International Journal of Engineering Research and General Science Volume 3, Issue 1, January-February, 2015
ISSN 2091-2730

A Text Sentimental Approach for Online Portals Using Hadoop

Revathi S,
PG Scholar,
SVS College of Engineering,
Coimbatore.
Tamil Nadu – India
revsbe@gmail.com

Rajkumar N,
Assistant Professor,
SVS College of Engineering,
Coimbatore.
Tamil Nadu – India
nrajkumar84@gmail.com

Sathish S,
Assistant Professor,
Karpagam college of Engineering,
Coimbatore.
Tamil Nadu – India
sathishmeped@gmail.com
+91-9003638180

Abstract: Big data is an emerging technology to process the vast amount of both structured and unstructured data. Now a day social media such as twitter, face book, blogs and forums are the well suitable source to gathering the huge amount of data. Text sentiment analysis for the online portals such as flip kart, Amazon, Godaddy, etc.. are very important to review about their product performance in the market. Sentiment analysis is a text analysis method which aims to contextualize the meaning of the social network data. In the existing work, sentiment analysis is done by polarizing the sentences which derived from the public opinion. However it cannot polarize the public opinion accurately where the sentiment analysis is performed over the social network data's. In this work, we target on finding an appropriate polarity recognition method for public opinion supervision system. In our method, we explore new feature extraction rules which extract emotional nouns, verbs, adjectives, and bigrams as representative features. Then, we apply Fuzzy Naïve Bias to classify these online opinions into positive and negative class. Also we introduce the new category of sentiment analysis namely called as 'Neutral'. The experimental conducted were proves that the proposed methodology provides better result than the existing methodology

Index Terms*:* Big Data, Fuzzy Naïve Bias, Hadoop, Rule Base Sentiment Analysis, Sentiment Analysis, Job Tracker, Task Tracker, Name Node, Data Node

1. INTRODUCTION

Big data (also spelled Big Data) is a general term used to describe the voluminous amount of unstructured and semi-structured data a company creates -- data that would take too much time and cost too much money to load into a relational database for analysis. Although Big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data. A primary goal for looking at big data is to discover repeatable business patterns. It's generally accepted that unstructured data, most of it located in text files, accounts for at least 80% of an organization's data. If left unmanaged, the sheer volume of unstructured data that's generated each year within an enterprise can be costly in terms of storage. Unmanaged data can also pose a liability if information cannot be located in the event of a compliance audit or lawsuit. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, duration, storage, search, sharing, transfer, analysis, and visualization.

Hadoop:

Hadoop is a data processing software framework. It is a recent big data technology. It handles structured as well as unstructured information. It is open source software. It is recognized in 2008 and supports multiple operating systems, which was developed by Google and popularized by yahoo.

For effective scheduling of work, every Hadoop-compatible file system should provide location awareness: the name of the rack where a worker node is. Hadoop applications can use this information to run work on the node where the data is, and, failing that, on the same rack/switch, reducing backbone traffic. HDFS uses this method when replicating data to try to keep different copies of the data on different racks. The goal is to reduce the impact of a rack power outage or switch failure, so that even if these events occur, the data may still be readable. A small Hadoop cluster includes a single master and multiple worker nodes. The master node consists of a Job Tracker, Task Tracker, Name Node and Data Node. A slave or worker node acts as both a Data Node and Task Tracker, though it is possible to have data-only worker nodes and compute-only worker nodes. These are normally used only in nonstandard applications. Hadoop requires Java Runtime Environment (JRE) 1.6 or higher. The standard start-up and shutdown scripts require Secure Shell (ssh) to be set up between nodes in the cluster.

Personal computers, smart phones, tablets, and an ever-growing number of embedded devices can now all connect and communicate with each other via the internet. Computing devices have numerous uses and are essential for businesses, scientists, governments, engineers, and the everyday consumer. What all these devices have in common is the potential to generate data. Essentially, data can come from anywhere. Sensors gathering climate data, a person posting to a social media site, or a cell phone GPS signal are example sources of data. The popularity of the Internet alongside a sharp increase in the network bandwidth available to users has resulted in the generation of huge amounts of data. Furthermore, the types of data created are as broad and diverse as the reasons for generating it.

