SENTIMENT ANALYSIS ON TEXTUAL REVIEWS

Mirsa Karim
PG Scholar, Department of Computer Science and Engineering St. Joseph's College of Engineering and Technology, Pala, Kottayam, India
E-mail: mirsakarim@gmail.com

Smija Das
Assistant professor, Department of Computer Science and Engineering St. Joseph's College of Engineering and Technology, Pala, Kottayam, India
E-mail: smija.das@sjcetpalai.ac.in

Abstract. These days there is an expansion in review websites. It has turned out to be considerably more intricate to mine fundamental data from survey sites and take proper choice. Using Natural Language Processing, there is need to identify sentiment of content or document. In this paper Sentiment Analysis is done in view of Rule based mechanism and machine learning approach. Both of these strategies are analyzed and discovered that machine learning is most appropriate for Sentiment Analysis in light of the exactness measurement. Sentiment Vader and Senti word net are the Rule based algorithms utilized and LDA analysis on Naive bayes is the machine learning strategy used.

1. INTRODUCTION
Sentiment Analysis is a forthcoming exploration progressing field that is becoming important because of utilization of different applications. Supposition Analysis is additionally called as sentiment mining. Audits are given by individuals in an unstructured way in type of forums, blogs etc. Then preprocessing of surveys is done and seen if the survey is certain, negative or nonpartisan. Order approaches like vocabulary and machine learning based methodologies are utilized for Sentiment Analysis. Vocabulary based approach is of word reference based approach and corpus based approach. Machine learning strategies are most broadly used to group and anticipate supposition as either positive or negative conclusion. Machine learning calculations are for the most part named directed or unsupervised approach. Regulated approach takes named dataset where each preparation set has effectively allotted its supposition. Unsupervised approach takes unlabelled dataset where audit isn’t characterized with its mark. Assessment investigation alludes to the undertaking of recognizing supposition from surveys.

1.1. Applications and challenges of sentiment analysis
The primary utilization of assumption investigation is in this manner giving the clients the possibility and suggestion in selection of products. A client is normally pulled in to certain part
of item if there should be an occurrence of picking an item. A solitary worldwide rating could be exact in decision making process. Estimation investigation can team up the conclusions of the analysis and assess evaluations on specific parts of the item. Another utility of feeling investigation is for organizations that need to know the interest of clients on their products. If client is unsatisfied with specific part of item organization can change the angle and furthermore help to discover what all viewpoints are more pulled in by clients. At last, Sentiment Analysis has been proposed as a part of different innovations. Sentiment Analysis for discovering conclusion or assessment of client about specific item.

Key challenges on opinion examination are tread precisely on exactness numbers, utiliize both machine learning and human knowledge, adopt a multi-technique look into plan, Stop regarding assessment investigation as a hobby, keep a receptive outlook about the discoveries.

2. RELATED WORK

Bhumika M. Jadav[4], performed Sentiment Analysis to order surveys in light of its assessment as either positive or negative category. Here unstructured information was changed over to organized information utilizing preprocessing and later to numeric score esteem. Rajat Bhat, Swapnil Gaonkar[15] utilized Hadoop Distributed File framework (HDFS) to store information set and keep running on MapReduce outline for performing Sentiment Analysis. Aruna Sathish[8] centers around the fundamental utilizations of Sentiment Analysis in web based business and in addition data security. By following the client remarks a reasonable thought of brand observation and consumer loyalty was acquired. Xiaojiang Lei[15], Xueming Qian propose a social client’s audits assessment estimation approach and compute every client’s supposition score on things/benefits and considered administration notoriety, which mirrors the clients’ thorough assessment. Finally, benefit notoriety factor is combined into suggestion framework to make a precise rating forecast.

Xueming Qian, He Feng[1] considered three necessary factors, personal attraction, interpersonal similarity consideration, and interpersonal impact, and combined them into a systemic accurate recommendation model probabilistic matrix factorization technique was used for recommendation model. Shengxiang Gao, Zhengtao Yu, Linbin Shi, Xin Yan, and Haixia Song[6] proposed a method to find expert rating by using previous rating histories, based on some association rules and result was found by creating probabilistic matrix factorisation for review experts.

