 until end of words Step 5: add the polarities of all words of a message i.e. total polarity of a message.
Step 6: Based on that polarity, message can be positive or negative or neutral.
Step 7: repeat step 1 until M€ NULL

Table 3: Compare the file size with Pre-processing techniques.

VII. RUNTIME EXECUTION

PRE-PROCESSING	FILE SIZE(KB)	TOTAL
Before preprocessing	82.9	100%
After removal of RT tweets	80.9	96.9%
After removal of url	80.0	96.5%
After filtering and removal of emotion icons	77.8	93.8%
After removal of WHquestions	76.7	92.5%
After removal of special symbols	76.7	92.5%

In the preprocessing the original message size is 100% so, after removing the RTtweets the file size is gradually decreased to 96.9%. after url removal the file size is decreased as well as in filtering technique and emotion icons, removal of WHquestions, removal of special symbols also decreases the file size of the messages. so, after pre-preprocessing the messages it gradually decreases to 92.5%.

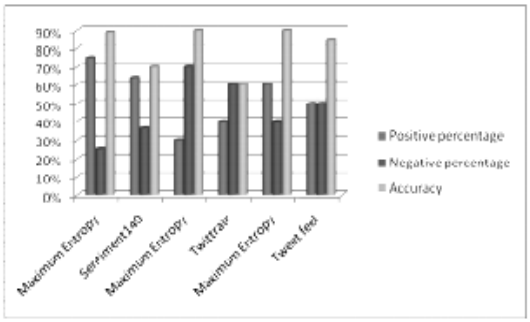

Fig. 7: Bar Graph

- Accuracy: Accuracy is the ratio of number of correctly classified documents and total number of documents

$$\text{Accuracy} = \frac{\text{Number of correctly classified documents}}{\text{Total Number of documents}}$$

VIII. CONCLUSION

We show that using emoticons as noisy labels for training data is an effective way to perform distant supervised learning. Machine learning algorithms (Naïve Bayes, maximum entropy classification) can achieve high accuracy for classifying sentiment when using this method. Although Twitter messages have unique characteristics compared to other corpora, machine learning algorithms are shown to classify tweet sentiment with similar performance. This paper illustrates the research area of Sentiment Analysis and its latest advances. It affirms the terminology, the major tasks, the granularity levels, and applications of sentiment analysis. It also discusses the impact of Web 2.0 applications on this research field. Most work has been done on product reviews – documents that have a definite topic. More general writing with varied domains, such as blog posts, tweets, posts and web pages, have recently been creating & receiving attention. Future work in expanding existing techniques to handle more general writings and crossing domains is an exciting opportunity for both academia and businesses.

REFERENCES

[1] Baker, R., & Yacef, K. (2009). "The State of Educational Data mining in 2009: A Remark Future Visions." Journal of Educational Data Mining.
[2] P.-Y. S. Chen, S.-Y.Wu, and J. Yoon, "The impact of online recommendations and consumer feedback on sales," in International Conference on Information Systems (ICIS), pp. 711–724, 2004.
[3] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book remarks," Journal of Marketing Research, vol. 43, pp. 345–354, August 2006.
[4] B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, 2012.
[5] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 168–177.
[6] Pang, B and Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieva,l 2008,(1-2),1–135
[7] Hatzivassiloglou, V. and Wiebe, J. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING), 2000.
[19] Riloff E. and Wiebe J., Learning extraction patterns for subjective expressions.Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) , 2003.
[8] Blum, Avrim and Shuchi Chawla. 2001. Learning from labeled and unlabeled data using graph mincuts. In Intl. Conf. on Machine Learning (ICML),pages 19.26.
[9] Tang H, Tan S, and Cheng X. A survey on sentiment detection of reviews. Expert Systems with Applications: An International Journal, September 2009, 36(7):10760–10773.
[10] Anderson, P. What is Web 2.0? Ideas, technologies and implications for education. Technical report, JISC,2007
[11] 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.
[12] Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S., Identifying sources of opinions with conditional random fields and extraction patterns. Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005.
[13] [21] Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D., Automatic extraction of opinion propositions and their holders. Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text, 2004.
[14] Yi, J., Nasukawa, T., Niblack, W., &Bunescu, R.,Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. Proceedings of the 3rd IEEE international conference on data mining (ICDM 2003):427–434
[15] [23] Hu, M. and Liu, B. Mining opinion features in customer reviews. In Proceedings of AAAI, 2004: 755–760.
[16] [24] Popescu A-M. and Etzioni O., Extracting product features and opinions from reviews, Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP),2005
[17] Boykov, Yuri, Olga Veksler, and Ramin Zabih.1999. Fast approximate energy minimization via graph cuts. In Intl. Conf. on Computer Vision (ICCV), pages 377.384. Journal version in IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 23(11):1222.1239, 2001.
[18] Martin, J. (2005). Blogging for dollars. Fortune Small Business, 15(10), 88–92.
[19] Popescu, A. M., Etzioni, O.: Extracting Product Features and Opinions from Reviews, In Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, 2005, 339–346.
[20] Hu and Liu ,2006; Konig & Brill ,2006 ; Long Sheng ,2011; 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.
[21] B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1{135, 2008.

[22] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79{86, 2002
[23] Twitter Sentiment Classification using Distant Supervision by Alec Go, Richa Bhayani, and Lei Huang.
[24] J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Association for Computational Linguistics, 2005.
[25] K. Nigam, J. Lafferty, and A. Mccallum. Using maximum entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61{67, 1999.
[26] [Mikheev, 1999] Andrei Mikheev. Feature lattics and maximum entropy models. Machine Learning, 1999.
[27] [Nigam et al., 1999] Kamal Nigam, Andrew McCallum,Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. Machine Learning, 1999.
[28] [Csisz_ar, 1996] I. Csisz_ar. Maxent, mathematics, and information theory. In K. Hanson and R. Silver, editors, Maximum Entropy and Bayesian Methods. Kluwer Academic Publishers, 1996.
[29] [Rosenfeld, 1994] Ronald Rosenfeld. Adaptive Statistical Language Modeling: A Maximum Entropy Approach. PhD thesis, Carnegie Mellon University, 1994.
[30] Hatzivassiloglou . V, J. Wiebe, Effects of adjective orientation and gradability on sentence subjectivity, in: Proceedings of the International Conferenceon Computational Linguistics (COLING), 2000, pp. 299–305.
[31] Jaynce Wiebe and Rada Mihalcea (2006), 'Word Sense and Subjectivity', Joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, pp. 73-99.

First Author I. Deepak Kumar Yadav is pursuing M.E. (Master of Engineering) in Computer Technology & Applications from SSTC, SSGI, FET Bhiai Chhattisgarh Swami Vivekanada University ,Bhilai,, India and has done his MCA (with Honors) from BIT, Durg, India. Also working as an Assistant Professor in Department of Computer Application SSTC, SSGI, FET Bhilai. His interest area is Data Mining and Knowledge Management Process, Information and data security.

Second Author II. Sampada Vishwas Massey has done MTech. In Computer Science from Chhattisgarh Swami Vivekanada University ,Bhilai,, India and has done her BE in Computer Science & Engineering from CSVTU, Bhilai, India. Also working as an Assistant Professor in Department of Computer Science & Engineering SSTC, SSGI, FET Bhilai. Her interest area is Image Processing and Neural Network.