II. RELATED RESEARCH

This section presents the development of sentiment analysis in recent years. Since the first study in this area focused on the analysis of the semantic orientation of adjectives [8], techniques of sentiment analysis have been extensively used in text filtering, tracking of public opinion, and customer relationship management [1], [4]–[6]. Sentiment analysis is mining affective information from data and recognizing the sentiment polarity contained in the information (e.g., happy or sad, approve or disapprove, and agree or disagree). The classification of former studies has been done by different standards [2], [7]. In accordance with the study by
Zhang *et al.*, the present study discusses previous studies by their level of granularity, type of analytical technique, and language [7].

1) Level of Granularity: Previous studies discuss the problem related to sentiment analysis at different levels of granularity, from the document level to the sentence level. For example, Pang *et al.* classified the sentiments of articles by adopting a standard bag-of-features framework, which features unigrams and bigrams of words [8]. Turney *et al.* proposed an unsupervised learning algorithm known as pointwise mutual information and information retrieval (PMI-IR) to predict these semantic orientations of an article by calculating the similarity of its contained phrases to two reference words: “excellent” and “poor” [9]. Several recent studies have also considered the spread, density, and intensity of polar lexical terms to improve the performance of sentiment classification [10].

2) Type of Analytical Technique: Existing approaches to sentiment analysis can be categorized into rule- and learning based approaches. Rule-based approaches often require an expert-defined dictionary of subjective words; this approach predicts the polarity of a sentence or document by analyzing the occurring patterns of such words in text [11]. For example, Wiebe *et al.* provided a lexicon source of subjectivity clues, such as verbs, adjectives, and nouns, with their polarity (i.e., positive, negative, or neutral) and strength (i.e., strong or weak) annotated [12]. However, this lexicon is able to define the original polarity of a word only, and the actual polarity of a word may be modified by its context in a sentence. Several approaches that consider the context of words have been proposed to determine the sentiment orientation of words. Yuen *et al.* proposed an approach to deriving the semantic polarity of words on the basis of morphemes [13]. Knowledge sources, such as WordNet, have also been used to measure the semantic polarity of adjectives [14].

As to learning-based approaches, Hu and Liu [15] developed an approach to extracting option features from product reviews based on linguistic patterns called class sequential rules, which can be mined from a set of labeled training sequences of words and part-of-speech tags. Pang *et al.* [8] represented reviews as a bag of unigram/bigram features and applied three machine-learning methods to predict their sentiment. However, they found that, for sentiment classification, machine learning algorithms did not perform as well as traditional topic categorization tasks. In addition, learning-based sentiment classification requires sufficiently large training data sets with positive and negative examples manually labeled, which are often very costly and time consuming [9].

3) Language: Most sentiment analysis studies have focused on the English language and achieved remarkable success in numerous applications. By contrast, Chinese sentiment analysis has not been sufficiently investigated [17]. The unique linguistic characteristics of the Chinese language pose several technical challenges for Chinese sentiment analysis. The primary challenge is that the Chinese language does not segment words by spaces in sentences. Therefore, word segmentation is often required as an additional step in Chinese language processing [16]. In addition, the Chinese language contains various adverbs. The use of these adverbs can lead to subtlety
and ambiguity in sentences. The English language mainly uses suffixes to express comparative and superlative words (-er and -est, respectively), whereas the Chinese language uses various adverbs in varying degrees such as “ /more” and “ /most.”

Thus, determining the sentiment polarity of Chinese sentences presents greater difficulty, particularly when multiple adverbs and subjectivity clues appear in one sentence. Moreover, considering the differences of contexts and the ambiguity of the Chinese language itself, a document that contains several positive words may indicate a strong negative tone, and vice versa.

III. PROPOSED WORK

We are proposed the Fuzzy Naïve Bias (FNB) classification algorithm. In this algorithm we are combined the Fuzzy clustering with Naïve Bias classification algorithm. Fuzzy clustering is that the set of text was consider here to find the similar data and those similar data was grouped into the separate set. This output is given as an input to the Naïve Bias classification. In Fuzzy Naïve Bias (FNB), multiple clusters have the similar data. We are calculating the conditional probability for all data set. Also here we can introduce the third category namely called as neutral polarity. Due to this the computation time is low. Also it shows the performance accuracy in higher.

FNB has some modules. They are,

Data collection:

Data set can be collected in social media such as face book, twitter, micro blogs and forum, etc… Now a day data is increasing up to zeta byte. So that it was too complicated to process this volume of data. Big data is an emerging technology to process volume of data.