Evangelos Psomakelis Konstantinos Tserpes[14] contemplated approaches like sack of words, n-grams. Order calculations and dictionary based methodologies were utilized for execution evaluation. Main arrangement calculations utilized for assumption investigation was Multilayer Perceptrons, Nave Bayes, SVM C4.5 as well as their mixes. The outcomes demonstrated that machine learning approach was most appropriate for anticipating assumption of tweet. Cataldo Musto, Giovanni Semeraro, Marco Polignano[9] proposed a vocabulary based approach for forecast of estimation in Twitter posts. This investigation depends on the lexical sources like SentiWordNet, WordNet-Affect, MPQA and SentiNet. Anas Collomb[10] proposed procedures for correlation of techniques used to assess the notoriety of things utilizing slant examination. The machine learning approaches utilize calculations to prepare dataset to anticipate the notion. The vocabulary based technique utilizes introduction of sentences for estimation examination. The administer based method characterize given sentence into positive and negative by feeling thought. Shreya Banker1 and Rupal Patel[11] played out a near report to discover the assumption related with audit. Measurable methods, like Naive Bayes are
better order strategies for supposition analysis. Machine learning methodology can take the valence of word for estimation in view of recurrence. Bruno Ohana [12] suggested that for conclusion investigation positive and negative word scores can be checked utilizing sentword net by assessment introduction.

3. METHODOLOGY
In this Paper Sentiment Analysis is done based on Rule based mechanism and machine learning approach. Both of these methods are compared and found that machine learning is best suited for Sentiment Analysis based on the accuracy measurement. Sentiment Vader and Sentiword net are the Rule based mechanisms used and LDA analysis on naive Bayes is the machine learning technique used.

3.1. DATA SET
Movie review dataset (http://www.cs.cornell.edu/people/pabo/movie-review-data/) from Cornell University labelled as positive and negative is taken.

3.2. RULE BASED MECHANISMS:
Rule based Sentiment Analysis effectively use rule mining algorithms to discover the features of a product and to find opinion associated with a particular product.

3.2.1. SENTIMENT VADER
VADER (Valence Aware Dictionary and Sentiment Reasoner) is a vocabulary and control based Sentiment Analysis instrument that is particularly receptive to assumptions communicated in social media. This calculation is utilized to group notion into four classes: positive, negative, neutral and compound. The compound score, is the entirety of all the dictionary evaluations which have been standardized to go between -1 and 1.

\[
\text{norm score} = \frac{\text{score}}{\sqrt{\text{score}^2 + \alpha}}
\]

The review is considered as positive, if compound score is \( \geq 0.5 \), is considered as neutral if score \( > -0.5 \) and \( < 0.5 \) and is considered as negative if score is \( \leq -0.5 \)

3.2.2. SENTIWORD NET
Sentiword net is Natural language processing tool for calculating score of word and to find opinion nature.

Sentiword algorithm is as follows:
Input file is created
Pos tagger is used to parse each sentence
Tag is allocated to each word i.e., whether noun, adjective, verb etc
Tag is then checked and assigned to sentiword net to calculate score
Sentiment type of word is then returned
Total no of positive and negative adjective words are counted for each sentence
If count is odd number then sentence is negative else positive

3.3. MACHINE LEARNING APPROACHES
Machine learning approaches learns the sentiment associated with the corresponding sentence but they require training data which is somewhat difficult to obtain and they are often much more computationally costly in associated with time taken for classification, cpu processing etc.
3.3.1. LDA ANALYSIS ON NAIVE BAYES:

3.3.2. Data Preprocessing for LDA

All the unique words are constructed in the vocabulary $V$, each word has a label $w_i$ 1, 2, . . . , Nd. For each document $d_j$, we choose a dimensional Dirichlet random variable $\theta_m \sim \text{Dirichlet}(a)$

(a) For each topic $z_k$, where $k \in [1, \gamma]$, we choose $\theta_k \sim \text{Dirichlet}(b)\text{.}$

(b) For each topic $z_k$, the inference scheme is based upon the observation that:

$$p(\Theta, \Phi | D^{\text{train}}, \alpha, b) = \sum_z p(\Theta, \Phi | z, D^{\text{train}}, \alpha, b) \times P(z | D^{\text{train}}, \alpha, b).$$

3.4. Extracting Product Features

Tags are added i.e., the symbol / before product features to distinguish other words in reviews.

LDA is a Bayesian model, which is utilized to model the relationship of reviews, topics and words.

The terminologies used in LDA model is described as:

1) $v$: vocabulary consisting of unique words
2) $w_i$: unique word in review
3) $d_m$: document associated with user
4) $T$: total number of topics
5) $\theta_m$: the multinomial distribution of topics
6) $\phi_k$: the component for each topic
7) $a, b$: Dirichlet priors associated with the multinomial distributions $m$ and $k$

3.5. User Sentimental Measurement:

Then for a review $r$ that user $u$ posts for the item $i$, Sentiment score is calculated as follows:

$$s(r) = 1 / N_c (Q * R_w * D_w)$$

where $c$ denotes the clause. $N_c$ denotes the number of clauses. $Q$ denotes the negation check coefficient. $D_w$ is determined by the empirical rule. The Naive bayes classifier works on Bayesian theorem and they perform classification using multinomial logistic regression.

The Naive bayes classifier is a simple classifier that relies on Bayesian probability and the naive assumption that feature probabilities are independent of one another. Training data give information about conditional and priori probabilities which feed to Naive bayes classification which form MachineLearning based classifier. In conditional probabilities look for words that appear more in positive and negative reviews. when a review is given it is split to words and look conditional probability whether they are likely to occur in positive or negative reviews.
3.6. Extract reviews
Extract reviews get reviews from positive and negative files and return 2 list one for positive reviews and other for negative reviews. For every review return tuple and label of review.

3.7. Build Vocabulary
For building Vocabulary iterate to all words in reviews and extract words. Then will get a unique collection of words. Add all words to a set. Then create a dictionary of words by iterating through all words in vocabulary. If a word is present in vocabulary dictionary hold true for value else show false.

3.8. Training
Extraction of features will give all the features needed for training data. It is then given to nltk classifier. Once we have trained classifier we will use classifier for individual reviews and then apply to test data, which individually classify positive and negative reviews and store it in list. Once prediction is complete see how many predicted labels are correct.

4. PERFORMANCE EVALUATION
SentiWord Net, Sentiment Vader and LDA Analysis on Naive Bayes was used in order to predict sentiment associated with review and found that machine learning technique is best suited for Sentiment Analysis. In Sentiwordnet and Vader, label 1 is assigned to positive review and 0 for negative review. Summing all values of positive and negative reviews with the total length of reviews is used for performance measurement. While in Naive bayes label 1 for positive reviews and -1 for negative reviews is used.

Based on the Accuracy measurement obtained from the same data set given it is shown that Sentiment Vader has 54.7 percent Accuracy which is low. Naive bayes has Accuracy of 75.2 which is high as compared to rule-based mechanisms which shows that machine learning techniques is more accurate in calculating Sentiment of reviews.

| Algorithm Used                  | SentiWord Net | Sentiment Vader | LDA on Naive Bayes |
|--------------------------------|---------------|-----------------|-------------------|
| Accuracy                       | 59.17         | 54.76           | 75.2              |

5. ARCHITECTURE
In the proposed system architecture Text Reviews ie, sequential data is given as input. Then Sentiment Analysis is done by Sentiwordnet, Sentiment Vader and LDA analysis on Naive bayes. Sentiword net use POS Tagging ie. assigning a unique tag to each word in review whether they are verb, adverb, noun or adjective. Then synset score is evaluated for Sentiment Analysis. Synsets are set of synonyms.