Preprocess:

In this module, the tool WordNet is used to remove the prepositions and discriminator. Stop words are words which are filtered out before or after processing of natural language data (text).There is no single universal list of stop words used by all tools processing of natural language. Non-significant words are removed from text such as articles, preposition and conjunction by using the "stemmer". **Stemmer**- Would reduce each word their "root". Example**: "funniest"** would become **"funny"**

Feature Extraction:

The choice of feature plays a key role in deciding precision. As illustrated in the front section, our target is to find a suitable approach to identify emotional trend of online public opinions. Of all sources of public opinions, micro blog and BBS occupy biggest share. And these two sources have distinctive features. First, it's short in length; then, emotional polarity is obvious; also, netizens use lots of phizs, this could be extracted as important feature for emotional trend identification. According to these features, we set a set of extraction object to select suitable feature. HowNet is a tool to process the Chinese text. So that it was not useful to process the English language. We extract the sentence features emotional noun, emotional verbs, adjectives, and adverbs by using the Natural Language Processing (NLP) tool. For an example, if the word is **Noun**, then it is represented as **NN.** If the word is **Pronoun**, then it is represented as **NNP.** If the word is an **adjective**, then it is represented as **JJ.** If the word is an **Adverb**, then it is represented as **RB.** If the word is **Conjunction**, then it is represented as **DT.** If the word is **Preposition**, then it is represented as **PRP.** If the word is **Verb**, then it is represented as **VV**

Opinion Classification:

After choosing the right characteristics to express the opinion of the short online text, we need to select an appropriate classifier to distinguish between different points. In a generalized Naive Bayesian classifier is proposed that uses the fuzzy partition of variables instead of them. It partitions the domain of each continuous variable into fuzzy regions. Therefore, each variable is a linguistic variable taking linguistic values. The training of Fuzzy Naive Bayesian classifier is done by performing an unsupervised fuzzy clustering in the feature space to obtain an optimal fuzzy partition. The conditional probabilities of each node in Fuzzy Naive Bayesian classifier are then estimated.

The proposed method is based on a fuzzy bayesian classifier over LR-type fuzzy numbers. Fuzzy bayes formula is introduced with the following equation:

$$p(w_j|\breve{x}) = \frac{p(\breve{x}|w_j)P(w_j)}{p(\breve{x})} \qquad (1)$$

In order to u se Bayesian classification for fuzzy numbers, it is required to compute for each class. In N-dimensional feature space, samples are in the form of $\breve{x} = \widetilde{x1}, \widetilde{x2}, \ldots \widetilde{xn}$.

In the current work, though, we present Bayesian classifier over fuzzy numbers as a quite new approach. In similar works with the proposed method [(1) and (2)], density estimation over fuzzy data (both discrete and continuous) have been studied with known density function according to following equation:

$$P(e_i|H_j) = \int \mu_{ei}(x) f(x|H_j) dx \qquad (2)$$

Where f(x|Hj) is the conditional probability density function at value x given Hj.

Sentiment Base

The extraction of properties is based on the sentiment, modifier, and rule bases. Here, we identify the updating issues of these bases, which are the key point in practice. Since topics and fashion terms discussed online are quickly changing, the rule and object bases need to be updated with time. In this work, we update the base semi automatically. With regard to the object base, given that the topics change quickly, we should summarize the related topics and objects, as well as their attributions and components.

$$\text{Weight P}c_i = \frac{fp_{c_i} / \sum_{i=1}^{n} fp_{c_i}}{fp_{c_i} / \sum_{i=1}^{n} fp_{c_i} + fn_{c_i} / \sum_{i=1}^{n} fn_{c_i}} \quad (3)$$

$$\text{Weight N}c_i = \frac{fn_{c_i} / \sum_{i=1}^{n} fn_{c_i}}{fn_{c_i} / \sum_{i=1}^{n} fn_{c_i} + fp_{c_i} / \sum_{i=1}^{n} fp_{c_i}} \quad (4)$$