Vader uses compound score to calculate sentiment associated with review. For training Naive Bayes build a Vocabulary and extract features from the review. HDInsight is a completely overseen cloud benefit that makes it simple, quick, and savvy to process enormous measures of information. Utilize prevalent open-source systems, for example, Hadoop, Spark, Hive, LLAP, Kafka, Storm, R and more. Sky blue HDInsight empowers an expansive scope of situations, for example, ETL, Data Warehousing, Machine Learning, IoT and more. The IPython Notebook is presently known as the Jupyter Notebook. It is an intelligent computational condition, in which we can consolidate code execution, rich content, arithmetic, plots and rich media.
Power BI is a suite of business investigation instruments that convey bits of knowledge all through your association. Interface with several information sources, disentangle information prepare, and drive specially appointed investigation. Deliver lovely reports, at that point distribute them for your association to devour on the web and crosswise over cell phones.

![Proposed System Architecture](image)

**Figure 1. Proposed System Architecture**

### 6. CONCLUSION

Sentiment Analysis, otherwise called assessment mining is utilized to distinguish estimation related with a sentence. This examination would help the client to perform fitting rating. Sentiment examination dependably manages extremity of sentence. In this paper feeling examination is finished by utilizing Rule based and Machine learning based systems and found that machine learning procedure is more precise and accurate in anticipating the conclusion of a sentence or finding sentiment associated with sentence. Future work can be improved as research in joining a classifier with different methodologies, for example, word vectors may create preferable outcomes over every individual classifier can deliver without anyone else.
References

[1] XuemingQian, He Feng, Guoshuai Zhao, Tao Mei, Personalized Recommendation Combining User Interest and Social Circle, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2007

[2] Yan-Ying Chen, An-Jung Cheng, and Winston, Travel Recommendation by Mining People Attributes and Travel Group Types From Community Contributed Photos, IEEE TRANSACTIONS ON MULTIMEDIA, VOL.15, NO.6, OCTOBER 2013

[3] Wenjuan Luol, Fu Zhenu Zhang, Ratable Aspects over Sentiments: Predicting Ratings for Unrated Reviews, 2014 IEEE International Conference on Data Mining

[4] Bhumika M. Jadav, Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis International Journal of Computer Applications 2015

[5] Xiaojiang Lei, Xueming Qian, Rating Prediction via Exploring Service Reputation, 2015 IEEE 17th International Workshop on Multimedia Signal processing

[6] Shengxiang Gao, Zhengtao Yu, Linbin Shi, Xin Yan, and Haixia Song, Review Expert Collaborative Recommendation Algorithm Based on Topic Relationship IEEE/CAA JOURNAL OF AUTOMATICA SINICA, VOL. 2, NO. 4, OCTOBER 2015

[7] Hao Ma, Haixuan Yang, Michael R. Lyu, Irwin King, SoRec: Social Recommendation Using Probabilistic Matrix Factorization 2015 IEEE International Conference on Data Mining

[8] Aruna Sathish, Sentiment Analysis in E-Commerce and Information Security, International Journal of Innovative Research in Computer and Communication Engineering 2016

[9] Cataldo Musto, Giovanni Semeraro, A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts, IEEE International Conference on Data Mining 2016

[10] Anas Collomb, A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation, IEEE TRANSACTIONS ON MULTIMEDIA, 2016

[11] Shreya Banker1 and Rupal Patel, A BRIEF REVIEW OF SENTIMENT ANALYSIS METHODS International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016

[12] Bruno Ohana, Brendan Tierney, Sentiment Classification of Reviews Using SentiWordNet 2016

[13] Xiaojiang Lei, Xueming Qian, Rating Prediction Based on Social Sentiment From Textual Reviews, IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 18, NO. 9, SEPTEMBER 2016

[14] Evangelos Psomakelis, Konstantinos Tserpes, COMPARING METHODS FOR TWITTER SENTIMENT ANALYSIS Conference: International Conference on Knowledge Discovery; 2017

[15] Rajat Bhat, Swapnil Gaonkar, Sentiment Analysis of Product Reviews using Hadoop, IJSRD - International Journal for Scientific Research Development—Vol. 4, Issue 12, 2017