$$\text{S}c_i = \text{WeightP}c_i - \text{WeightN}c_i \quad (5)$$

In the above formula, the polarity $\text{S}c_i$ depends on morphemes c_i, and the absolute value of $\text{S}c_i$ is the degree of tendency of morphemes C_i. The steps for calculating the sentiment polarity of words are as follows. Scan the positive and negative word lexicons; if the word w appears in the positive word lexicon, $S_w = 1$; if the word appears in the negative word lexicon, $S_w = -1$. Otherwise, the sentiment polarity is computed using morphemes by

$$S_w = \frac{1}{p} \sum_{j=1}^{p} S_{c_i} \quad (6)$$

Where S_w represents the sentiment polarity of the word w, which consists of $c_1, c_2 \ldots c_p$. If $S_w > 0$, the sentiment polarity of the word is positive; otherwise, the sentiment polarity of the word is negative. If the value obtained is close to zero, the word can be considered neutral.

FNB Algorithm Steps:

Step1: Create two classes of LR-type fuzzy numbers with arbitrary distribution.

Step2: Consider test samples from one of the created classes.

Step3: Apply K-NN algorithm in order to estimate likelihood density function by using a distance metric. (e.g. Hausdorff, Hathaway and Yang distance).

Step4: Using the estimated likelihood density function, compute a confusion matrix that includes the probability of belonging test samples into classes.

Step5: Finally, the recognition rate is achieved from obtained confusion matrix.

Modifier Base:

Negation adverbs cause sentiment polarity reversal to mean the opposite (e.g., "awesome" is positive, but it becomes negative if the word 'not' presented before it). Similarly, degree adverbs that either strengthen or weaken the intensity of the sentiment polarity must be considered as well. In addition, sentence structure also affects the sentiment polarity value of a sentence. A complex sentence is modified by relational schema.

Performance Evaluation:

We can compare the FNB method with the R-BSA Algorithm which was shown in the following graphs. Due to this our FNB had achieve the higher performance when compared to the R-BSA.

Accuracy

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}} \quad (7)$$

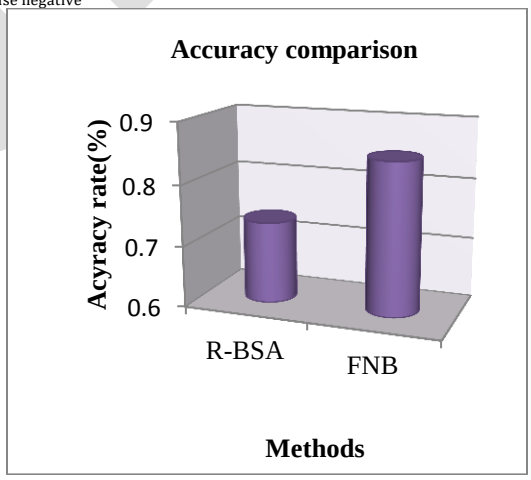

Fig 1. Comparing accuracy between R-BSA and FNB

Precision

Precision value is calculated based on the retrieval of information at true positive prediction, false positive .In healthcare data precision is calculated the percentage of positive results returned that are relevant.

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive+False positive}} \quad (8)$$

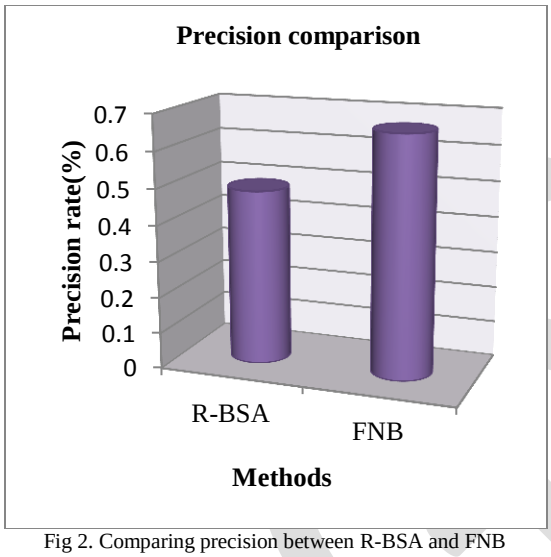

Fig 2. Comparing precision between R-BSA and FNB

Recall

Recall value is calculated based on the retrieval of information at true positive prediction, false negative. In healthcare data precision is calculated the percentage of positive results returned that are Recall in this context is also referred to as the True Positive Rate. Recall is the fraction of relevant instances that are retrieved,

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive+False negative}} \quad (9)$$

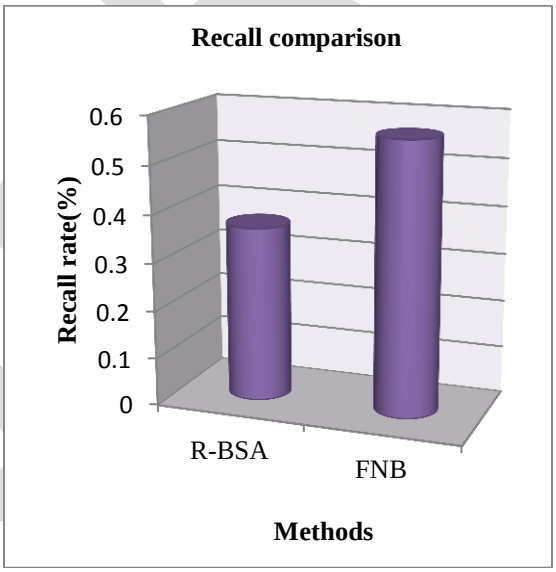

Fig 3. Comparing Recall between R-BSA and FNB

By using the confusion matrix we can calculate the TP, TN, FP and FN.

CONCLUSION

We proposed the FNB method to find the sentiment analysis, sentiment polarity and also evaluate the performance. We introduced the new category namely called as 'Neutral Polarity'. Also we were compared our FNB method with R-BSA method. So that our FNB method has been achieved the performance accuracy in higher. In our future work we can decide to find some of the classification method to improve better performance.

REFERENCES:

[1] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on theWeb," in *Proc. 14th Int. Conf. World Wide Web*, 2005, pp. 342–351.
[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retrieval*, vol. 2, no. 1/2, pp. 1–135, Jan. 2008.
[3] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proc. 8th Conf. Eur. Chapter Assoc. Comput. Linguist.*, 1997, pp. 174–181.
[4] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proc. 2nd Int. Conf. Knowl. Capture*, 2003, pp. 70–77.
[5] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Natural Language Processing and Text Mining*. New York, NY, USA: Springer-Verlag, 2007, pp. 9–28.
[6] E. Riloff, S. Patwardhan, and J. Wiebe, "Feature subsumption for opinion analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2006, pp. 440–448.
[7] C. L. Zhang, D. Zeng, J. X. Li, F. Y. Wang, and W. L. Zuo, "Sentiment analysis of Chinese documents: From sentence to ocument level," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 12, pp. 2474–2487, Dec. 2009.
[8] B. Pang, L. Lee, and S.Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.*, 2002, vol. 10, pp. 79–86.
[9] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, 2002, pp. 417–424.
[10] B. K. Tsou, R. W. Yuen, O. Y. Kwong, T. La, and W. L. Wong, "Polarity classification of celebrity coverage in the Chinese press," in *Proc. Int. Conf. Intell. Anal.*, 2005, pp. 137–142.
[11] K. Bloom, N. Garg, and S. Argamon, "Extracting appraisal expressions," in *Proc. HLT-NAACL*, 2007, pp. 308–315.
[12] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning subjective language," *Comput. Linguist.*, vol. 30, no. 3, pp. 277–308, Sep. 2004.
[13] R. W. Yuen, T. Y. Chan, T. B. Lai, O. Kwong, and B. K. T'sou, "Morpheme-based derivation of bipolar semantic orientation of Chinese words," in *Proc. 20th Int. Conf. Comput. Linguist.*, 2004, pp. 1008–1014.
[14] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using WordNet to measure semantic orientations of adjectives," in *Proc. Int. Conf. Lang. Resourc. Eval.*, 2004, pp. 1115–1118.
[15] M. Hu and B. Liu, "Opinion feature extraction using class sequential rules," presented at the AAAI Spring Symposium Computational Approaches Analyzing Weblogs, Palo Alto, CA, USA, 2006, Paper AAAI-CAAW-06.
[16] D. Zeng, D. Wei, M. Chau, and F. Wang, "Chinese word segmentation for terrorism-related contents," in *Intelligence and Security Informatics*. New York, NY, USA: Springer-Verlag, 2008, pp. 1–13.
[17] W. Che, Z. Li, and T. Liu, "LTP:AChinese language technology platform," in *Proc. 23rd Int. Conf. Comput. Linguist., Demo.*, 2010, pp. 13–16